From Micro to Macro:

Essays on Rationality, Bounded Rationality and Microfoundations

Mohammad Reza Salehnejad

The London School of Economics and Political Science

Submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

2005

Abstract

This thesis examines some issues at the heart of theoretical macroeconomics, namely the possibility of establishing a predictive theory of individual behaviour *and* transforming it into a theory of the economy using aggregation. As regards individual behaviour, the basic idea in economics is that *homo economicus* follows the prescriptions of the expected utility theory. The thesis argues that the expected utility theory takes the agent's view of the economy as given, and is silent about how he models his choice situation and defines his decision problem. As a consequence, it is of only a minor contribution to the analysis of economic phenomena.

To explain how the agent learns about the economy and thus models his choice situation, new classical economists have relatively recently proposed that the agent behaves like a statistician. That is, like a statistician, he theorises, estimates, and adapts in attempting to learn about the economy. The usefulness of this hypothesis for modelling the economy depends on the existence of a 'tight enough' theory of statistical inference.

To address this issue, the thesis proposes a preliminary conjecture about how a statistician perceives and models a choice situation: the statistician regards measurable features of the environment as realisations of some random variables, with an unknown joint probability distribution. He first uses the data on these quantities to discover the joint probability distribution of the variables *and* then uses the estimate of the distribution to uncover the causal structure of the variables. If the resulting model turns out to be inadequate, the initial set of variables is modified and the two phases of inference are repeated. This setting allows the separation of probabilistic inference issues from those of causal inference.

The thesis studies both stages of learning from data to argue why there cannot be a 'tight enough' theory of statistical learning. As a result, the marriage of the hypothesis that the agent behaves like a decision scientist with the one that he behaves like a statistician is not of much help in predicting behaviour and modelling the economy. The thesis next turns to the other issue relating to the move from a theory of individual behaviour to a theory of the economy. It argues that to explain economic phenomena it is necessary to view the economy as a society of interactive, and heterogeneous, agents. However, the regularities emerging in such a society are not directly related to the laws operating at the micro level. The connection between the individual and the aggregate levels is highly complex.

Acknowledgements

The completion of this thesis has taken many years of study in a corner of the LSE library. During these years I continuously wrestled with the topic and often suffered with doubts. One thing that more than anything else has helped me through these years, and enabled me to bring the thesis to its present state were the encouraging comments made by Professor Nancy Cartwright on parts of the thesis. Her incisive, and reassuring comments often energized me for months, and made me believe that with persistence and hard work it would be possible to bring the thesis to a satisfactory end. So, my genuine special thanks and gratitude goes to her.

Equally, my sincere thanks go to Professor Colin Howson. The decision to devote my PhD thesis to the current topic was a gradual one. Narrowing down the topic and precisely defining it took longer than usual. The literature on bounded rationality, which forms a main part of the thesis, soon became extremely technical, and often beyond reach. Even more challenging was locating the real difficulty with the literature. Professor Howson's unbounded patience made it possible for me to acquire an adequate background for understanding this literature, and finding my way through it. If this thesis deserves to be dedicated to anyone, it should surely be to Professor Howson.

Among others I would like to express my special gratitude to Dr Cliburn Chan. Though by profession a biologist interested in nonlinear dynamics, Dr Chan was generous enough to read various parts of my thesis to teach me how to write and even how to restructure my arguments. His lessons have been a highly valuable source of intellectual wealth that will enrich my life for many years to come.

The ideas in the thesis have benefited from correspondence with many of those whose works are cited throughout the chapters. They often helped me understand the content of their papers, correct my understanding of their reasoning, and guided me to new and unknown territories. They are too many to be mentioned here, but I am grateful for their generous support. Their advice is present on every page of the thesis.

Contents

Chapter 1

Theoretical Versus Atheoretical Macroeconomics

Concepts and Controversies

Chapter 2

Rational Behaviour and Economic Theory

Chapter 3

Homo Economicus as an Intuitive Statistician (1)

Model Free Learning

Chapter 4

Homo Economicus as an Intuitive Statistician (2)

Bayesian Diagnostic Learning

Chapter 5

Homo Economicus as an Intuitive Statistician (3)

Data Driven Causal Inference

Chapter 6

The Economy as an Interactive System

An Appraisal of the Microfoundations Project

List of Tables

Introduction

> "As economics pushes on beyond 'static' it becomes less like science, and more like history." (Hicks, 1979: xi)

Modern economies consist of millions of heterogeneous decision-making units interacting with each other, facing different choice situations and acting according to a multitude of different rules and constraints. The interaction of the decision-making units at the micro level gives rise to certain regularities at the economy level, which form the subject matter of macroeconomics. The complexity of modern economies makes it impossible to build an analytic model that represents the behaviour of all the basic decision-making units populating the economy. In modelling the economy, it is inevitably necessary to leave many details out, introduce aggregate variables, and focus on the relations among the aggregates. Macroeconomics is therefore primarily the study of aggregates.

The study of the economy at the aggregate level presents a number of difficulties. For practical reasons, the economy cannot be subjected to controlled experiments to establish causal relations true of economic aggregates. Economists have to rely on statistical analysis of aggregate data to study the causal relations true of the aggregates. Statistical analysis alone, however, is inadequate for causal inference, and must always be supported with substantive information regarding the underlying structure to yield new causal conclusions. Moreover, aggregate economic data are inherently imprecise, rendering the outcomes of statistical analysis in macroeconomics even more uncertain. These difficulties raise the issue of how it is possible in macroeconomics to acquire the non-sample information needed for modelling the structure of the economy.

In response to this question, several competing approaches to macroeconomics have emerged. The so-called theoretical macroeconomics, championed by new classical economists, suggests that none of the methodological problems arise at the individual level. So, we can start by establishing a theory of individual behaviour, which explains how the agent interacts with the economy, defines his choice situation, and thus makes a decision. Once we have established a theory of behaviour, we can transform it into a theory of the economy as a whole using aggregation procedures. Since the theory is directly derived from the rules of individual behaviour, it correctly specifies the causal structure of the economy. Aggregate data can then be used to transform the theory into a quantitative model of the economy, describing the economic structure.

The enterprise of deriving the correct macroeconomic theory from microeconomic theory – or the *microfoundations project* – rests on two assumptions. The first is that we have, or it is possible to establish, an empirically adequate theory of individual behaviour. The other is that the theory can be transformed into a theory of the economy using aggregation procedures, without having to introduce any substantive assumption about the structure of the economy.

As regards individual behaviour, the basic idea in economics is that *homo economicus* follows the prescriptions of decision theory, understood in terms of one or another expected utility theory, in particular the theory of subjective expected utility. The theory of expected utility, in all the variants on offer, takes the agent's view of the economy as given, and says nothing about how he predicts future values of economic variables. To fill this theoretical vacuum, new classical economists have set forth the rational expectations hypothesis, which identifies the agent's subjective beliefs about the economy with the mathematical expectations implied by the true economic model. This gives rise to a view of the economy as a society in which everyone, except the econometricians, knows the structure of the economy. On this basis, the new classical paradigm defines economics as the enterprise to derive observable economic phenomena from two fundamental assumptions: (1) people are expected utility maximisers and (2) they maximise their expected utility with respect to the true model of the economy.

Theoretical problems with the rational expectations hypothesis have led to a slow paradigm shift in new classical economics that aims to remove the information asymmetry existing between the econometricians and people in a rational expectations economy by suggesting that, like econometricians, market participants also lack knowledge of the true economic model, and must learn it from available economic data. The new paradigm has been dubbed the bounded rationality paradigm, after Herbert Simon (1955, 1956). Though the idea of bounded rationality is relatively old, a unanimous interpretation of the paradigm is yet to emerge. One leading interpretation, set forth by new classical economists, conceives of the economy as a society of 'intuitive statisticians', where everyone, like econometricians, theorises, estimates, and adapts in attempting to learn about probability distributions that, under rational expectations, they already know (Sargent, 1993:3). So understood, the paradigm replaces the second principle of new classical economics with the assumption that agents maximize their expected utility with respect to models that, like econometricians, they construct from economic data. We will refer to the proposal that homo economicus behaves like an econometrician as the *intuitive statistician* hypothesis of bounded rationality.

This thesis studies several foundational issues relating to the theoretical approach to macroeconomics or, more specifically, the microfoundations project.

Chapter 1 begins with defining some key concepts in macroeconomics, outlines several arguments for the necessity of theory in modelling the economy, and characterises the theoretical approach in some detail. To pave the way for defining some basic controversies in macroeconomics, the chapter then reconstructs the so-called atheoretical approach to macroeconomics, which stands at the other extreme end of the spectrum of views on the nature and scope of macroeconomics. The view rejects both suppositions of theoretical macroeconomics. It argues that current theories of individual behaviour lack precision and substantial difficulties face any attempt to make them precise. More importantly, because of individual heterogeneity and interaction among decision-making units in the economy, the view rules out any simple, and useful, relation between the individual and economy levels. As a consequence, it confines the scope of macroeconomics to establishing models that efficiently summarise data, and are useful for short run predictions. In one reading, the approach rejects the very existence of aggregate relations suitable for a causal account. The contrast between these views reveals that the issues regarding theories of economic behaviour and those about the link between the micro and macro levels are the most basic topics in macroeconomics. Of equal importance is the conjecture that one can sensibly talk of structural relations at the economy level.

Chapter 2 studies the contribution of rational choice theories to economic theorising by concentrating on Savage's theory of subjective expected utility. Using the framework of the theory, the chapter distinguishes between several phases of human decision making, which include (i) modelling the choice situation, (ii) defining the decision problem, and (iii) solving the problem. In light of this, the chapter distinguishes between two possible types of theories of behaviour: choice-based theories of behaviour and learning-based theories of behaviour. The rational choice theories on offer fall into the category of choice-based theories of behaviour; they take for granted how the agent models his choice situation and defines his decision problem, and only explain how he solves a well-defined decision problem. So, in modelling behaviour using these theories, a host of substantive assumptions are needed to specify the agent view of his choice situation, and the problem he is trying to solve. These assumptions concern the agent's view of the causal structure of the environment, his values, beliefs, needs, and goals.

The chapter then demonstrates that the resolution of economic controversies *primarily* hinges on how the agent models his choice situation and defines his decision problem, rather than on the *specific* method by which he solves the problem. In fact, the rational choice theories are consistent with all sides of any substantive controversy in economics, and contribute very little to economic analysis. Substantial results, usually attributed to these theories, are the implications of the assumptions made about how people specify their choice situation, and redefine it when faced with changes in the economy or new information. As a consequence, a theory of economic behaviour cannot take as given the structure of the choice situation and how the agent defines his decision problem. Economics is thus in need of a learning-based theory of behaviour, which explains how the agent models his choice situation, defines his decision problem, and redefines it as a result of experience. The chapter ends with an analysis of the rational expectations hypothesis to explain why the hypothesis fails to eliminate the necessity of a learning-based theory of behaviour in theoretical economics.

Chapter 3 begins studying the intuitive statistician hypothesis. The usefulness of this hypothesis for modelling and thinking about the economy depends on whether there exists a 'tight enough' theory of statistical inference. To address this issue, the thesis proposes a preliminary conjecture about how a statistician perceives and models a choice situation: The statistician regards measurable features of the environment as realisations of some random variables, with an unknown joint probability distribution. He first uses the data on these quantities to discover the joint probability distribution of the variables *and* next uses the estimate of the distribution to uncover the causal structure among the variables. If the resulting model turns out to be inadequate, the initial set of variables is modified and the two phases of inference are repeated. This setting allows the separation of probabilistic inference issues from those of causal inference.

Central to learning the joint probability distribution of a set of variables is model specification (formulation), not estimation or hypothesis testing. Whether there is a 'tight enough' theory of statistical learning critically thus depends on whether or not there is a 'tight enough' theory of model formulation. Having said this, to study the issue of model formulation at its most general level, the chapter turns to nonparametric inference, which theoretically seeks to design algorithms that receive data on a set of variables and yield the model that, given the data, best approximates the underlying mechanism. The framework explains why there cannot be such algorithms, highlights intrinsic limitations of model-free inference, and establishes the necessity of probabilistic background information for building interpretable statistical models. With the data samples normally available, one must begin with a parametric model to obtain an interpretable model of the data, which raises the question of where the parametric models come from.

Chapter 4 studies the topic of statistical learning from the perspective of the Bayesian theory, which is said to allow the incorporation of background information into inference. The chapter first looks at some critical issues at the foundation of the Bayesian theory to explain why the theory, as stands, cannot be a theory of learning, and is only concerned with coherent analysis. As a result, to explain central aspects of inference such as model specification, empirical model assessment, and re-specification analysis, one has to go beyond the boundaries of the Bayesian theory. Having done this, the chapter draws on several themes in the relatively recent statistical literature to reconstruct a broader theory of Bayesian inference, which takes some steps in explaining the central aspects of inference traditionally left out in the Bayesian literature, including model formulation. Reflecting on the requirements and limits of the broader theory, the chapter considers the possibility of establishing a 'tight enough' theory of parametric inference, and brings to the fore some important implications for the bounded rationality program.

Chapter 5 studies the second phase of statistical learning relating to inference about causal structure. The chapter concentrates on the graph-theoretic approach to causal inference in order to investigate the possibility of a data-driven approach to causal inference. By data-driven we mean any effort to draw causal conclusions from probabilistic data using only general subject-matter-independent principles supposedly linking causation and probability. A claim for a data-driven approach to causal inference raises two separate issues. The first is whether there are universal principles connecting probabilistic and causal dependencies. The other is whether the principles are sufficient for inferring from the joint probability distribution of a set of variables the causal structure generating the distribution. The chapter takes up both topics, and by reflecting on the limits of data-driven causal inference outlines an account of causal inference from observational data.

The analysis in these last three chapters helps us judge if there can be a 'tight enough' theory of statistical learning, explaining all the general phases of learning from data.

Chapter 6 studies the other element of the microfoundations project that has to do with the move from a theory of individual behaviour to a theory of the economy. The chapter starts with a critique of the representative agent modelling approach to the study of the economy, explaining why understanding large-scale economic phenomena calls for thinking of the economy as a society of interactive heterogeneous individuals. Having done so, it investigates some basic issues that individual heterogeneity and interdependencies create for theoretical study of the economy. Individual heterogeneity and interdependencies existing in modern economies fundamentally undercut the conception of the economy underlying the microfoundations project. In fact, they sever any simple, direct, and useful link between the micro and macro levels, casting doubt on the very existence of stable relations at the economy level, which are suitable for a causal account.

The thesis concludes by highlighting some of the implications that arise from the chapters for the bounded rationality program in particular, and for the study of the economy in general. The marriage of the hypothesis that the agent behaves like a decision scientist with the hypothesis that he behaves like an intuitive statistician is not of much help in predicting the course of the economy.

Several disclaimers apply to this thesis. A good part of this thesis concerns modelling bounded rationality. However, it is only concerned with certain foundational issues regarding one interpretation of the bounded rationality project. The thesis does not touch on the literature on learning in games. It does not cover the literature that models adaptive learning without assuming that the agent forms a mental model of his choice situation. Nor does it explain how introducing bounded rationality helps explain economic phenomena not explainable within the framework of full knowledge and rationality. Though our analysis highly relevant to these matters, they are subjects for future research.

Chapter I

Theoretical versus Atheoretical Macroeconomics

Concepts and Controversies

1 Introduction

> Human beings in society have no properties but those which are derived from, and may be resolved into, the laws of the nature of the individual man. In social phenomena the Composition of Causes is the universal law (J.S. Mill, 1974 [1874]: 879).

The study of fluctuations in aggregate measures of economic activity and prices over relatively short periods (business cycle theory) and development of the economy over the long run (growth theory) constitutes what we call macroeconomics. The objective is to understand the causes of economic fluctuations and growth, forecast the future course of the economy, and aid analysis of state policies. Following tradition, we may categorise major objectives of macroeconomics under the headings of explanation, forecasting, and policy analysis. These objectives call for a quantitative model, which for most purposes must represent the causal structure of the economy. In macroeconomics, a major methodological issue is, therefore, the understanding of the causal structure of the economy.

A modern economy consists of millions of individual decision makers, firms, and institutions, each solving different decision problems under diverse circumstances and subject to distinct social and economic constraints. This complexity makes it practically impossible to build a model which represents the behaviour of all the basic decision making units of the economy. In modelling the economy, we inevitably need to leave out many details of the decision-making units, introduce aggregate variables, and focus on the relations of the aggregates. For this reason, macroeconomics is primarily the study of aggregates.

The study of the economy at the aggregate level presents a number of methodological difficulties. To begin with, for practical reasons, it is impossible to subject the economy to controlled experiments. This impossibility deprives macroeconomics of a reliable method for causal inference which is available to other scientific fields. Economists inevitably have to rely on statistical analysis of aggregate data to study the causal structure of the measured variables. However, statistical analysis is inadequate for causal inference, and has to be supported with domain-specific information regarding the structure under study to yield new causal conclusions. This raises the issue of how, in macroeconomics, we can acquire the subject-matter information necessary for a causal analysis of aggregate data.

There are two major reactions to these methodological limitations. On the one hand, there is the view that acknowledges the limitations involved in directly studying aggregate data but argues that they can be overcome indirectly. We can, so goes the argument, start by studying how individuals make their decisions, and thereby establish a theory of individual behaviour. Once we have established a theory of individual behaviour, we can transform it through aggregation into a qualitative theory of the economy. Having done this, with the help of statistical methods, we can use aggregate data to estimate the model parameters and thus establish a quantitative model. Since the model is derived from the rules of individual behaviour or other basic decision making units, it correctly specifies the causal structure of the economy, which is needed for the purposes of explanation and policy evaluation. We call this approach *theoretical* macroeconomics.

The other response argues that current theories of individual behaviour lack precision and substantial difficulties face any attempt to make them precise. Moreover, many complex issues arise in aggregating individual behaviour, relating to the heterogeneity of behaviour and interactions among decision makers. The aggregation difficulties rule out any simple relationship between the individual and the economy level, undermining the proposal of theoretical economics for establishing causal structure. Econometricians can build models that summarise economic data but cannot establish whether a model represents the economic structure. Aggregate data are also imprecise. This imprecision further weakens the reliability of macroeconomic models as a tool for policy analysis. Economists must accept irresolvable differences among themselves and limit their claims to conclusions that follow from the analysis of imprecise aggregate data, which is the only objective ground in macroeconomics. We call this view *atheoretical* macroeconomics.

This chapter explains some central notions in economics, elaborates on some of the limitations of statistical analysis for causal inference, and briefly reconstructs the above two approaches to macroeconomics. The objective is to provide a glimpse of some basic controversies in macroeconomics and highlight the importance of the questions studied in this thesis.

This chapter is organised as follows: Section 2 defines macroeconomics, its subject matter and objectives. Section 3 discusses three arguments for the necessity of theory in macroeconomic analysis; all concern the boundaries of statistical analysis for causal inference. Section 4 sketches out the theoretical approach to macroeconomics, also known as the microfoundations program, and describes the assumptions underlying the program. Section 5 outlines the main features of the atheoretical approach and contrasts it with theoretical economics. Section 6 concludes with an outline of the issues studied in the thesis.

2 Macroeconomics

Macroeconomics studies recurring patterns discernible at the aggregate economic level, aiming to build a model that serves to explain why such patterns occur, predict their future developments, and analyse how policy interventions might affect them. This section first defines the fundamental notions of *structure* and *structural model.* It then characterises the objectives of macroeconomics and the features that a model should possess to serve as a means for achieving the objectives. In doing this, we follow the view of macroeconomics developed by the early econometricians working in the Cowles Commission.

2.1 Macroeconomic Structure

The variables used to describe an economy originate in the decisions made by its components - numerous individuals, firms, institutions, and governments.[1] Families make decisions about what to consume and when, how many hours to work, and what

[1] The following view of the economy is borrowed from Granger 1990a.

to invest in; firms make decisions about unemployment, production, pricing, marketing, borrowing, investment, and so forth. We can consider the input and output variables of these decision-makers as the basic variables of the economy. Often, the eventual outcome of the decisions is not quite what was planned: poor health may disrupt work, or a supply shortage may result in a lower production than expected. It is therefore sensible to model the decision outputs of an individual or institutional decision-maker as a function of its prominent input variables plus a stochastic residual vector. The vector of all input and output variables of all participants in the economy and their linking functions form the *microstructure* of the economy. We denote the microstructure by (***m***, ***r***), where ***m*** and ***r*** stand for the vector of micro-variables and micro-relations respectively. In place of (***m***, ***r***), the economic literature often refers to the triple (***m***, ***r***, ***p***) as the true economic data generating mechanism, where ***p*** is the joint probability distribution of the micro-variables ***m*** (Granger, 1990a:7; Hendry and Ericsson, 1991:18; Spanos, 1986:661-72).

The immense number of micro-variables and relations in any modern economy makes it impossible to consider the behaviours of all decision-making units in a model. Modelling an economy requires introducing aggregate variables and focusing on the patterns that emerge at the aggregate level. A central assumption in economics is that the microstructure (***m***, ***r***, ***p***) leads to a unique *macrostructure* (***M***, ***R***, ***P***), where ***M*** stands for the set of aggregate input and output variables, ***R*** for the relations among the aggregates, and ***P*** for the joint probability distribution of the aggregate variables (Epstein, 1987:65). This macrostructure is the subject matter of macroeconomics. Thus, macroeconomics is defined as the study of the aggregative relations that emerge from decisions and interactions of the basic decision-making units of the economy. By contrast, microeconomics is defined as the study of the behaviour of the basic decision-making units of the economy in which no aggregation is involved. The presence of aggregation is what separates these fields of study from each other (Keynes, 1936:292-3).[2]

[2] See Janssen (1993, ch. 1) for various notions of macroeconomics.

The above view of the subject matter of macroeconomics is consistent with mainstream economics that postulates the existence of recurring and stable relations at the aggregate level. However, there are several notions of macroeconomic structure and a host of views on the nature of the relation between the micro and macro structure. We continue by explicating a notion of structure found in the writings of the early econometricians working in the Cowles Commission, which will serve as a benchmark for defining alternative notions of structure and characterising some key controversies in macroeconomics.[3]

2.1.1 Structural Models

Early econometricians rarely defined the notion of structure explicitly. Instead, they focused on the related notion of *structural model.* It is convenient to first describe this notion and then use it to define the concept of structure. Consider the simple stochastic equation

$$Y = \alpha + \beta X + \varepsilon, \qquad (2.1)$$

where Y is the response variable, X is the regressor, and ε is the error term with mean zero.[4] This equation is commonly used to represent the regression of Y on X, giving the mean of the distribution of Y conditional on a particular value of X, i.e., $E(Y / X = x)$ (Greene, 1990, ch.10). As a regression equation, (2.1) describes the association between X and Y in the population from which the data are sampled. As opposed to this usage, the equation may also be used for predicting the effects of (hypothetical) interventions in X on Y. If the equation correctly predicts how the values of Y change as we intervene to change the values of X, it is called *structural* (Hurwicz, 1962:236-7). A difference between (2.1) as a regression equation and (2.1) as a structural equation is that in the former case the equation may cease to hold as

[3] For the history of the Cowles Commission Foundation see Darnell et al. (1990) and Epstein (1987).

[4] We adopt the usual convention of denoting random variables by upper-case letters, and their values by the corresponding lower-case letters. Likewise, we denote random vectors by bold upper-case letters and their values by the corresponding bold lower-case letters.

soon as changes are made to X whereas in the latter case the equation is invariant to interventions made to the values of X. Another way to state this notion of structural equation in the context of simple equation (2.1) is the following, which is borrowed with small changes from Pearl (2000:160):

Definition: An equation $Y = \alpha + \beta X + \varepsilon$ is said to be structural if in an ideal experiment where we control X to x and any other set Z of variables (not containing X or Y) to z, the value of Y would be independent of z and is given by $\alpha + \beta x + \varepsilon$.

This notion of structural equation captures the core of the manipulability conception of causation (Woodward, 1999). On this view of causality, variable X causes variable Y if it is possible at least in a hypothetical experiment to change Y by manipulating X. So, the claim that equation (2.1) is structural means that it expresses a causal relation. In that case, the parameter β in (2.1) reflects the causal effect of X on Y, contrary to a regression equation in which β only represents the degree of association between X and Y in the population. The terms 'structural' and 'causal' are used interchangeably in what follows.[5]

Econometricians refer to the variables on the right hand side of a structural equation as *exogenous*; variable X in equation (2.1) is exogenous if intervening to set $X = x$ gives the same result for Y as observing $X = x$. Similarly, the variable on the left-hand side of a structural equation is called *endogenous*. Exogeneity also has weaker meanings in the literature. It sometimes refers to a variable whose value is not explained within the model but is supplied to it and sometimes refers to a variable which is statistically independent of the error term in the equation. We use 'exogeneity' to refer to the first notion, i.e., as an independent variable in a structural equation (Engle, Hendry, and Richard, 1983).

The early econometricians generalised the notion of a structural equation to systems of equations. An equation system forms a *structural model* if each equation in the

[5] See Hurwicz (1962) for the connection between causal and structural relations.

system is structural and remains invariant to changes that invalidate other equations in the model.[6] In light of the definition of a structural equation as a causal relation, each equation in a structural model represents an autonomous causal mechanism that can be modified without undermining the mechanisms represented by other equations in the model. As an illustration, consider a simple version of the model of demand and price determination in economics, which has been discussed by many authors including Goldberger (1992) and Pearl (2000:27):

$$Q = \alpha_1 P + \beta_1 I + \varepsilon_1 \tag{2.2a}$$

$$P = \alpha_2 Q + \beta_2 W + \varepsilon_2 \tag{2.2b}$$

where Q is the quantity of household demand for product A, P is the unit price of A, I is household income, W is the wage rate for producing A, and ε_1 and ε_2 are error terms, representing unmodelled factors that affect quantity and price respectively. This model is structural if equation (2.2a) correctly forecasts the effects on Q of (hypothetical) interventions in P or I, and equation (2.2b) correctly predicts the effects on P of interventions in Q or wage W. Moreover, interventions invalidating (2.2a) must not invalidate (2.2b) and vice versa.[7] If we change the values of the parameters α_1 and β_1 by intervening in the mechanisms determining the household income I, the change must not affect α_2 and β_2. The mechanisms represented by these equations must be unrelated. In short, what makes this model structural is that each equation characterises an autonomous causal mechanism, one equation describing the causal process determining the demand for A and the other the process determining the price for A.[8]

[6] A broader notion of structural model is found in Koopmans (1949 [1971]).

[7] A requirement necessary for this exercise is that ε_1 be independent of P, ε_2 independent of Q, and ε_1 independent of ε_2.

[8] For further discussion see Goldberger, 1992; Aldrich, 1987; Pearl, 2000; and Woodward, 1999.

This concept of structural model reveals the notion of structure implicit in the writings of the Cowles Commission econometricians. According to these researchers, a structure consists of a set of autonomous causal relations that can be utilised separately for intervening in the state of the economy. Koopmans, a leading member of the Commission, recapitulates this concept of structure in the following passage:

> The study of an equation system derives its sense from the postulate that there exists one and only one representation in which each equation corresponds to a specific law of behaviour (attributed to a specific group of economic agents) ... Any discussion of the effects of changes in economic structure, whether brought about by trends or policies, is best put in terms of changes in structural equations. For these are the elements that can, at least in theory, be changed one by one, independently. For this reason it is important that the system be recognisable as structural equations (quoted in Epstein, 1987:65).

From this perspective, the subject matter of macroeconomics is the study of autonomous causal relations true at the economy level, emerging from the interactions of individual decision makers with each other and with the environment. In what follows, we will refer to this viewpoint as the *received view*, and use it as a benchmark to define and compare some alternative views on the nature and scope of macroeconomics.

2. 2 Macroeconomic Objectives

To complete the description of the received view, it is also vital to expound on the objectives traditionally set for macroeconomics. This demands an understanding of the framework within which economic analysis is usually carried out. In its simplest form, consider an economy whose state at time t can be described by an endogenous variable Y_t and an exogenous variable X_t.[9] The dynamics of the economy is described by a difference equation:

[9] The description to follow draws on Lucas (1976) and Cooley et al. (1984).

$$Y_{t+1} = f(Y_t, X_t, \theta, \varepsilon_t) \tag{2.3}$$

where θ is a parameter vector defining the function *f*, and the disturbance term (random shock) ε_t has probability distribution $P(\varepsilon_t)$. The description of the economy is completed by specifying the mechanism generating the exogenous variable X_t, shown by

$$X_t = g(Z_t, \lambda, e_t) \tag{2.4}$$

where Z_t denotes the only variable affecting X_t, λ a parameter vector defining the function *g*, and e_t a disturbance term with probability distribution $Q(e_t)$.[10] The functions *f* and *g* are taken to be fixed but not directly known or at least not fully known. Data on X_t and Y_t is used to estimate θ and λ, as well as the parameters of the distributions $P(\varepsilon_t)$ and $Q(e_t)$. This fitted model is then used for prediction, policy analysis and explanation.

2.2.1 Prediction

Given an estimate of the parameters in the model (2.3), the task in prediction is to estimate the expected value of Y_{t+1} when the values of Y_t and X_t are given. For the time being, the presence of Y_t can be overlooked. Depending on how the value of X_t is given, three categories of prediction can be distinguished. First, there are cases where the actual value of X_t is known and the model is used to predict Y_{t+1}. Such predictions are called *ex post* predictions, since the actual value of X_t is *already* known. Secondly, there are cases where the actual value of X_t is not yet known, and one instead uses *guessed* values of X_t to predict future values of Y_t. Such predictions

[10] Z_t may be the same as Y_t..

are called *ex ante* forecasts. In both *ex ante* and *ex post* prediction, one acts as an observer. Thus, if the model closely approximates the associations in the population during the periods for which the predictions are made, it will correctly forecast future values of Y_t, regardless of whether it is structural or not.[11] A regression model is sufficient for *ex ante* and *ex post* prediction, and no understanding of the causal mechanism generating Y and X is required.

Besides these types of predictions, the concern in economics is often *conditional* prediction; that is, the objective is to predict the likely value y_{t+1} that would arise if X_t could be set at a value different from its actual value. Since conditional prediction implies setting the values of an exogenous variable through intervention, the model must be structural or, in other words, invariant to the intervention to predict the outcome of the intervention. A regression model of the observed regularities is not adequate; one needs an understanding of the structure generating the data (Lucas and Sargent, 1979:298).

2.2.2 Policy Analysis

The objective in policy analysis is to design changes in the economy that take it to a desired state. In its simplest form, a policy consists of a change in the value of a policy variable, say X_t, to alter the value of the target variable Y_{t+1} – a policy variable is an exogenous variable whose values can be modified through state intervention. Analysis of such a policy involves predicting alternative values of Y_{t+1} that would arise if X_t were set at values different from its actual value. If such predictions were possible, the future values of Y_t could be estimated for various values of X_t to find a value that would yield the desired result.

More often, a policy change is defined as a change in the *mechanism* that determines a policy variable. In the context of our simple economy, this concept of policy intervention amounts to a change in the mechanism

[11] Fair (1987:271) defines various notions of predictions.

$$X_t = g(Z_t, \lambda, e_t)\,. \qquad (2.4)$$

The underlying idea is that each set of possible values for the parameters λ defines a possible mechanism for X_t. A policy change then consists of a change in the values of these parameters to influence the course of the economy (Tinbergen, vol.2 1939:18). The analyst considers a different set of values for λ than the actual values to define an alternative mechanism for X_t. Having done so, he uses the rule to generate a sequence of hypothetical values $\{x_t\}$ and recursively inserts them in the model (2.3) to simulate the course of the economy under the rule. The exercise is repeated for plausible values of λ to select a rule with the desired outcome. A crucial requirement for the success of this exercise is that the model (2.3) remains invariant to changes in the policy rule (2.4). If a change in the mechanism generating X_t undermines the relation (2.3) that governs the behaviour of the endogenous variable Y_t, then an estimated version of (2.3) will not correctly predict the course of the economy under alternative policy rules (Lucas, 1976).

Policy analysis in either sense involves setting the values of the exogenous variables by intervention, and, for that reason, the relations in the model must be causal. Another equally important point is that if policy change involves a regime (rule) change, other relations in the model must be invariant to the modification in order to be of any use in predicting the policy outcomes. If one modifies equation (2.4), equation (2.3) must be invariant to the modification in order to be of any use in simulating the course of the economy.

2.2.3 Explanation

Another related concern in economics is to understand why certain *particular* facts are as they are. For instance, it is of utmost significance for policy analysis to understand why the inflation rate was at 2.5% in the UK last year or why an increase in the interest rate by 1% did not have the expected effect on the housing market. Still another concern, also related to policy analysis, is to understand the regularities that emerge at the economy level. To give an example, the Western economies post World War II displayed a trade-off between the rate of inflation and unemployment, known as the Phillips curve. The curve suggests that an increase in inflation is followed by a decline in unemployment. Any use of this regularity as a means for designing and evaluating employment policies demands understanding why the relation exists and, equally important, under what circumstances it may cease to hold. These queries fall under the general heading of explanation. We draw on some basic issues in the philosophical literature to find out what constitutes an adequate explanation of a particular fact. This will help us understand the features that a model should have in order to play a role in the explanation of economic phenomena.[12]

An early theory of scientific explanation, put forward by Hempel and Oppenheim ([1948] 1965), defines an explanation of a particular fact as an argument to the effect that the phenomenon to be explained was to be expected by virtue of certain explanatory facts (Hempel, 1965:336).[13] The premises in the argument constitute the *explanans* (that which does the explaining) and the conclusion is the *explanandum* (that which is explained). Hempel requires that the explanans include at least one lawful generalisation. This view of explanation has become known as the *inferential* view, since it identifies an explanation with an argument. Schematically an explanation in this approach takes the form:

True statements of initial conditions Laws	} *Explanans*
Statement of what is to be explained.	*Explanandum*

[12] It is beyond the scope of this thesis to touch on the question of what constitutes a satisfactory explanation of a law or regularity. A seminal paper on this topic is Friedman (1974). Chapter 6 can also be viewed as an exercise in explanation of regularities.

[13] The following analysis of explanation is based on Salmon (1998, chapter 19).

Hempel distinguishes two models of explanation of particular facts. For contexts in which universal laws are available, such as the physical sciences, he proposes his deductive-nomological (D-N) model of explanation. A D-N explanation is a valid argument whose premises include at least one universal law and which deductively entails the explanandum. To give a well-known example, according to Hempel, we can explain why John has Down's syndrome by deducing the fact from the initial condition that John's cells have three copies of chromosome 21 and the law that any person whose cells have three copies of chromosome 21 has Down's syndrome. It is an essential characteristic of a D-N explanation that its explanans should include at least one lawful generalisation; accidental generalisation, Hempel says, do not explain.

For contexts such as macroeconomics in which generalisations are usually statistical, Hempel introduces his inductive statistical (I-S) model of explanation which proceeds by subsuming the event-to-be explained under a statistical law. An I-S explanation is also an argument, with the difference that its premises do not logically entail the explanandum, and are only required to confer a high probability on it. In a simple example from Hempel, if one asks why John rapidly recovered from his streptococcus infection, an I-S explanation is that he took a dose of penicillin, and almost all strep infections clear up quickly upon administration of penicillin. This, in Hempel's view, forms an adequate explanation since the explanans are true and confer a high probability on the explanandum.

Unlike in a D-N explanation, the explanans in an I-S explanation do not logically entail the explanandum but are required to confer a high probability on it. This difference leads to a host of distinct issues for the I-S model of explanation which originate from the reference class problem. According to Hempel, an inductive statistical argument must refer to at least one statistical law to be an adequate explanation. The law gives the probability of the explanandum *E given* that it is a member of a *reference* class *C*. The problem is that the reference class *C* is usually inhomogeneous in the sense that it can be partitioned in a way that affects the probability of the explanandum. For instance, the generalisation in the explanation of

John's recovery is stated in terms of the class of people who take penicillin after having strep infection. In this class, the frequency of quick recovery from strep infection is high and thus, according to Hempel, we can explain John's recovery by the fact that he has taken penicillin. But this class can be partitioned into two subclasses: one subclass consisting of people with strep infection who are resistant to penicillin and the other consisting of those who are not. If John is a member of the former subclass, taking penicillin no longer explains his recovery, despite the fact that in both explanations the premises can be true. In general, an event is a member of many classes. Depending on the class chosen to subsume the event, we can explain the same event (John's recovery) with different degrees of probability, or even explain contradictory events (non-recovery). This raises the question as to how to choose a reference class to explain an event.

To address this question, Hempel adds that an I-S explanation must refer to a statistical generalisation that is stated in terms of a 'maximally specific' reference class to be satisfactory. By this, he essentially means that given our background knowledge it must not be possible to partition the class C in a *nontrivial* way that affects the conditional probability of the explanandum. To be precise, the class C is said to be maximally specific if the probability of the explanandum E is the same in any of its subclasses; that is, if $P(E/C_i) = P(E/C_j)$ for any i and j, where $(C_1, C_2, ..., C_n)$ is any partition of C. According to Hempel, an I-S explanation is then adequate if the explanans are true, confer a high probability on the explanandum, and satisfy the requirement of maximal specificity.[14]

A problem with the maximal specificity requirement is that not all partitions affecting the probability of the explanandum are permissible. For example, given that John has lung cancer, that he has worked in a chemical factory where many employees have contracted lung cancer, that he has nicotine stained fingers, and that the frequency of lung cancer among people with nicotine stained fingers is higher, the requirement calls for using the statistical generalisation that provides the probability of lung cancer

[14] On this account, the original explanation given above for John's quick recovery fails to satisfy the maximal specificity requirement because the class of people who take penicillin can be partitioned in a way that changes the conditional probability of John's recovery.

among employees with nicotine stained fingers. But this partition is not permissible, since having nicotine stained fingers and having lunge cancer could be the effects of a common cause (heavy smoking), and an effect of a common cause does not explain another effect of the same cause. The moral of this story is that for an argument to be explanatory the explanans must be causally relevant to the explanandum. Mere statistical association is not sufficient (Salmon, 1998:309).

Another problem with the I-S model of statistical explanation is that it is symmetrical. Given a statistical association between Gaussian random variables *X* and *Y* which have no common causes, but *X* causes *Y*, the I-S model implies that one can explain *X* by *Y* or *Y* by *X*. This implication goes against the intuition that, while causes explain effects, effects do not explain causes (Granger, 1988:17; Sobel, 1995:13). The conclusion is that the mere presence of a stable statistical correlation is not adequate for a statistical argument to be explanatory. An adequate explanation must relate the explanandum to its causes.

These considerations expose the difficulty of developing a theory of explanation of particular facts that makes no reference to causal relations. An explanation of a particular fact must give information relating to the causal process that has generated it. As Lewis (1986) notes, to explain a particular fact is to give information about its causal history. In general, whenever we try to explain a particular phenomenon, it must be shown that (1) the explanatory events are actually true, (2) the events are causes of the explanandum, in that if they were present and there were no preventative causes, the explanandum would occur too, and in addition (3) the events are *actually* the causes of the explanandum in the sense that if they had not been present in the situation under study the explanandum would not have occurred.[15] The reason for the inclusion of this last condition is that for any event there might be several sets of sufficient causes that could bring it about. Explanation of particular facts, therefore, calls for knowledge of the causal structure, and a model must be structural to be useful in explaining particular phenomena.

[15] This is true if no other sufficient set of causes is present.

To summarise the received view of macroeconomics, there is a structure behind the aggregate data, consisting of autonomous causal relations that, at least hypothetically, can be manipulated independently of each other. The prime task of macroeconomics is then defined to be understanding and modelling of the structure. In addition, all the objectives traditionally set for macroeconomics, namely *ex ante* and *ex post* prediction, conditional prediction, policy analysis and explanation, are considered as achievable.

3 The Need for Theory

A query for the received view is how the structure of the economy can be discovered. Natural sciences usually appeal to the method of controlled experiments to uncover causal relations. Economists are not in a position to subject the economic system to controlled experiments and must resort to statistical methods to analyse aggregate data. Yet, statistical analysis is inadequate for causal inference from sample data. There are three lines of arguments in the literature for this inadequacy of statistical methods, and hence the necessity of economic theory in macroeconomics. A brief study of these arguments sheds light on the reasons behind the emergence of competing approaches to macroeconomics.

3.1 Statistical Control

A major argument for the necessity of theory in macroeconomics is based on the inadequacy of the regression method for causal inference (henceforth, RMCI). The method of regression stands at the heart of econometrics and many controversies in macroeconomics relate to this method. It is, therefore, worthwhile explaining in some detail how the method is used for causal inference and why it fails in establishing causation. There are many discussions of the method as well as its limitations. We will draw on Simon (1954), Clogg and Haritou (1997), Spirtes, Glymour and Scheines (1998), Pearl (2000), and Spirtes (2000) to describe the method, and explain why it fails.

We first concentrate on the simple regression equation (2.1), and then extend the analysis to cases where there are several regressors involved. Since in the following the first moment of the variables is of no interest, we assume that the variables are measured around their mean, and drop the intercept from the equation. Equation (2.1) then becomes:

$$Y = \beta X + \varepsilon \tag{3.1}$$

Regression analysis is concerned with estimating the parameter β, and the conditions under which an unbiased, efficient (minimum variance) and consistent estimate can be obtained from the data. To use this as a method of causal inference, one has to explain the conditions under which such an estimate of β can be taken as an estimate of the effect of *X* on *Y*, as well as how the conditions can be established in practice. Accordingly, thee are three issues to address for a full view of the possible role of regression in causal inference. The first concerns the conditions under which an unbiased, efficient and consistent estimate of β can be obtained from the data. The second concerns the conditions under which the estimate can be taken as an estimate of the effect of *X* on *Y*. Finally, the third concerns the possibility of establishing these conditions in practice. We will review the answers given to these questions by econometricians, and then explain why the regression method is unable to establish causation.

To estimate β, users of the RMCI turn to the theory of ordinary least squares. This theory makes a number of assumptions about the error term ε to ensure an efficient, unbiased and consistent estimation. To begin with, it assumes that the expected value of ε_i conditional on observation X_i is zero; that is, $E(\varepsilon / x_i) = 0$. This implies that the unconditional mean $E(\varepsilon)$ is zero. Likewise, the same condition implies that ε_i and X_i are uncorrelated; namely, $Cov(x_i, \varepsilon_i) = 0$. This last implication is known as the *orthogonality* condition. Given the linearity of (3.1), the orthogonality condition ensures that a least squares estimate of β is unbiased. The theory also requires that observations on X provide no information about the variance and covariance of the error term ε. This means that the errors associated with the observations must have constant variance σ^2 and be uncorrelated with each other. Under these conditions, a least-squares estimator of β is shown to be efficient, unbiased and consistent.

Econometricians and social scientists add one or two requirements to the orthogonality condition to identify an unbiased estimate of β with the effect of X on Y. Herbert Simon, in his celebrated article (1954), requires X to precede Y. By this, he intends to rule out bi-directional causation between X and Y. Others also require that X can indeed be a causal variable so as to exclude nonsense inferences like inferring that having nicotine stains on one's finger causes lung cancer. Therefore, according to the users of the RMCI, a least squares estimate of the coefficient of X coincides with the effect of X on Y if X is uncorrelated with ε, X precedes Y, and X can indeed be a causal variable. The validity of this answer, Simon maintains (1954), can easily be shown in the context of the simple regression equation (3.1). Suppose β in (3.1) represents the effect of X on Y. If we multiply the equation through by X and take expectations of both sides, we will have

$$Cov(X,Y) = \beta V(X) + Cov(X,\varepsilon), \tag{3.2}$$

where $Cov(X,Y)$ is the covariance of X and Y, $V(X)$ is the variance of X, and $Cov(X,\varepsilon)$ covariance of X and ε. If X and ε are uncorrelated, the least squares estimate $\hat{\beta}_{XY}$ will be equal $Cov(X,Y)/V(X)$, which is the same as β, the effect of X on Y. That is,

$$\hat{\beta}_{XY} = Cov(X,Y)/V(X) = \beta$$

However, if the condition fails, $\hat{\beta}_{XY}$ and β will no longer be the same. Implicit in this analysis is that every correlation has a causal explanation, in the sense that it arises either because of a direct causal connection or because of unmeasured common causes. So, if the existence of unmeasured common causes is ruled out by assuming the orthogonality condition, a correlation between X and Y reveals the presence of a direct causal connection. Evidently, if the world contained spurious correlations, which could not be explained by reference to latent common causes, the orthogonality condition would not justify inferring from a correlation between X and Y that either X causes Y or Y causes X. Such a conclusion would demand first ruling out all possible non-causal explanations.[16]

This brief description of the RMCI shows how the regression method is used to establish causation. Given the conditions, one simply regresses Y on X. If the least squares estimate $\hat{\beta}_{XY}$ differs from zero, X is said to cause Y, and if it is zero, X is said not to cause Y. The success of this method depends, on the one hand, on the adequacy of the conditions and, on the other, on the possibility of establishing them in practice. The first topic, that is, the adequacy of the conditions, falls outside the scope of this chapter.[17] Instead, we study the second issue. All the three conditions demand a careful analysis. But, to keep the discussion short, we confine ourselves to an examination of the orthogonality condition, as this will suffice to explain why the RMCI fails.

The RMCI comes with a method for establishing the orthogonality condition. To explain the method, it is vital to note that this condition differs from other familiar statistical assumptions underlying a regression model, such as the linearity of the function linking X and Y or the normality of the distribution of Y. The validity of these assumptions can be checked by using observations on X and Y. In fact, for arbitrarily large samples, there are statistical algorithms that discover the functional

[16] See N. Cartwright (1989) for a full discussion.

[17] Woodward (1999) and Pearl (2000) have argued that the orthogonality condition is neither necessary nor sufficient for causal interpretation of a regression equation.

form of the relation between X and Y, and estimate the correct density function of Y. In contrast, observations on X and Y contain no information on the validity of the orthogonality condition. This follows from the fact that the true disturbances ε_i associated with observations (x_i, y_i) are never known. In practice, we can only estimate the residuals $e_i = (y_i - \hat{y}_i)$, with $\hat{y}_i$ being the predicted value of y_i. However, if we use, for example, the least squares method to estimate β, the residuals e_i are automatically uncorrelated with x_i. Thus, one cannot uses the residuals to establish the condition (Clogg, et al., 1997:94).[18] In this sense, the condition is not a statistical assumption.

Faced with this limitation, econometricians have tried to establish, or at least support, the validity of the orthogonality condition by bringing in variables other than X and Y. To understand the philosophy behind this attempt, one should note that the error term ε in equation (3.1), when taken as a structural relation, stands for the effects of omitted variables on Y. Any correlation between X and ε is, therefore, said to indicate the presence of latent common causes for X and Y. Such variables are referred to as confounders.[19] This interpretation suggests that the correlation between X and ε can be eliminated by including all the confounders of X and Y in the regression of Y on X. In that case, the error term ε will be uncorrelated with X, and if other conditions are in place, an estimate of β will coincide with the effect of X on Y. It has thus been suggested that the orthogonality condition can be established by searching for all the confounders of X and Y, and including them in the regression of Y on X. To estimate the effect of X on Y, it is not enough to estimate the simple regression equation (3.1). Instead, it is necessary to regress Y on X and all the confounders of X and Y. An estimate of the regression coefficient of X in this

[18] The ordinary least squares regression coefficient of X is given by $E(YX)/E(XX)$. If we define β as equal to $E(YX)/E(XX)$, we have

$$\varepsilon = Y - \beta X,$$

$$X\varepsilon = XY - \beta(XX),$$

$$E(X\varepsilon) = E(YX) - \beta E(XX) = 0.$$

[19] In other words, a confounder of X and Y is a variable that is a cause of both X and Y.

equation corresponds with the effect of X on Y. The process of regressing on confounders is often called *conditioning* or *statistical control.*

The reasoning behind this claim can be illustrated by considering the case, where there is only one confounder Z for X and Y.[20] Suppose the process generating Y can be described by model (3.3):

$$X = \alpha Z + \varepsilon_1 \tag{3.3a}$$
$$Y = \beta X + \gamma Z + \varepsilon_2 \tag{3.3b}$$

where $Cov(\varepsilon_1, \varepsilon_2) = 0$, and α, β, γ are different from zero. Z in this model is a common cause of X and Y. If we estimate (3.1) in place of equation (3.3b), X and ε will be correlated, and a least squares estimate of the coefficient of X will differ from β. To see this, we simply need to multiply (3.3b) through by X and take expectations of both sides to get

$$Cov(X,Y) = \beta V(X) + \alpha\gamma V(Z).$$

We then have

$$\hat{\beta}_{XY} = \frac{Cov(X,Y)}{V(X)} = \frac{\beta V(X) + \alpha\gamma V(Z)}{V(X)} \neq \beta.$$

However, if Z is included in the regression of Y on X, the orthogonality condition is satisfied and the least squares estimate of β can be equated with the effect of X on Y, as shown below:

$$\hat{\beta}_{XY/z} = \frac{Cov(X,Y/Z)}{V(X/Z)} = \frac{Cov(X,Y)V(Z) - Cov(X,Z)Cov(Y,Z)}{V(X)V(Z) - Cov(X,Z)^2}$$

[20] This example is borrowed with some changes from Spirtes et al. (1998).

$$= \frac{\beta(V(X) - \alpha^2 V(Z))}{V(X) - \alpha^2 V(Z)} = \beta$$

The example illustrates that regression on a confounder eliminates bias; it turns an otherwise biased estimate into an unbiased one.

The problem with this reasoning, of course, is that the set of confounders of X and Y is not known. In practice, statisticians inevitably replace the set of confounders of X and Y with a set of *potential* confounders, namely, a set of measured variables that precede X and Y and can possibly affect them. It is held that by controlling for potential confounders, one is likely to control for the real confounders, and eliminate possible correlation between X and ε (Black, 1982:31). Thus, one is advised to control for as many potential confounders as one can to achieve a reliable estimate of the effect of X on Y. The longer the list of potential confounders included in the regression of Y on X, the more reliable is said to be the estimate:

> One must include in the equation fitted to data every 'optional' concomitant [potential confounder] that might reasonably be suspected of either affecting or merely preceding Y given X – or if the available degrees of freedom do not permit this, then in at least one of several equations fitted to the data (Pratt and Schelifer, 1988:44).[21]

In this way, multivariate regression has come to dominate macroeconomics. To estimate the effect of X on Y, Y is regressed on X and a few other variables thought likely to affect both X and Y. The estimate of β in the equation with all the potential confounders, whose inclusion affects the estimate of β, is taken to represent the effect of X on Y. The RMCI can also be generalised to regression equations with multiple regressors. For a causal interpretation of a multivariate regression equation, all the regressors are required to precede the response variable Y and to be uncorrelated with the error term. Similarly, to establish the orthogonality condition,

[21] The phrase inside the bracket is added.

one has to control for all the confounders of the regressors and the response variable (Clogg et al., 1997:94).

3.1.1 Limitations of Statistical Control

Critiques have questioned the adequacy of the RMCI from a variety of perspectives. Most of these criticisms relate to the limitations of statistical control in practice. One limitation arises from the small number of variables measured in practice. To state the critique precisely, let $\boldsymbol{C}$ be the complete set of potential confounders for X and Y. The plausible idea of statistical control is that if we could control for all the variables in $\boldsymbol{C}$, we would be able to control for all the real confounders of X and Y, and estimate the effect of X on Y. But the set $\boldsymbol{C}$ is never completely known. What one measures in practice is a *proper* subset of $\boldsymbol{C}$, which may exclude some or even all of the actual confounders of X and Y. As a result, conditioning on measured confounders can never guarantee the truth of the orthogonality condition, and a non-zero estimate of β can always be due to latent common causes. The RMCI on its own fails to distinguish between cases of genuine causal connection and spurious correlation (Pearl, 2000:186).

Another problem is that conditioning on a measured variable, which is not a confounder but is taken as a potential confounder, can turn a consistent estimate of the effect of X on Y into an inconsistent estimate. This occurs whenever one controls for a *barren proxy*; that is, a variable Z that is correlated with factors that influence X and Y but itself has no effect on X and Y. As an illustration, consider the following example studied by Pearl (2000), Spirtes et al. (1998), and Spirtes (1997). Suppose that our set of measured variables consists of $\{X, Y, Z\}$, X precedes Y, and that Z

precedes both X and Y. Also, suppose that the causal structure of these variables is given by the model below (Figure 1),

$$
\begin{aligned}
X &= U_1 + \varepsilon_x \\
Z &= \alpha U_1 + U_2 + \varepsilon_z \\
Y &= \beta X + \gamma U_2 + \varepsilon_y
\end{aligned}
$$

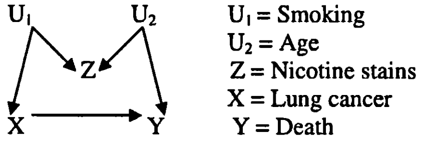


Figure 1

where ε_x, ε_z and ε_y are independent error terms, U_1 is an unmeasured common cause of X and Z, and U_2 is an unmeasured common cause of Y and Z. Further, suppose that the unmeasured variables are uncorrelated with the error terms. In this setting, if Y is regressed on X alone, the least squares estimate of β is consistent. However, if Y is regressed on both X and Z, the estimate of β is no longer consistent, and normally differs from the effect of X on Y. This can be seen from the least squares estimate of β in the regression equation of Y on X and Z. To this end, we first note that $Cov(X,Z) = \alpha V(U_1)$ and $Cov(Y,Z) = \beta(\alpha V(U_1)) + \gamma V(U_2)$. Let $Cov(X,Z) = \rho$ and $\gamma V(U_2) = \tau$. Then we have

$$
\begin{aligned}
\hat{\beta}_{XY/Z} &= \frac{Cov(X,Y/Z)}{V(X/Z)} = \frac{Cov(X,Y)V(Z) - Cov(X,Z)Cov(Y,Z)}{V(X)V(Z) - Cov(X,Z)^2} \\
&= \frac{\beta V(X)V(Z) - \rho(\rho\beta + \tau)}{V(X)V(Z) - \rho^2} = \beta - \frac{\rho\tau}{V(X)V(Z) - \rho^2}
\end{aligned}
$$

which generally differs from β. The estimate is consistent only if either ρ or τ is zero. Otherwise, β might be zero but the least squares estimate $\hat{\beta}_{XY/Z}$ significantly different from zero. "[T]here is no sense in which one is 'playing safe' by including rather than excluding 'potential confounders' in the conditioning set; conditioning on these variables could change a consistent estimate into an inconsistent estimate" (Spirtes, 1997:7). One cannot simply condition on any measured variable that precedes X and Y. Before conditioning, it must be ensured that the variable is not a barren proxy and doing this obviously necessitates some knowledge about the causal relation between the measured and unmeasured variables affecting *X* and *Y*. Such knowledge cannot be obtained from statistical analysis of data on the measured variables.

Another related point is that a potential confounder must itself satisfy the orthogonality condition (Clogg, et al., 1997:98). To be precise, to control for a potential confounder *Z* in estimating the effect of *X* on *Y*, *Z* must also be uncorrelated with the error term. Since this requirement cannot be assumed *a priori*, one needs to bring in new variables to ensure that *Z* and ε are uncorrelated. This, of course, requires making new orthogonality assumptions. It is thus never possible to establish the orthogonality condition by controlling for measured potential confounders (Freedman, 1987:307). To terminate the regression, one must rely on substantive domain specific information. This necessity of subject-matter information in establishing causation is viewed as a key reason for the essential role of theory in macroeconomics.

3.2 The Identification Problem

A second argument for the necessity of theory in macroeconomics relates to the conditions under which the parameters of a model are fully determined by the joint probability distribution of the observables in the model – the so-called identification problem. Historically, a common belief in economics has been that the values of important economic variables such as demand and supply for a good are simultaneously determined, and for that reason the economy is said to be best represented by a simultaneous equations model. Because of feedback, the error terms across the equations in a simultaneous equation model are not usually uncorrelated. Consequently, applying the ordinary least squares method to the model does not yield consistent estimates of the parameters. To achieve consistent estimates, the model must first be transformed into a system of regression equations or, in other words, a *reduced form* model, where the errors across the equations are independent. In this context, the identification problem is the problem of inferring the parameters of the underlying simultaneous equations model (structural model) from the parameters of the regression model (Manski, 1995). However, it is usually the case that a large, and often infinite, set of parameter values of the structural model is consistent with the parameters of the reduced form model, making it impossible to infer the parameters of the structural model from those of the reduced form model.[22] The identification problem thus has no purely statistical solution.

In the context of simultaneous equations models, the identification problem can be resolved by imposing restrictions on the variables that enter into each equation. In a linear structural model, if one can exclude from each equation one variable that enters into other equations, none of the model equations can be written as a linear combination of the others, and the model becomes identifiable (Koopmans, 1949 (1971):169). One arguably needs to rely on non-sample (domain-specific) information to decide which variable to exclude from or include into an equation. Likewise, in recursive models, identifiability calls for the orthogonality condition and the independence of the disturbance terms across the equations (Boudon, 1968:208; Blalock, 1968:167).[23] These conditions, as argued earlier, are not statistical assumptions; they can be justified only by means of domain-specific information, another reason for the essential role of theory in modelling the economy.

It should be noted that the identification problem is different from the causal inference problem, as there could be more than one identifiable causal model consistent with a data set. Recursive models are always identifiable if the disturbance terms satisfy the orthogonality condition and are independent across the model equations. But there are usually many identifiable recursive causal models consistent with a data set. This lack of uniqueness leads to a quandary regarding which is the true causal model. A solution to the identification problem is not, therefore, a solution to the causal inference problem.

[22] For an example, see Koopmans (1949 [1971]:169).

[23] A thorough analysis of the identification problem is given in Manski (1995).

3.3 The Lucas Argument

The last two arguments reveal some of the limitations of statistical analysis for causal inference. The economic literature also provides a third important argument for the necessity of theory, which is based on the inadequacy of knowledge of the *existing* structure for policy analysis. Various statements of this argument are found in the writings of early econometricians, including Haavelmo (1944), Koopmans (1947), and Hurwicz (1962). However, it was Lucas who most forcefully stated the argument in his critique of econometric policy evaluation (1976), and supported it by means of various graphic examples. Lucas' critique is primarily directed at what he calls the conventional theory of policy evaluation built around a model of the economy essentially similar to the one described earlier. Recall the economy had a single endogenous variable Y_t, whose law of motion was given by the difference equation

$$Y_{t+1} = f(Y_t, X_t, \theta, \varepsilon_t), \tag{3.5a}$$

and the rule (law) governing the policy (exogenous) variable X_t by

$$X_t = g(Z_t, \lambda, e_t). \tag{3.5b}$$

In this setting, Lucas interprets a policy change as a change in the rule governing the policy variable X_t, which involves setting the parameters λ at values different than their actual values. Thus, in evaluating policies, the analyst contemplates a set of values for λ different than the actual values, derives a sequence of values for X_t using the new rule, and recursively inserts them into the fitted model (3.5a) to predict future values of Y_{t+1}. The exercise is carried out for different values of λ to select a rule that produces the desirable result.

This practice, Lucas argues, is fundamentally flawed, as it assumes that the economic structure $(\mathbf{f}, \theta)$ *prior* to the policy change (call it the *old* structure) and the structure *afterwards* (call it the *new* structure) are the same. The structure, however, emerges from and is sensitive to the decision rules (supply and demand functions) of the agents. As we change a policy regime, we change the environment in which the agents operate, altering the constraints restricting their choice behaviour. The agents recognise the change and modify their decision rules. This invalidates the structural model fitted to the data collected prior to the intervention, rendering it useless for predicting the course of the economy under the new policy regime. And so, statistical analysis alone, Lucas argues, cannot provide the information requisite for policy evaluation, because econometric analysis can at most offer knowledge of the prevailing structure from which the data have been sampled. Evaluating non-trivial policies, however, calls for knowledge of the new structure emerging from the policy change, for which no data are yet available, and hence whose discovery falls outside the domain of statistics.

To elaborate on what is involved in inferring the new structure, let f^i_{old} stand for a structural relation true of the economy prior to the policy change and f^i_{new} stand for the relation true after the change. Each new relation f^i_{new} theoretically depends on the old relation f^i_{old} and the policy; i.e.,

$$f^i_{new} = \phi_i(f^i_{old}, Policy). \tag{3.6}$$

Hurwicz (1962) calls these mappings modification function. Thus, in addition to the knowledge of the old structure, inferring the new structure requires knowledge of the *modification* functions ϕ_i, which map each old relation f^i_{old} into a new relation f^i_{new}. These functions cannot be inferred by statistical analysis of the data generated by the old structure, and one needs a theory to explain how people react to the policy and how that affect the structural relations currently true of the economy. To put it somewhat differently, Lucas' argument points to another important aspect of the inadequacy of statistical analysis for causal inference. It states that for evaluating actions and policies not only it is essential to know whether a relationship of interest is causal but it is also essential to know the conditions under which it remains invariant. Such knowledge requires more than analysis of the data generated by the current structure. It demands understanding the chance-setup leading to the relation. The issues of whether a relation is causal and the conditions under which it remains invariant are distinct matters (see also Cartwright, 1995).

The Lucas critique has been criticised on several grounds. Notably, it has been argued that the critique only applies to interventions involving a change in a policy rule but such changes are very rare (Sims, 1982). Or it has been said that people are slow in absorbing the effects of policy changes and, therefore, in practice, statistical models closely approximating the prevailing structure give reliable short-run forecasts of the policy outcomes. These criticisms, even if true, do not weaken the logical force of Lucas' argument. The point is that if a policy change could shift the structure, for predicting the policy outcomes, one would first have to predict the emerging structure, which cannot be achieved by statistical analysis of the data sampled from the structure prior to the change.

This section has studied three arguments for the necessity of subject matter information in modelling the structure. A central question in macroeconomics is whether there is an alternative procedure for acquiring the theoretical information necessary for modelling the structure. Reflection on this question has led to several rival approaches to macroeconomics. The remainder of this chapter reconstructs two competing approaches, which span the spectrum of views on the scope and nature of macroeconomics.

4 The Theoretical Approach

The first approach is theoretical macroeconomics. The general idea of this approach is present in the writings of the early members of the Cowles Commission, including Koopmans (1947b) and Marshack (1953), as well as other early economists such as Jevons (1871) and Hicks (1939). A rigorous and systematic statement of the approach, however, emerged as a result of new classical economists' reflection on the failure of Keynesian macroeconomic models in the 1970s (Lucas and Sargent, 1979). To put it quite generally, according to these economists, economics can acquire the theoretical information necessary for modelling the structure by adopting a bottom-up approach to the study of large-scale economic phenomena. This involves establishing a theory of microeconomic behaviour and transforming it into a theory of the economy using aggregation methods. The theoretical approach rests on two basic assumptions regarding human behaviour and the relation between the micro and macro levels in the economy.

The first assumption is that we have, or it is possible to establish, a satisfactory theory of individual behaviour. Early in the history of modern economics, economists thought that *intuition*, *introspection*, and *interview* were reliable means of understanding behaviour. Tjalling Koopmans held that through these means it would be possible to establish the motives of consumers, firms and investors, and hence understand how they make decisions. The information, he added, could eventually be turned into a theory of economic behaviour as precise as the laws of motion of material bodies known to Kepler (Koopmans, 1947b:166).[24] In recent years, more emphasis has been placed on experimentation. It has been suggested that neither the ethical prohibitions nor the unbearable costs involved in experimenting with the economy are encountered in studying individual behaviour. Therefore, even if current theories of behaviour are imprecise, with adequate research it should be possible to establish an adequate theory of behaviour (Lucas; 1980:288-90).

The second assumption is that the laws of the society are the same as the laws of the individual, and hence a theory of the economy as a whole can be inferred from a theory of individual behaviour. In this regard, theoretical economists follow John Stuart Mill, who wrote:

> "Human beings in the society have no properties but those which derived from, and may be resolved into the laws of the nature of the individual man. In social

[24] A similar view regarding the obviousness of the laws of economic behaviour is explicit in Mill's Principles of Political Economy, where he writes "Happily, there is nothing in the laws of Value which remains for the present writer or any future writer to clear up; the theory of the subject is complete" (1848 [1990]:420).

> phenomena the Composition of Causes is the universal law"(John Stuart Mill (1974 [1874]): 879)

There are at least two interpretations of this doctrine in the literature. In the early days of economics, Jevons (1965 [1871]:16) and Hicks (1939:245) held that the general form of the laws of economics is *the same* in the case of a single decision maker and a nation, and thus the laws of the economic system can be derived from the laws of a single decision making unit, be it a household or a firm. Another interpretation of the doctrine, on the other hand, emphasizes the importance of competition. It says what characterizes an economy is competition over scarce resources, and to understand the laws of the economy it is vital to understand how individuals compete against each other. The interpretation, therefore, identifies the laws of the economy with the laws of a group of competitive agents, not with the laws of a single individual (Lucas, 1981:289).

The implications of these assumptions for modelling the economy are clear. By observation and experimentation, the economist can establish a theory of individual behaviour, replace the variables in the micro theory with their corresponding aggregate variables to obtain a qualitative model of the economy, and use aggregate data to estimate the model, hence transforming it into a quantitative model. Since the model is derived from the laws governing the basic decision making units of the economy, it correctly represents the structure. Accordingly, the main methodological objective of modern theoretical economics, at least as understood by the new classicals, has been to incorporate aggregative problems into the framework of microeconomics, eliminate the distinction between microeconomic and macroeconomic theory, and speak of economic theory in general. Robert Lucas vividly states and defends this claim in the following passage:

> "The most interesting developments in macroeconomic theory seem to me describable as the reincorporation of aggregative problems such as inflation and the business cycle within the general framework of 'microeconomic' theory. If these developments succeed, the term 'macroeconomic' will simply disappear from use and the modifier 'micro' will be superfluous. We will simply speak, as did Smith, Ricardo, Marshall and Walras, of economic theory." (Lucas 1987:107-8)

The project of inferring the patterns emerging at the economy level from a theory of individual or group behaviour is known as the *microfoundations* project. The project is claimed to enable the economist to establish a reliable theory of the economy without having to subject it to costly and prohibitive experiments:

> Suppose that we have some ability to predict how individual behaviour will respond to specified changes. How, if at all, can such knowledge be translated into knowledge of the way an entire *society* is likely to react to changes in its environment? ... We clearly need to know something about the way a group of monkeys interacts, in addition to their individual preferences, in order to have any hope of progress on this complicated question.... The ingredient omitted so far is, of course, competition... Notice that, having specified the rules by which interaction occurs in detail, and in a way that introduces no free parameters, the ability to predict individual behaviour is *nonexperimentally* transformed into the ability to predict group behaviour. ... This is exactly why we care about the "microeconomic foundations" of aggregate theories (Lucas, 1981:289-91).[25]

The derived theory is believed to specify variables relevant for describing the economy, draw a line between endogenous and exogenous variables, determine the sign of relevant regression coefficients, and impose constraints on the form of the functions relating the aggregates. This information will be adequate for modelling the structure, and all the conventional goals of macroeconomics are thus claimed to be attainable. Most importantly, the microfoundations project is said to make it possible to predict the outcomes of policies that could shift the structure. One begins by analysing how a proposed policy might affect the way in which basic decision-making units of the economy interact with each other and make decisions. Knowing this, one will be able to infer through aggregation the impact of the policy on the economy as a whole, and derive the *new* structure that will prevail after the policy change. Since the same can be carried out for any policy, one will be able to help the state officials to select an optimal policy (Lucas and Sargent, 1979). Theoretical economics confines the role of statistical methods to estimating and testing of economic theories, and leaves no role for the regression method in causal inference. If a theoretical model fails to accord with the data, the road to progress is held to lie in searching for better theories, not in sophistication of statistical procedures or in collection of more aggregate data.

[25] Italics are added. Also see the same article, footnote 11.

The call for microfoundations is characteristic of all schools of theoretical economics. It is, nevertheless, in the new classical school that the search for microfoundations has most systematically been pursued. In this school, one definition of microeconomic theory takes the basic unit of economic analysis to be a single decision maker, either a consumer (household) or a producer (firm). The consumer is modelled as an expected utility maximizer whereas the firm as an expected profit maximizer. Moreover, when there is uncertainty, the individual is assumed to act according to the true probabilistic model of the economy. From this perspective, a call for microfoundations is a call for a model of the economy in which the starting point is an expected utility or profit maximisation problem. To model the relation between aggregate variables of interest, such as aggregate income and consumption, a utility maximization problem for a single consumer is set up and solved subject to his budget constraint. The solution defines the relation between the relevant micro variables, say, individual income and consumption. The same relation is hypothesized to be true at the aggregate level, and the corresponding aggregate variables are inserted into the model to derive a model of the economy. Aggregate data are then used to estimate the model. This approach usually goes by the name of the "representative agent" or "per capita" modelling approach.

Another definition of microeconomics in the new classical school stresses competition. In a competitive environment, the outcomes of an agent's decision depends on the actions of others in the economy, which means agents must form expectations about the actions of others and, indeed, expectations about the expectations of others, and so forth. This feature of the economy is believed to be best captured by assuming equilibrium (Chari, 1999:3). Therefore, new classical economists have mainly equated microeconomic theory with the theory of Walrasian general equilibrium, or its successor the Arrow-Debreu competitive equilibrium theory, supplemented with the rational expectations hypothesis. In light of this interpretation, the laws of the economy are identified with the laws derived from the general equilibrium theory joined with the rational expectations hypothesis (Howitt, 1986:273). A model is called structural if it is built on an appropriate microeconomic theory. Models that lack microeconomic foundations are viewed as non-structural (Sims, 1991:923, footnote; 1982:115-6).

5 Atheoretical Macroeconomics

The view of theoretical economics given above characterises one extreme side of the spectrum of competing views on the nature and scope of macroeconomics. The view stands on two types of assumptions. Metaphysically, it assumes that the microeconomic structure leads to a unique set of stable relations among economic aggregates, suitable for a causal account. And methodologically, it assumes that an accurate theory of individual behaviour can be established and no substantial difficulties arise in transforming it into a theory of the economy. We now present an alternative view that stands on the other extreme side of the spectrum of opinions about macroeconomics.[26] The approach is due to Christopher Sims, who, in one way or another, challenges all the assumptions underlying theoretical economics. Following Cooley and LeRoy (1985), we refer to this approach as *atheoretical* macroeconomics.

Sims' atheoretical approach also emerged in response to a general discontent with the performance of macroeconomic models during the 1970s and 1980s. Most economists of the time, including Sims, blamed the failure on the identifying restrictions underpinning the models, which were supposedly derived from economic theory. Sims termed the restrictions as 'incredible' (Sims, 1980:1; 1982a:108). However, contrary to the theoretical economists, he did not think that the key to improving the state of macroeconomics was to search for better theories. In his view, the problem with macroeconomics was more profound. Consequently, he called for a far-reaching revision of the field and its objectives. Sims' revision is open to more than one interpretation. Two possible interpretations will be discussed here, one methodological and the other metaphysical.

[26] For a history of atheoretical macroeconomics see Simkins (1999).

Our accounts of atheoretical macroeconomics differ from a dominant interpretation of the approach criticised in a well-known paper by Cooley and LeRoy (1985). According to these authors, Sims altogether dispenses with the role of economic theory or in general domain specific information, and seeks to achieve the traditional objectives of macroeconomics by means of statistical analysis of aggregate data alone.[27] A reason put forward for this reading is the use of Granger's test of causality by Sims and his followers, which is nothing but a statistical procedure for determining whether a variable helps predict another variable. Another reason is the claim by Sims that atheoretical models are useful for policy analysis. Since a model must be structural to be useful for policy evaluation, any claim for usefulness of atheoretical models for policy analysis presumes a structural interpretation of the models. Both reasons can be challenged. To begin with, Sims rejects that the Granger test of causality alone can ever establish causality (1977:29 and 42; 1986:3). In his view, it is always necessary to rely on non-sample information to conclude that a relation that passes the test is actually structural. Moreover, according to Sims, atheoretical models, as long as they remain uninterrupted, are of no use in policy analysis (1986:3); Sims simply challenges the claim that the interpretation derives from a well-founded economic theory (1982a:138). I shall rely on Sims' own writings as well as Cooley and LeRoy's paper (1985), Leamer (1985), Pagan (1987), and Swanson and Granger (1997) to give a brief review of formal aspects of Sims' modelling approach.

5.1 Methodological Interpretation

On this interpretation, Sims agrees with the theoretical economists that the relations discernable at the economy level are in principle suitable for a structural account but challenges the existence of a reliable method for discovering the structure. Sims argues that economic theories are bound to remain imprecise due to the lack of controlled experiments and the fact that the economic structure is non-stationary. The structure continuously shifts through natural, social and political changes, and critically through accumulation of experience by people. As people learn about the

[27] Cartwright (1989), Epstein (1989) and Leamer (1985) suggest a similar interpretation.

economy and discover the outcome of their actions, they modify their behaviour, and this shifts the structure. As a consequence, a theory approximately true of a situation might no longer be true of a new situation, making it difficult to tell whether the failure of economic theories is due to changes in the structure or to mistakes in theorising about the structure:

> ... dynamic economic theories must inherently be incomplete, imprecise, and therefore subject to variation over time. One reason for this is that economic cause-effect relations involve a "recognition delay" about which theory has little to say and may be expected to be variable.... It is wrong, then, to expect economic theories to be complete, mechanical, and divorced from reference to specific historical circumstances (Sims, 1981:579).[28]

Sims argues that this inherent imprecision renders economic theory personal and subjective (2004:282). And, as a consequence, uninterpreted statistical models of aggregate data are the only yardsticks of objectivity in macroeconomics, forming a basis around which economists may come to narrow down their differences (Sims, 1987:53). These models are not, however, suitable for policy analysis, which requires classifying the variables into exogenous and endogenous categories, and deciding whether a variable can be influenced by a policy. In making such decisions, the analyst must inevitably rely on his personal view of the economy. Two economists, with different views of the economy, can arrive at conflicting interpretations of a single model of the data, and there is no objective ground to decisively resolve the disagreement.

Sims discerns three general stages in modelling aggregate data. The first stage is to build a model that best fits the data, which gives one possible characterisation of the economic structure. The second stage is to search for alternative models equally fitting the data, which provide different views of the structure that might have generated the data. In the final step, the analyst relies on his personal view of the economy to select one of the models that, in his view, is most likely to approximate the structure. An appropriate modelling approach, Sims says, must distinguish between those aspects of a model that are based on the analysis of data, and those based on subjective judgements about the economic structure (1982b:317; 1987:51).

[28] Similar remarks are found in Sims (1996:113).

Such a distinction, he argues, saves economics from the Lucas critique (1976) and that of Freedman (1981), as these critiques are directed at the subjective features of large-scale economic models (Sims, 1982b:317).

5.1.1 Vector Autoregression

Sims therefore abandons the framework laid down in the Cowles Commission that requires a theory to specify relevant variables, divide them into exogenous and endogenous variables, and determine the variables entering in each equation in the model. As an alternative, he puts forward a framework in which there is initially no division of the variables into exogenous and endogenous categories and every variable enters into the equation of every other variable (Hendry, 1993:128).[29] The modeller relies on his view of the economy to select relevant variables and then uses the data as well as subjective and pragmatic considerations to select a model. To describe Sims's formal approach, we adopt a study from Swanson and Granger (1997), which models the behaviour of four aggregate variables consisting of money M_t, consumption C_t, investment I_t, and gross domestic product Y_t. Let $\mathbf{Y}_t$ be the vector of current variables (M_t, C_t, I_t, Y_t) and $\mathbf{Y}_{t-i}$ the vector of lagged variables $(M_{t-i}, C_{t-i}, I_{t-i}, Y_{t-i})$. Theoretically, the point of departure in Sims's approach is a structural model of the following form:

$$B\mathbf{Y}_t + \sum_{i=1}^{p} \Gamma_i \mathbf{Y}_{t-i} = \varepsilon_t \tag{5.1}$$

where B and Γ_i's are 4×4 matrices whose terms are polynomial in the lag operator, p is the lag length, and ε_t is a column vector of stochastic error processes with elements ε_{it}. The matrices have no zero element and all the variables are of identical lags. Also, the model contains only current and lagged endogenous variables. This contrasts with a structural model of theoretical economics in which the theory dictates

[29] Before Sims, T.C. Liu (1960) had argued that in a macroeconomic setting no variable could be regarded as exogenous.

variables to be either endogenous or exogenous, and sets some elements of the coefficient matrices to zero.

In practice, Sims works with a vector autoregression (VAR) representation of (5.1), in which each current variable is regressed on its own past values and past values of other variables under study. The transformation into a VAR model leads to a model of the form

$$\mathbf{Y}_t = \sum_{i=1}^{p} A_i \mathbf{Y}_{t-i} + u_t \qquad E(u_t u_t') \equiv \sum \qquad (5.2)$$

where the A_i's are 4×4 matrices, u_t is a 4×1 column vector of stochastic error processes, Σ is the contemporaneous covariance matrix, and *E*(.) is the expectation operator. Every current variable in (5.2) is a function of two components: its best linear predictor based on past values of all the variables considered and its unpredictable error u_t, which is also called 'innovation' (Darnell, 1990:120). The innovation terms satisfy the orthogonality condition, and the least squares method can be used to estimate the parameters A_i.[30]

A VAR model can effectively captures patterns existing in the data and, so long as the mechanism generating the data remains the same, is useful for *ex ante* and *ex post* prediction. A VAR model, however, sweeps all the (exogenous) variables that can affect the contemporaneous variables under the blanket of the disturbance (innovation) terms and is only driven by random shocks. As a result, it is not suitable for policy analysis in the traditional sense which involves tracing out the effects on the endogenous variables of changes in the exogenous variables or the rules governing them. Sims and his followers inevitably redefine a policy as a random shock to a variable in the system, and interpret policy analysis as the task of tracing out the reaction of the system to that shock. But even in this narrow sense, a VAR model cannot be used for policy analysis. In general, the contemporaneous covariance

[30] This follows from the assumption that the present does not influence the past and the fact that all the variables on the right hand side of (5.2) are lagged except for u_t.

matrix $\sum$ is not diagonal. The non-zero off-diagonal elements entail that one variable, say Y_{it}, cannot be shocked through its corresponding error term, u_{it}, without having to simultaneously deliver correlated influence on other variables (Demiralp, et al., 2003:746). Without independence, it will not be possible to use the model to trace out the evolution of the system caused by a shock to a single variable. Sims and other VAR modellers advocate orthogonalizing the shocks using a Choleski decomposition to diagonalize the error covariance matrix $\sum$ by pre-multiplying (5.2) with the unique triangular matrix *T*. This generates a Wold causal chain among the current elements of $\mathbf{Y}_t$ vector:[31]

$$T\mathbf{Y}_t = T\sum_{i=1}^{n} A_i \mathbf{Y}_{t-i} + \eta_t, \qquad E(\eta_t \eta_t') = D \tag{5.3}$$

where $\eta_t = Tu_t$ and $D = T\Sigma T'$, a diagonal matrix. The errors η_t are termed as the *orthogonalized innovations* (Sims, 1987:52-3). A problem with this exercise is that the causal ordering is arbitrary, since for any ordering of the variables in model (5.2) there is a unique triangular matrix which orthogonalizes the covariance matrix of the errors. Generally speaking, if we have *k* endogenous variables in the model, we can order them in *k*! ways, resulting in *k*! different causal chain models equally fitting the data. These models describe alternative causal relations among the variables, and if no way can be found to select an ordering, any policy analysis based on a model like (5.3) will be arbitrary. A crucial matter facing the VAR methodology is how to transform a VAR model into a causal chain model in a non-arbitrary way.

In principle, Sims thinks "There is no unique way to do this" (1980:21). However, he suggests that some confidence in an ordering can be gained by checking the performance of the model against the data. In this line, if, for instance, we partition the data containing a shock to a variable into two parts, fit the model to one part, and the model closely approximates the impact of the shock in the other part, we gain

[31] A Wold causal chain is a system of equations in which the shock to Y_1 contemporaneously affect Y_2, $Y_2,\ldots,Y_n$ while the shock to Y_2 contemporaneously affect Y_3, Y_4, ..., Y_n but into Y_1 with a lag, and so on.

some confidence in the model. If a model is fitted to all the data, and there is no other data to check the performance of the model outside the sample period, the reliability of the model remains in doubt.

5.1.3 Selecting a Causal Chain Model

Since Sims' paper (1980), there have been several attempts to reduce the subjectivity involved in transforming a VAR model into a causal chain model. A proposal, which is in line with Sims' view, is found in Swanson and Granger (1997), who aim to devise a *data-driven* method for causally ordering the error terms.[32] These authors begin by estimating the VAR model (5.2) and using it to compute the residuals associated with the observations on the variables. The residuals form the data in their study of the causal relations among the errors. An assumption underlying Swanson and Granger's method is that the causal relations among the errors are recursive such that the error in the first equation in the appropriate model is exogenous and only affects the errors in the following equations (although generalisation to non-recursive models is possible in principle). Having said this, a possible ordering of the errors associated with model (5.2) is the following:

$$\begin{aligned}
m_t &= v_{Mt} \\
i_t &= \alpha m_t + v_{It} \\
c_t &= \gamma i_t + 0m_t + v_{Ct} \\
y_t &= \lambda c_t + 0i_t + 0m_t + v_{Yt}
\end{aligned} \tag{5.4}$$

where the lower case letters stand for the error terms; for instance, m_t stands for the error term in the equation for money M_t and so forth.[33] Swanson and Granger assume that the errors v_{it} in (5.4) have zero expectation, are contemporaneously uncorrelated, and have a non-singular covariance matrix. Given these conditions, they

[32] Swanson and Granger's work has been recently pursued and extended by Demiralp and Hoover (2003).

[33] In (5.4), each error, except for m_t, is a linear function of the innovation terms appearing earlier in the model and a stochastic component.

prove that a recursive model like (5.4) entails certain zero partial correlations (vanishing partials). In particular, if in the true recursive model m_t causes c_t and c_t causes i_t, the partial correlation of m_t and i_t given c_t is zero. If the partial correlation $\rho(m_t, i_t / c_t)$ is found to be zero or close to zero in the data, then the variable c_t in the appropriate causal ordering lies between m_t and i_t. Swanson and Granger exploit this and similar results to specify an ordering of the errors that is compatible with the data. Their method involves using the estimates of the residuals to compute the correlation matrix of the error terms, which is used to compute all possible partial correlations among the errors. The method then searches for a model that is compatible with the vanishing partials discerned in the data.

There are twelve partial correlations among the error terms associated with the variables under study here. In the data considered by Swanson and Granger $\rho(y_t, m_t / c_t)$, $\rho(y_t, m_t / i_t)$, and $\rho(i_t, m_t / c_t)$ are lowest in absolute value, and thus the most appropriate candidates for zero partial correlations. The first vanishing partial $\rho(y_t, m_t / c_t) \approx 0$ suggests that in the appropriate causal ordering c_t lies between y_t and m_t, the second $\rho(y_t, m_t / i_t) \approx 0$ suggests that i_t lies between y_t and m_t, and the third $\rho(i_t, m_t / c_t) \approx 0$ implies that c_t lies between i_t and m_t. Altogether, these vanishing partials suggest that a causal ordering as in the model below is compatible with the data:

$$\begin{aligned} m_t &= v_{Mt} \\ c_t &= \alpha m_t + v_{Ct} \\ i_t &= \gamma c_t + 0 m_t + v_{It} \\ y_t &= \lambda i_t + 0 c_t + 0 m_t + v_{Yt} \end{aligned} \qquad (5.5)$$

The zero partial correlations are not, however, compatible with the ordering expressed by model (5.4).

Swanson and Granger's method shares with other statistical approaches to causal inference a number of assumptions about the connection between probability and causation, which will be studied in chapter 5. Here, the relevant point to make is that the method fails to entirely eliminate the arbitrariness involved in transforming a VAR model into a causal chain model. As recognised by the authors, partial correlation is invariant to the reversal of causal directionality in the sense that it does not matter whether m_t causes c_t and c_t causes i_t or i_t causes c_t and c_t causes m_t. In either case, the partial correlation $\rho(i_t, m_t / c_t)$ is zero. Thus, besides model (5.5), a recursive model in which the causal influences proceed from y_t through i_t and c_t to m_t is also compatible with the vanishing partials in the data. In consequence, there is usually more than one causal ordering compatible with any set of vanishing partials found in the data, and one must draw on other considerations to select an ordering. In the present example, Swanson and Granger eventually favour the ordering in model (5.5) on the grounds of a conjecture that money, consumption, or investment is a leading indicator of GDP (1997:363).

Also, a correlation between two variables can be due to latent common causes. If the correlation between m_t and c_t given any possible combination of other variables under study is different from zero, it cannot still be concluded that either variable causes the other. The possibility of latent common causes enormously widens the class of models compatible with the vanishing partials, making it impossible for the present approach to distinguish between cases of causal and spurious relations. If no outside knowledge is available, the choice of a particular causal ordering of the innovation terms and hence the choice of a VAR model remains arbitrary. Empirical evidence alone is inadequate for specifying the privileged transformation that corresponds to the data generating structure.

5.1.4 Revising the Objectives

According to Sims, economists are never in a position to eliminate the need for personal judgement in selecting a model as a representation of the structure. Owing to the unreliability of personal judgments, Sims argues for revising the conventional objectives of macroeconomics (Sims, 1982:39-40). In particular, he urges economists to be sceptical about the analysis of policies that have no historical precedent. If a policy had a precedent in the data, and enough data were available, it would be possible to fit a model to part of the data, and use it to investigate how it performs in predicting the course of the economy in the rest of the data. If the model performed well in mimicking the effect of the policy, assuming that the structure of the economy was still the same, it would also most likely predict the outcomes of the policy in the new situation (Sims, 1982:122). However, if a policy had no historical precedent, the choice of a model for evaluating it would be entirely subjective. In that case, there would be no guarantee that the model would correctly predict the policy outcomes. The more a policy differs from those that have precedents in the data, the less reliable will be the analysis. Sims therefore questions the objective of evaluating novel policies, which he claims to fall outside the reliable domain of macroeconomics (Sims, 1982:119). For him, economists are observers of the economy, not engineers of reform (Lucas, 1987:8).

Equally, Sims is sceptical of the reliability of explanations in macroeconomics. In his opinion, "economists must accept that a single view of the causal structure of the record they examine will never emerge (Sims, 1977:30; 1981:583). Explanations of large-scale economic phenomena are consequently "stories" that the modellers can envision about what is going on inside their models (Sims, 2004:282). The choice of a story is based on personal considerations, and must be viewed with scepticism (Sims, 1981:583). Economists can be helpful in *ex ante* and *ex post* predictions over a short period of time. Analysis of radical polices and explanation of macroeconomic events, however, falls beyond the boundaries of their field (Sims, 1987:50).

Finally, Sims argues that the lack of controlled experiments and the inadequacy of statistical inference are not the only sources of the uncertainty of economic models. Aggregate economic data are also inherently inaccurate, and this fundamentally adds to the uncertainty of the models. This uncertainty casts doubt even on the choice of a model for *ex ante* or *ex post* prediction. And so, he advocates avoiding the choice of a single model and instead working with a group of models best fitting the data. If the task at hand is just *ex ante* or *ex post* prediction, it will be more reliable to average the predictions of all the models and act accordingly. In general, it is more reasonable both in prediction and policy analysis to compare the predictions of a number of plausible models and take a decision that is close to the predictions of all the models (Sims, 2004:282). Sims' view is consistent with a Bayesian approach to model selection in which the analyst expresses his uncertainty regarding the models in form of a probability distribution over the models, thereby avoiding acceptance of any of the models as the true model.[34]

5.2 Metaphysical Interpretation

The interpretation above of atheoretical macroeconomics assumes that it makes sense to speak of the causal structure of the economy, i.e., a web of structural relationships true of economic aggregates. Sims' early writings often suggest a more radical view that challenges the very existence of causal relations at the economy level. He time and again argues that economic variables can be aggregated in many different ways, and, all different levels of aggregation are theoretically arbitrary, and hence acceptable:

> Almost every kind of data used in economics... is an aggregate or index number of some sort. We deal with accounting data. Household budget studies divide expenditure into a finite number of categories with somewhat arbitrary bounds. Studies of firms use the firm's own books to construct measures of input, output, and prices. This is not just a matter of aggregation of fine-grained truth in which arbitrary accounting conventions would not be necessary. ... The degree of arbitrariness in classifying production into two-digit industries is not convincingly greater than that in classifying it into four-digit industries (Sims, 1987:50).

Sims, in addition, argues that, as the level of aggregation is varied, quite different and conflicting models of the system are achieved. And since there is no non-arbitrary or natural level of aggregation, it is wrong to attribute any causal interpretation to

[34] For a Bayesian perspective on inference from aggregate data see Leamer (1991).

aggregate models. In his view, there is no truth about price indices, national income accounts, or the money stock in the way there is truth about falling objects, electrical currents or the stars:

> The contribution of econometric probability models may be to make the process of economic data cheaper; more explicit, and more easily responsible. In doing so, it might also succeed in improving decision-making. But econometricians will not find truth the way physicists do. There is no truth about price indexes, national income accounts, expenditure of household *j* on meat, or the money stock the way there is truth about falling objects, electrical currents, or the stars (Sims, 1987:51).

The search for truth in macroeconomics is, therefore, misguided. An implication of this rejection of a causal structure at the economy level is that the tools of structural modelling are irrelevant to macroeconomic modelling. Large-scale economic models become black boxes useful for summarising data, and making short run *ex ante* and *ex post* predictions. They are not, however, suitable for the kind of policy analysis economists have traditionally been after. The emergence of a pattern at the aggregate level may have an explanation but the explanation is not causal. The pattern is explained by showing how it emerges from an attempt to summarise the data. On this alternative reading, Sims deprives macroeconomics of its traditional subject matter. Macroeconomics is atheoretical because there are no truths at the economy level for a theory to represent. To him, economists are closer to accountants than natural scientists (Sims, 1987).

This view of macroeconomics has precedents in the history of economics. Hayek (1979) argued that the 'wholes' studied in the social sciences are constructs of our mind; they do not represent any thing in the external world, and are not therefore subject to scientific laws (1979:96).[35] Also, recently, some new Keynesian economists have emphasized the vital importance of individual heterogeneity and direct interactions among market participants in explaining economic phenomena. Individual heterogeneity and direct interaction enormously complicate the relation between the micro and macro levels, making it impossible to attribute any theoretical interpretation to the relations emerging at the economy level. This has led these

[35] See Hoover (2001, ch. 5) for an appraisal of Hayek's position.

economists to favour an atheoretical view of macroeconomics, similar to Sims' approach (Colander, 1996:66).

6. Conclusion

This chapter began by defining the subject matter and objectives of macroeconomics, and outlined three arguments on the limitations of a purely statistical approach to macroeconomics. The arguments showed the necessity of domain specific information for modelling the structure and achieving macroeconomics' objectives. Hence, the most fundamental question in modelling the economy is concerned with the feasibility of obtaining such information in economics. In response to this question, we reconstructed two general approaches to macroeconomics. The theoretical approach suggests that the necessary information can be obtained by incorporating aggregative phenomena into the framework of microeconomic theory. In contrast, the atheoretical approach rejects the credibility of domain-specific knowledge in macroeconomics and, as an alternative, uses certain general principles concerning the connection between probability and causation to narrow down the number of admissible models compatible with the data. It, therefore, calls for revising macroeconomics' objectives. The contrast between these two competing views reveals that the issues regarding theories of economic behaviour and those about the link between the micro and macro levels are the most basic topics in macroeconomics. Of equal importance is the conjecture that one can sensibly talk of structural relations at the economy level.

Chapter 2

Rational Behaviour and Economic Theory

1 Introduction

> Unfortunately, the general hypothesis that economic agents are Bayesian decision makers has, in many applications, little empirical content: without some way of inferring what an agent's subjective view of the future is, this hypothesis is of no help in understanding his behaviour. ... To practice economics, we need *some* way (...) of understanding *which* decision problem agents are solving. (Lucas, 1981:223)

The difficulties in atheoretical study of aggregate data have led economists to propose a bottom-up approach to the study of macroeconomic phenomena that involves establishing a theory of individual behaviour and transforming it into a theory of the economy using aggregation methods. Thus, even though macroeconomics is primarily concerned with aggregate phenomena such as the unemployment level or general price movements, issues of individual behaviour have come to occupy a central place in theoretical economic analysis. The chief conjecture about *homo economicus* is that he behaves rationally. This conjecture is believed to be an 'engine of truth', serving to establish the laws of economic behaviour. Market forces are said to eliminate irrational behaviour, justifying focusing exclusively on the study of rational behaviour. This chapter studies the contribution of various rationality hypotheses to theoretical economics.

Although the literature provides a host of definitions of rational behaviour, the leading definition is based on the theory of subjective expected utility, best developed in Savage's book (1954 [1972]), *The Foundations of Statistics.* Savage's theory provides a general framework for studying possible contributions of the behavioural rationality hypotheses to economic analysis. We build our analysis around this theory and then explain how it applies to other rational choice theories. Savage's theory identifies rationality of behaviour with subjective expected utility (SEU) maximisation. Since its inception, the theory has been criticised on several grounds. It has been argued that the postulates of the theory are empirically wrong, its computational requirements exceed those of human beings, and people are not simply after their own utility (Camerer 1955; Sen, 1987; and Suppes, 1961). Nevertheless, these criticisms have not yet seriously shaken the central status of the theory in modern theoretical economic analysis.

This chapter argues that the rational choice theories on offer are inadequate for explaining and predicting behaviour, regardless of whether they are true or not. The theories give no explanation of how the agent models his choice situation, and defines his decision problem. They only state how, given a fully specified choice situation, the agent makes a choice that maximises his expected utility with respect to the situation. In using the theories to model behaviour, a host of substantive assumptions are needed to specify the agent's view of his choice situation and the problem he is trying to solve. These assumptions concern the agent's view of the causal structure of the environment, his values, beliefs, needs, and goals. It is only then that the theories become relevant and can predict how the agent solves the decision problem.

However, a theory of economic behaviour cannot take as given the structure of the choice situation and how the agent defines his decision problem, since the resolution of economic controversies critically hinges on how he models his choice situation and defines his decision problem than on the specific *method* by which he solves the problem. The expected utility maximization hypotheses are consistent with all sides of any substantive controversy in economics, and are therefore of a minor contribution to economic analysis. Substantial results, attributed to these hypotheses, are in fact the implications of the substantive assumptions made about how people specify their choice situation and how they re-specify it when faced with changes in the economy. The minor contribution of the hypotheses also explains why economists have not been very worried about their failure in individual choice experiments. In practice, because of the silence of the expected utility theories, economists have turned to econometric analysis to settle economic controversies. But, for several reasons, the success of the econometric method is very limited.

As an attempt to specify how the agent views the environment and defines his choice situation, new classical economists have set forth the rational expectations hypothesis. The hypothesis identifies the agent's view of the economic environment with the true model of the economy, suggesting that he maximises his expected utility with respect to the true model. Accordingly, as soon as the economist knows the structure of the economy, he also knows what the agent thinks of the economy. He then only needs to discover the agent's preferences to specify the decision problem he is trying to solve, and predict his behaviour. We will briefly review the rational expectations hypothesis to further our understanding of the current state of microeconomic theory. The chapter ends with a characterisation of the sort of theory of behaviour that is needed for thinking about the economy.

The plan of this chapter is as follows. Section 2 outlines Savage's theory of subjective expected utility. Section 3 restates some key questions concerning the theory's role in economic analysis. Section 4 discusses the postulates underpinning the theory by looking at some empirical counterexamples. Section 5 examines the adequacy of Savage's theory for predicting and explaining economic behaviour, and considers whether econometric methods can settle behavioural issues left unresolved by the theory. Section 6 takes up the rational expectations hypothesis. Section 7 concludes the chapter by arguing for the necessity of a learning-based theory of behaviour in economics.

2 Rational Choice

A rational choice theory of behaviour consists of a characterisation of rationality and a claim that a rational individual only chooses acts that satisfy the description. The oldest characterisation of economic rationality defines rational behaviour in terms of pursuit of self-interest – rational behaviour is self-interested behaviour.[1] Economists flesh out the idea of pursuit of self-interest by stating that a producer prefers more profit to less profit or a consumer prefers more money to less money. Another notion identifies behavioural rationality with the requirement that choices from different subsets of the universal set of available options be maximising solutions from the respective subsets according to some binary relation R. A person is then rational if his choice from any subset of the set of options available to him is the R-maximal element of the subset (Sen, 1987: 69). These definitions do not take into account the fact that full knowledge of the states of the world is never available, and one therefore always has to make decisions whose outcomes

[1] See Sen (1987) for an analysis of these rationality notions, and historical references.

are uncertain. A theory of rational behaviour should take this ubiquitous feature of real life decision-making seriously. and characterise rational behaviour under uncertainty.

Attempt at establishing a theory of rational choice under uncertainty demands a formal characterisation of uncertainty and a description of how the uncertainty thus characterised is taken into account in making decisions over alternative courses of actions (Sen, 1987:72). The theory used in this context is the *expected utility* theory, which weighs the value of each of the outcomes of an action by the respective probabilities of the different outcomes of the action. According to this theory, behaviour is rational if it is the outcome of expected utility maximisation. Depending on one's interpretation of probability, two general classes of expected utility theories can be defined. An interpretation, known as the objective interpretation, takes probability to be a measure of relative frequency. This interpretation underpins the Von Neumann-Morgenstern theory of expected utility. Another interpretation, called the subjective interpretation, takes probability to be a measure of the degree of belief that a person has in the occurrence of an event. This interpretation lies behind Savage's subjective expected utility theory, which will be the focus of analysis in what follows.

2.1 Savage's Theory of Subjective Expected Utility

As in any axiomatic system, there are three parts to Savage's theory. The first concerns definition of primitive and constructive notions. The second involves introduction of the axioms, and the third involves establishing the main result of the postulates. We describe the first two parts of Savage's theory in some detail, as they play a critical part in our understanding of the role of the rationality hypotheses in economic theorising.

2.1.1 Small Worlds

Savage starts with defining the primitives of his theory, including a choice set and a formal description of what the decision-maker is uncertain about.[2] To this end, Savage has a colourful example. Imagine you have just broken five good eggs into a bowl to make an omelette. A sixth egg, which for some reason you must either use for the omelette or throw it away, lies unbroken besides the bowl. You are about to decide what to do with this unbroken egg, which you do not know is good or rotten. Savage calls the sixth egg, the object about which you are concerned, the *World.* A description of the world, leaving no relevant aspect un-described, is called a *state* (of the world), herein good or rotten. Of these two states one does in fact obtain. It is called the *true* state. A set of states is called an *event.* The event that has every state of the world as its element is called the *universal* event, and denoted by S. There are at least three actions available to you: you may break the egg into the bowl containing the other five good eggs, you may break it into a saucer for inspection, and you may throw it away without inspection. Depending on the state of the egg, each of these acts will have some *consequences*, say, wasting five good eggs or making a clean saucer dirty. Let Z denote the set of all the consequences about which you are concerned. In deciding on an act, you must take into account possible states of the world and also the consequences that may follow from each act under each state of the world. Accordingly, an *act* is formally defined as a function that attaches a consequence to each state of the world, i.e., a mapping from S to Z. Let F denote the set of available acts. The set F is the choice set. In making a decision you prefer one act to others. A binary relation $\prec$ expresses your (strict) preferences over set F; thus for two acts f and g in F, $f \prec g$ means g is (strictly) preferred to f. The term 'world' is also used to refer to the pair (S, Z). Table 1 below gives a schematic representation of a world, corresponding to Savage's example (Savage, 1954 [1972]:14).

Although this example illustrates the basic notions of Savage's theory, it does not describe a typical situation to which the theory is intended to apply. To be precise, Savage develops his theory around an ideal agent whose guide in life is the

[2] For discussions of Savage's theory see Fishburn (1970, ch 14) and (1981).

proverb "Look before you leap" as opposed to "You can cross the bridge when you come to it" (Savage, 1954 [1972]:16). That is, in making a decision, he not only considers the consequences of his immediate acts but also those of the acts that he might need to take given the consequences of the immediate acts, and so forth. The objects about which he contemplates are not then simple acts but sequences of acts (Savage, 1954 [1972]:15). Savage carries the maxim "Look before you leap" to the extreme, assuming that the agent behaves as though he has only one decision to make in his entire life. "He must, [...], decide how to live, and this he might in principle do once for all" (Savage, 1954 [1972]:83). Consequently, the world $(\mathbf{S},\mathbf{Z})$ that he considers to represent his choice situation has an extremely large (potentially infinite) number of states and an ultimately refined description of the consequences of the acts under each state. Savage refers to an ultimately refined pair of states and consequences $(\mathbf{S}^*,\mathbf{Z}^*)$ as the *grand world.*

Table 2.1
Savage's world

Act	State	
	Good	*Rotten*
Break into bowl	Six-egg omelette	No omelette and five good eggs destroyed
Break into saucer	Six-egg omelette and a saucer to wash	Five-egg omelette and A saucer to wash
Throw away	Five-egg omelette and one good egg destroyed	Five-egg omelette

In reality, no matter how refined a world $(\mathbf{S},\mathbf{Z})$ is, it still does not include every conceivable state or consequence. Even if a person is currently considering a lifetime decision, he may not bother with the price of oil on 25th June 3500. Thus, the world $(\mathbf{S},\mathbf{Z})$ he actually considers to represent his choice situation is, in Savage's terms, a *small world* in the sense that each element in S can still be partitioned into smaller states and Z can still be replaced with an even more refined description of the consequences.

2.1.2 The Postulates

Savage's theory is based on seven postulates regarding the preference relation on *F*. The postulates can be stated in several equivalent ways. The statement below follows Fishburn (1971:191). Savage's first postulate asserts that the strict preference relation $\prec$ on *F* is a *weak order.* That is to say that $\prec$ is asymmetric and negatively transitive. The preference relation $\prec$ is asymmetric just in case, for every act *f* and *g* in *F*, if *f* is preferred to *g*, *g* is not preferred to *f*. And it is negatively transitive just in case, for every act *f*, *g*, and *h* in *F*, if *f* is not preferred to *g*, and *g* is not preferred to *h*, then *f* is not preferred to *h*:

Postulate 1: For every *f*, *g* and $\mathrm{h} \in \mathrm{F}$

a) if $f \prec g$ then not $g \prec f$;

b), if not $f \prec g$ and not $g \prec h$ then not $f \prec h$.

Let '~' denote indifference, which is defined as absence of strict preference. That is, for every *f* and *g* in *F*,

$$\mathrm{f} \sim \mathrm{g} \text{ if and only if neither } f \prec g \text{ nor } g \prec f .$$

It follows from Postulate 1 that the relation $\prec$ is transitive, $\sim$ is reflexive, symmetric, and transitive, and the preference relation on *F* is complete in the sense that, for every pair of acts *f* and *g* in *F*, exactly one of $f \prec g$, $g \prec f$, or $\mathrm{f} \sim \mathrm{g}$ holds (Fishburn, 1970, Theorem 2.1).

The second postulate says that states with similar consequences do not affect preferences. If acts *f* and *g* have different consequences over event *A* but agree over the complementary event A^c, they are ranked only on the basis of their differences on *A*. Similarly, if act f^* agrees with *f* and act g^* agrees with *g* on *A*, and further f^* and g^* agree on A^c, f^* and g^* are ranked in the same way that *f* and *g* are ranked. Let $f(s)$ be the consequence that *f* assigns to state *s* in *S*. The postulate can then be stated as follows:

Postulate 2: Suppose acts *f, g,* f^*, and g^* are such that:

a) $f(s) = g(s)$, $f^*(s) = g^*(s)$ for all s ∈ A^c

b) $f(s) = f^*(s)$, $g(s) = g^*(s)$ for all s ∈ A
then $f \prec g$ iff $f^* \prec g^*$.[3]

The third postulate states that the relative value of consequences is invariant across the states. To make this idea precise, two further notions are needed: *null* event and *constant* act. An event *E* is considered as *null* by a person if he is indifferent to acts that only differ on *E*. And an act is said to be constant if it yields the same consequence over every state of the world. Savage uses the notion of constant act to identify a consequence *x* with an act that leads to *x* over all the non-null states of the world. The third postulate then says that if a person prefers *y* to *x* given non-null event *A*, he prefers *y* to *x* in general and if he prefers *y* to *x* in general, he prefers *y* to *x* given *A*:

Postulate 3: If event *A* is not null, and

$f(s) = x$, $g(s) = y$ for all s ∈ A, $f(s) = g(s)$ for all s ∈ A^c,
then $f \prec g$ iff $x \prec y$.

So, for Savage, the set ***F*** not only does include the concrete acts but also, for every *z* in ***Z*****,** contains a constant act *f* that produces *z* in every state of the world, where by concrete acts we refer to acts that lead to different consequences over different states of the world. In this way, the postulate extends the preference relation $\prec$ from acts to consequences **Z**.

The fourth postulate supposes that the consequences following from an act under a state do not affect belief about the state. Suppose a person prefers consequence *y* to *x* and y^* to x^*. Then, if he prefers *y* to *x* when event *A* obtains rather than when event *B* obtains, he also prefers y^* to x^* when *A* obtains rather than when *B* obtains. Formally,

[3] "iff" stands for if and only if.

Postulate 4: Suppose $A, B \subseteq \mathbf{S}$; $x, y, x^*, y^* \in \mathbf{Z}$; $f, g, f^*, g^* \in \mathbf{F}$ are such that

a) $x \prec y$ and $x^* \prec y^*$

b) $f(s) = y$ for all $\mathrm{s} \in \mathrm{A}$ $\quad f(s) = x$ for all $\mathrm{s} \in \mathrm{A}^c$

$g(s) = y$ for all $\mathrm{s} \in \mathrm{B}$ $\quad g(s) = x$ for all $\mathrm{s} \in \mathrm{B}^c$

c) $f^*(s) = y^*$ for all $\mathrm{s} \in \mathrm{A}$ $\quad f^*(s) = x^*$ for all $\mathrm{s} \in \mathrm{A}^c$

$g^*(s) = y^*$ for all $\mathrm{s} \in \mathrm{B}$ $\quad g^*(s) = x^*$ for all $\mathrm{s} \in \mathrm{B}^c$

then $f \prec g$ iff $f^* \prec g^*$

This postulate paves the way for defining a qualitative likelihood relation $\prec *$ over S. Suppose y is preferred to x. Further, suppose acts f and g are such that f is equal to y on A and equal to x on A^c, and g is equal to y on B and equal to x on B^c. If g is preferred to f, then the only explanation for the ordering is that B is considered to be more probable than A; that is:

$$A \prec *B \text{ if and only if } f \prec g . \tag{2.1}$$

Thus, the preference ordering over $\boldsymbol{F}$ induces a likelihood ordering over S. These four postulates capture all the behavioural content of Savage's theory. Savage's remaining three axioms are technical postulates to ensure the existence of a mathematical representation of preferences and likelihood judgements (Kreps, 1988:128). We mention two of these postulates here. The first is the *non-triviality* postulate, which says:

Postulate 5: There is at least one pair of acts f and g such that $f \prec g$.

The other postulate states that, for every two non-indifferent acts in $\boldsymbol{F}$, and for every consequence x in $\boldsymbol{Z}$, the set $\boldsymbol{S}$ can be partitioned into arbitrarily small events so that altering either act to equal x *on just one of these events* does not reverse the preference ordering of the acts:

Postulate 6: For all f, $g \in \mathbf{F}$ such that $\mathrm{g} \prec \mathrm{f}$, and for all $x \in \boldsymbol{Z}$, there is a finite partition of S such that for every event A in the partition

a) if $f^*(s) = x$ for $s \in A$, $f^*(s) = f(s)$ for $s \in A^c$ then $g \prec f^*$

b) if $g^*(s) = x$ for $s \in A$, $g^*(s) = g(s)$ for $s \in A^c$ then $g^* \prec f$.

This postulate excludes infinitely desirable consequences. It also implies that if event *B* is less likely than event *C*, there is always a partition of *S* such that the union of each element of the partition with *B* is still less likely than *C*. Thus, *S* can endlessly be partitioned into smaller events. The postulate ensures that the preference relation $\prec$ has a property corresponding to the Archimedean property of natural numbers (and hence called the continuity postulate).

2.1.3 The Representation Theorem

These postulates lead to Savage's representation theorem. Savage shows that when preferences among acts in ***F*** satisfy the postulates, there exists, a unique finitely additive probability measure P on the set of all subsets of ***S*** such that

$$A \prec {}^*B \text{ if and only if } P(A) < P(B) \tag{2.2}$$

and, with P as given, there exists a real valued utility function *u* on **Z** such that for a finite **Z**,

$$f \prec g \text{ if and only if } \sum P(s)u(f(s)) < \sum P(s)u(g(s)). \tag{2.3}$$

According to (2.3), act *g* is preferred to act *f* if and only if the subjective expected utility of *g* exceeds the subjective expected utility of *f*. From this perspective, individual behaviour is rational if it is the outcome of subjective expected utility maximization.

3 Restating the Issues

Savage distinguishes between a normative and an empirical interpretation of his theory. The normative interpretation takes the postulates to be norms of rationality, providing a standard for actual people to follow. The empirical interpretation suggests that people's actual preferences among acts by and large obey the postulates and hence agree with a ranking of subjective expected utility. Here, we are concerned solely with the empirical interpretation that is taken for granted in positive economics.[4]

To analyse Savage's theory in its capacity as a descriptive theory of behaviour, it is vital to have, at least, an intuitive view of various phases of human decision-making. To this end, we rely on the common sense view implicit in Savage's discussion of human decision-making, which is also found elsewhere (Simon 1960). Reading the *Foundations*, one gets the impression that, according to Savage, there are two general phases in human decision-making. In the first phase, the decision maker draws on his view of the causal structure of the world to specify a small world, or more generally, the acts available to him, the states affecting the outcomes of the acts, and the consequences following from each act under each state. After that, he evaluates the likelihood of each state of the world and assesses the desirability of the consequences. We refer to a small world, the likelihood ranking of the states of the world, and the preference ranking of the consequences of the world as a *choice situation*:

$$\text{Choice situation} = \begin{cases} \text{Small world} \\ \text{Likelihood judgements over the states of the small world} \\ \text{Preferences over the consequences in the small world} \end{cases}$$

The choice situation defines the decision problem that the agent is trying to solve. In the second phase, the decision maker solves the problem by comparing the acts in the light of the likelihood of the states of the world and the desirability of their consequences to identify an act that is mostly likely to yield that which is desired the most.

This general characterisation of human decision-making is certainly imprecise. It, nevertheless, helps us to make a distinction between two types of theories of

[4] Savage favours the normative interpretation (1954:20).

behaviour. One possible class of theories of behaviour is those theories that are concerned with both phases of decision-making: They explain both how a person models his choice situation and defines his decision problem, *as well as* how he solves the problem. In contrast, a second possible class of theories of behaviour consists of those theories that take the structure of the choice situation and the definition of the decision problem as *given* and *exclusively* focus on how a person solves an already well-structured decision problem. We call the former group of theories *learning-based* theories of behaviour and the latter group *choice-based* theories of behaviour.[5]

A highly important point regarding Savage's theory is that it is a choice-based theory of behaviour. The reason for this classification can be explained by considering the restrictions that the postulates impose on the various stages of decision-making. According to the view just outlined, the process of decision-making starts with construction of a small world. The postulates impose two restrictions on the admissibility of a small world. Postulate 6 requires the set of the states of the world to be such that they can be partitioned indefinitely into smaller elements. On the other hand, the second, third, and fourth postulates necessitate the description of the consequences to be such that preferences among them can be stated without regard to beliefs about the states and that likelihood judgements about the states can be expressed without regard to the preference ranking of the consequences. These restrictions are surely nontrivial but leave the specific structure of the small world undetermined. Formation of a small world lies outside the theory:

> "I believe … that decision situations can be usefully structured in terms of consequences, states, and acts in such a way that the postulates of F. of S. [The Foundations of Statistics] are satisfied. Just how to do that seems to be an art for which I can give no prescription and for which it is perhaps unreasonable to expect one – as we know from other postulate systems for application" (Savage, 1971:79).

Now, consider beliefs and preferences. As our description of the theory reveals, Savage's postulates only require a certain correspondence between different parts

[5] A similar classification is found in Lane et al. (1996).

of a preference (choice) function. They make no reference to anything *outside* preference (choice) such as information, experience, goals, needs, and motivations. Nor do they presuppose any specific hypothesis about how values and beliefs are formed (Sen, 1993:495). The theory permits any internally consistent preference and likelihood ranking, thus taking the content of beliefs and values as *exogenous*. In Suppes' terms, both a cognitive and a moral idiot can be rational in the sense prescribed by Savage's theory:

> [Savage's theory] can be satisfied by cognitive and moral idiots. Put another way, the consistency of computations required by the expected utility model does not guarantee the exercise of judgement and wisdom in the traditional sense' (Suppes, 1984:207-8).

Moreover, since Savage's theory is silent about how a rational person models his small world and forms beliefs and values, it also is silent about how he defines his decision problem. The contribution of the theory to analysis of behaviour comes very late in positing how a person solves an *already well-structured* choice problem. The same point applies to other rational choice theories on offer, including the Von Neumann-Morgenstern theory; they too concentrate on the final phase of decision-making, and fall into the category of choice-based theories of behaviour.[6]

With these preliminaries, it is now possible to distinguish between two entirely different questions about Savage's theory in its capacity as a descriptive theory of economic behaviour: The first is whether it closely describes the process of human choice. And the other more critical question is whether a choice-based theory of behaviour is ever adequate for predicting and explaining economic behaviour, regardless of being true or false.

4 A Discussion of the Postulates

The first question, which has to do with the realism of the postulates, has been mostly investigated in experimental psychology. The second question, which

[6] For a review of major rational choice theories see Kreps (1988).

concerns the adequacy of choice-based theories of behaviour, has mostly been taken up in economics. Both approaches from psychology and economics are complementary. This section looks at some well-known findings from experimental psychology (Kahneman, 2003). The objective of examining them here is not simply to reiterate that the postulates fail. It is rather to explain why they fail, state a view of the nature of preferences that emerges from the findings, highlight the implications of the view for economic analysis, and set the stage for defining the kind of theory of behaviour needed in economics.

4.1 The Constructive Nature of Preferences

The first postulate implies that the decision maker has a complete preference ordering among acts in $\boldsymbol{F}$. To spell out what this implication really means, we need to note a distinction between 'indifference' and 'indecision.' Indifference refers to a case when the decision maker neither prefers f to g nor g to f but is ready to replace one of the options with the other in his preference ordering. Indecision, however, refers to a case when the decision maker neither prefers f to g, g to f, nor is ready to substitute one for the other in his preference ordering. Completeness, therefore, means that there are no cases of indecision. Thus understood, the weak order postulate is most consistent with the view that people have definite and ready-made preferences and as soon as they need to reveal them they can do so instantaneously and simultaneously (Thrall, 1954:183).

This view of preferences is incompatible with a large body of empirical evidence. In an early study, Frederich Mosteller and Philip Nogee (1951) observed that subjects on repeated elicitation of preferences would not always give the same answers. Similarly, Simonson and Tversky (1992) observed that varying the choice set could produce different patterns of preferences. In a set of experiments, they presented two groups of subjects with descriptions and pictures of microwave ovens taken from a catalogue. They invited one group of 60 individuals to choose between an Emerson microwave priced at $110 and a Panasonic priced at $180. The subjects were told that both items were on sale, one third off the regular price. Of these individuals, 57% chose the Emerson oven and

43% the Panasonic. In contrast, they presented the second group of 60 individuals with the same items together with a $200 Panasonic at a 10% discount. Only 13% of the people in the second group chose the more expensive Panasonic oven but its presence among the alternatives increased the percentage of the subjects who selected the less expensive from 43% to 60%. A similar pattern of preference variation has been found in a host of other experiments reported in Tversky and Shafir (1992).

If the subjects had definite and ready-made preferences or if they simply read preferences off "some master list" (Slovic, 1995:569), the introduction of the new expensive oven would not alter the percentage of people preferring the Emerson oven to the cheaper Panasonic one, and the subjects would exhibit an almost similar pattern of preferences in both experiments. Thus, the observed variation is most consistent with the view that people do not have ready-made preferences. Rather, when they need to choose among options, particularly among complex alternatives, they start in a sense from a state of indecision. They identify the features of the options relevant to the decision task at hand, compare the options in accordance with the attributes, and construct pro and con arguments for each option. The pro and con arguments are then used to *construct* a preference ranking of the options. From this perspective, since varying the choice set can make different attributes appear relevant or provide new information about the attributes already noted, a change in the choice set can give rise to construction of new pro and con arguments, and hence a different preference ranking. In the above example, the introduction of the more expensive microwave probably brought with it new useful clues that were not available before. When choosing among the ovens, the subjects most likely looked at the quality and the price of each brand. Since, in the first scenario, the quality difference between Emerson and Panasonic ovens appeared less dominant than the price difference, most subjects opted for Emerson. However, when the more expensive Panasonic was introduced because of a maintained correlation between price and quality, the subjects were led to think that the $180 Panasonic oven was of a much higher quality than previously thought. This additional clue rendered the quality difference more dominant than the price difference, driving more subjects to choose the $180 Panasonic oven, thinking that it was a bargain (McFadden, 1999:86).

If people do not have ready-made preferences but construct them from pro and con arguments, it is natural to expect that they sometimes fail to establish arguments necessary for transforming all cases of indecision into a definite preference ranking. There may not be enough information available about the options. The options may be complex, multi-dimensional, newly invented, and so forth. Or gathering information may be costly. There is thus every reason to expect that completeness can fail in practice.

The emphasis on the constructive nature of preferences is the hallmark of psychologists' view of preferences. There is, however, more to the claim in the psychological literature that preferences are constructed than revealed. To further our understanding of the constructive view of preferences, we look at another body of evidence termed preference reversals. The discovery of the preference reversal phenomenon goes back to a study by Slovic and Lichtenstein (1968), where they noticed that selling prices of gambles were more highly correlated with payoffs than with probabilities of winning but choices among lotteries were more highly correlated with probabilities of wining than with the payoffs. The observation led the researchers to the conjecture that if subjects were offered two gambles with the same expected returns, one featuring a high probability of winning a modest sum of money (called *H* for high chance of wining) and the other featuring a low probability of wining a relatively large amount of money (called *L* for low chance of winning), the subjects would most likely choose the high probability bet *H* but price higher the low probability bet *L*. Lichtenstein and Slovic (1971) tested this conjecture by confronting a group of subjects with pairs of gambles like the one depicted in Table 4.2:

Table 2.2

Preference Reversal Phenomenon

H-bet	**L-bet**
99 percent of winning $4	33 percent of winning $16
1 percent of loosing $1	67 percent of loosing $2

The subjects were asked to state the cash equivalent of the *H* bet, (i.e., the minimum price at which they would be willing to sell the bet if they owned it), the cash equivalent of the *L* bet, and make a choice between the two bets. Most subjects, as conjectured, chose the *H* bet but assigned a higher selling price to the *L* bet. In an experiment, 127 out of 173 subjects (or 73.4%) assigned a higher selling price to the *L* bet in every pair in which they chose the *H* bet, even though the expected value of both bets were the same.

As with any empirical finding, the preference reversal phenomenon is subject to competing explanations, arising from various assumptions that one is ready to make about preferences. There are several assumptions in the economic and rational choice literature that are relevant to the explanation of preference reversals. One assumption, which we have been discussing, is that people possess well-defined and stable preferences (Stigler and Becker, 1977). A second assumption is *description invariance* that says preferences among options do not depend on the manner in which they are represented or displayed. A third assumption is *procedure invariance* that says strategically equivalent methods of elicitation give rise to the same preference order; specifically, it does not matter whether choice questions or evaluation inquiries are used to elicit information about preferences.[7] Let C_H and C_L denote, respectively, the cash equivalent of *H* and *L*. Procedure invariance implies that the decision maker prefers *H* to a cash amount *X* if and only if his cash equivalent for *H* exceeds *X*, and that he is indifferent between *H* and *X* if and only if $C_H = X$. Finally, a fourth assumption is monetary consistency, which says people prefer more money to less. If *X* and *Y* are sure cash amounts, then $X > Y$ implies $X \succ Y$, where > refers to the ordering of the cash amounts. Given these assumptions, preference reversal implies violation of transitivity, as shown below:

[7] A principle of economic thinking is that opportunity costs and out of pocket costs should be treated alike. This implies that preferences should depend only on relevant differences between options, not on how these differences are represented.

$$\left.\begin{array}{l}1.\ H \succ L\\ 2.\ C_L > C_H\end{array}\right\}\ \text{Preference Reversal}$$

$$\left.\begin{array}{l}3.\ C_H \sim H\\ 4.\ C_L \sim L\end{array}\right\}\ \text{Procedure Invariance}$$

$$5.\ C_L \succ C_H \qquad \text{Monetary consistency}$$

....

$$\begin{array}{ll}\therefore\ L \succ C_H & (4 \text{ and } 5)\\ \quad H \succ C_H & (1, 4 \text{ and } 5)\end{array}$$

which contradicts $C_H \sim H$ (hence, intransitivity). Economists initially interpreted the reversals as violations of transitivity, and called for establishing a theory of expected utility that could account for intransitive choices (Machina, 1987). In contrast, psychologists saw more in the phenomenon than intransitivity, and began investigating whether it could have arisen from the failure of any of the other assumptions, in particular procedure invariance. This led to the definition of two rival hypotheses concerning the causes of preference reversals – the intransitivity and non-invariance hypotheses.

To investigate these hypotheses, Tversky, Slovic, and Kahneman (1990) extended Lichtenstein and Slovic's initial experimental setting by including an option of receiving a pre-specified sure cash amount *X*. In this setting, they asked the subjects to state their preferences between each of the pairs in the triple $\{H, L, X\}$ and also announce their cash equivalent for bets *L* and *H*. The researchers then focused on the preference reversal cases in which *X* fell between the cash equivalents C_L and C_H announced by the subjects; that is, the cases in which the reversals had the pattern:

$$H \succ L \text{ and } C_L > X > C_H. \qquad \text{(PR)}$$

The hypotheses of intransitivity and non-invariance give rise to different testable implications for preference orderings of those subjects whose preferences satisfy the PR pattern. To spell out some of these implications, note that procedure invariance can fail either because of *overpricing* of *L*, *underpricing* of *H*, or both overpricing of *L* and underpricing of *H*. Overpricing of *L* is said to occur if a person offers cash equivalent C_L for L that is greater than X but in a direct choice between C_L and L he prefers C_L, (i.e., $C_L \succ L$). Underpricing of H is said to occur if a person announces cash equivalent C_H for H that is less than X but in a direct choice between the H and C_H he prefers H to C_H (i.e., $H \succ C_H$). Thus, there are at least four potential explanations of the preference reversals. Below we derive the implications of the explanations based on the failure of transitivity, overpricing of *L*, and underpricing of *H*:[8]

Hypothesis I:	**Hypothesis II:**	**Hypothesis III:**
Intransitivity	**Overpricing of *L***	**Underpricing of *H***
$1. H \succ L$	$1. H \succ L$	$1. H \succ L$
$2. C_L > X > C_H$	$2. C_L > X > C_H$	$2. C_L > X > C_H$
$3. C_L \succ X \succ C_H$	$3. C_L \succ X \succ C_H$	$3. C_L \succ X \succ C_H$
$4.\ C_L \sim L$	$4.\ C_L \succ L$	$4.\ C_L \sim L$
$5. C_H \sim H$	$5. C_H \sim H$	$5. H \succ C_H$
....	....	....
$\therefore\ L \succ X$	$\therefore\ X \succ H$	$\therefore\ L \succ X$
$X \succ H$	$X \succ L$	$H \succ X$

Tversky et al. (1990) looked at the relative frequencies of these implied preference patterns among the preference orderings announced by the subjects. Their findings were astounding. In the study, 40% to 50% of the participants showed preference reversals consistent with the PR pattern. Of these subjects, only 10% had preferences consistent with the intransitivity hypothesis while the remaining 90% had preferences consistent with the non-invariance hypotheses. In particular, nearly two-thirds of the reversals were consistent with overpricing of the *L* bet. The researchers, therefore, concluded that the failure of procedure invariance (overpricing of the *L* bet) was the major cause of the preference reversals.

Several hypotheses have been proposed to explain the procedure invariance failure, including the *scale compatibility hypothesis*. The hypothesis suggests that an attribute of an object is given more weight when it is compatible with the

[8] The case where procedure invariance fails because of both overpricing and underpricing is similar.

response mode than it is not. Since the cash equivalence of a bet is stated in, say, dollars, compatibility implies that payoffs, which are also stated in the same units, are weighted more heavily in pricing than in choice. As a result, the L bet is overpriced relative to the H bet, leading to the observed preference reversals (Tversky, 1996:189-190).

The conclusion that preference reversals are to a large extent due to the failure of procedure invariance fits particularly well with another significant body of evidence, called 'framing effect', which points to the systematic failure of description invariance (Tversky and Kahneman, 1986). These findings altogether lend strong support to the viewpoint that there are no ready-made, well-defined and stable preferences; preferences are constructed on demand and, more importantly, are *endogenous* to the decision process. Moreover, the findings indicate that formation of preferences is *sensitive* to the manner in which options are framed and questions are posed (Fisher, et al., 1999:1074). Owing to this sensitivity, behaviour is likely to vary across situations considered as identical by the rational choice theory (Tversky and Thaler, 1990:210).

4.2 The Entanglement of Values and Beliefs

Savage's remaining behavioural postulates require a small world where preferences among the consequences and beliefs about the states are completely disentangled. Consider the third postulate, which says consequence x is preferred to y given a non-null event A if and only if x is preferred to y in general. Specialising A to a single state, it says that the relative value of x is invariant across the states. If beliefs about the states of the world affected the desirability of x, the relative value of x could vary across the states. In that case, the postulate would no longer apply. For the postulate to hold there must be a small world refined enough to permit expressing preferences among the consequences without regard to beliefs about the states and expressing likelihood judgements about the states without regard to preferences among the consequences (Shafer, 1986:743). The second and fourth postulates are also predicated on the existence of a small world where beliefs and values are entirely disentangled.

In reality, the value of the consequences that a person includes in his small world may depend on his beliefs about the likelihood of the states of the world, and as his beliefs about the states change so does his preference ordering of the consequences. Savage was aware of this fact. Considering a person who is about to decide whether to buy a bathing suit or a tennis racket, he acknowledged that whether the preson prefers 'possessing a bathing suit' to 'possessing a tennis racket' might depend on whether he expects to go on a picnic at a beach or in a park (1954:25). Nevertheless, he took such dependence as an indication of the inadequacy of the person's description of his choice situation. Possessing a bathing suit and a tennis racket, he argued, should be regarded as acts, not consequences. Appropriate consequences in this case would be things like 'having a refreshing swim with friends at a beach in a sunny day' and 'sitting on a shadeless beach twiddling a brand new tennis racket while one's friends swim.' Evaluation of these consequences does not depend on which of the two states "picnic at the beach" or "picnic in the park" occurs. In general, he conjectured that it would be possible to completely disentangle values from beliefs by carrying the refinement of the consequences to *"its limits"* (Savage, 1954: 25). In an adequately (ultimately) refined world, Savage suggested, there would be no link between one's values and beliefs.

The difficulty with this proposal is that an attempt at refining the consequences in $\boldsymbol{Z}$ can force a refinement of the states in $\boldsymbol{S}$. This is because the states $\boldsymbol{S}$ must be detailed enough to determine which element of $\boldsymbol{Z}$ will be achieved by each act in $\boldsymbol{F}$ (Shafer, 1986:474). Savage's suggestion to take "refreshing swim with friends" as the appropriate consequence, rather than "possession of a bathing suit", requires refining $\boldsymbol{S}$ to contain states such as whether friends come, whether the temperature is warm enough for a refreshing swim, whether the beach is clean, and so forth. These additional states can render one's evaluation of the elements in the refined $\boldsymbol{Z}$ dependent on ones' beliefs about which element in the refined $\boldsymbol{S}$ is true. Perhaps you would prefer twiddling a brand new tennis racket while your friends swim if you knew that your friends would bring along someone whom you don't like. There is *a priori* no reason to think that, for any set of acts, there is always an ultimately refined world in which preferences among the consequences can be completely disentangled from beliefs about the states. Even if such a world existed, it would not be anything similar to a description that a typical individual would have of his choice situation. Later in his life, Savage acknowledged that an "ultimate" analysis might not after all exist and if it existed it might be quite "cumbersome":

> "A nickel is itself a lottery ticket, and one objection to getting miserably drenched is that it seems conducive to illness. If the problem were concerned with illness or the possibility of accidentally buying poisoned food, then of course the notion of consequences would have to be further analysed. An ultimate analysis might seem desirable, *but probably it does not exist and certainly threatens to be cumbersome*" (Savage, 1971:79; Italic added)

The conclusion is that in the small worlds we normally construct to represent our choice situations, our evaluation of the consequences of the world depends on our beliefs about the states of the world. This dependence of preferences on beliefs defines another aspect of the constructive view of preferences. Finally, it is equally important to emphasise the constructive nature of small worlds; they are also the outcome of our beliefs and models about the world, and evolve with the evolution of our beliefs and models. Small worlds, beliefs and preferences are not 'there like the Rocky Mountains', to use Stigler and Becker's phrase (1977:76); they are all constructed.[9]

These remarks, though self-evident may seem, have a profound implication for modelling behaviour. Since different constructions of beliefs, small worlds and preferences can give rise to systematically different choices, no theory can accurately predict or explain (dynamic) behaviour without carefully taking into account the *factors* affecting formation of beliefs, small worlds and preferences (Bowles, 1998:75). Therefore, a satisfactory theory of behaviour should explain how beliefs, small worlds and preferences are formed; it cannot take them as *exogenous*. To illustrate the point, let us return to the preference reversal phenomenon. The phenomenon shows that payoffs and probabilities of wining have quite different effect in pricing gambles and choosing among them. Payoffs are weighted more heavily in pricing gambles whereas probabilities of winning

[9] We did not discuss the constructive nature of beliefs, as it is indisputable in economics (Aumann, 1987:13).

are weighted more heavily in choosing among gambles. This means a theory of behaviour cannot correctly predict or explain pricing and choice behaviour in such cases without taking into consideration the dominance of payoffs in pricing and probabilities in choice. A theory that pays no attention to the different roles of these factors is surely bound to yield wrong predictions. In light of this analysis, the real difficulty with Savage's theory is not simply that it gives a wrong description of human choice. Rather, the real difficulty is that the theory takes things as exogenous that cannot be taken as exogenous by a theory of behaviour. In general, because of the constructive nature of beliefs, small worlds and preferences, no choice-based theory of behaviour can ever adequately explain or accurately predict behaviour.

5 The Limited Role of Rational Choice Theories

In economics, critics of the rational choice theories have until recently paid less attention to the realism of the postulates. Instead, they have mainly disputed the contribution that Savage's theory, or similar rational choice theories, can make to economic theorising, whether they are true or not. This section draws on the works of economists such as Kenneth Arrow (1986), Arthur Goldberger (1989), Robert Lucas (1976), Herbert Simon (1984, 1986), and the philosopher Patrick Suppes (1961) to argue why choice-based theories of behaviour are in principle inadequate for dealing with substantive economic controversies. The analysis complements the lessons of the investigations in behavioural psychology. We continue working within the framework of Savage's theory but the relevance of the analysis to other choice-based theories will be evident.

5.1 Choice-based Theories and Economic Controversies

Savage's theory takes the structure of the small world as well as the content of beliefs and preferences as given, and only says how an agent solves a well-structured decision problem. This means, in modelling behaviour using the theory, a host of *exogenous* assumptions are needed to specify the agent's choice situation and decision problem. These assumptions are made through specification of a utility function, the variables entering the function, the physical or socio-economic laws determining the variables, their joint probability distribution, and so forth. Without such assumptions, the theory makes no *concrete* prediction about observed behaviour.[10]

Now, one way to reconstruct the critique in the economic literature of Savage's theory is that these assumptions are not like the auxiliary assumptions necessary for making a general theory speak about the world. Quite the opposite, they essentially assume the solutions to the very same questions that a theory of behaviour is expected to answer. The reason is that, by varying the exogenous assumptions, every conceivable side of any substantive economic controversy can be rationalised or, in other words, derived as the outcome of subjective expected utility maximisation. The key to settling an economic controversy, therefore, lies in correctly specifying the exogenous assumptions. However, correct specification of the assumptions necessary for making Savage's theory to have any implication about a substantive controversy requires nothing less than knowing the correct side of the controversy. As a result, when the necessary exogenous assumptions in a given situation are fully specified, nothing essential remains for the theory to predict; the predictions are already in the assumptions. Savage's theory simply repackages them in terms of subjective expected utility maximisation. But a theory of behaviour cannot take for granted the answer to the very same questions that is expected to address. Consequently, regardless of whether it is true or not, the theory cannot function as a theory of economic behaviour.

We defend these points by examining a rational choice-based model of economic behaviour to demonstrate how by varying the exogenous assumptions in the model any side of an economic controversy can be rationalised. We will then explain why the analysis generally holds.

[10] This is not to deny the conditional restrictions that Savage's postulates impose on observed behaviour, such as those tested in the Allais's paradox (Allais, 1953, 1997).

5.1.1 The Effect of Compensatory Educational Programs

We borrow our model from a paper by Arthur Goldberger (1989), who scrutinizes Gary Becker's claim about the effectiveness of public compensatory educational programs. The effectiveness of these programs is still a matter of controversy.[11] On one hand, there is the view that such programs positively contribute to the well being of the children participating in them and improve their future earnings. On the other, there is the view that the programs are ineffective, since parents whose children participate in them reallocate the portion of their income that they would have otherwise spent on their children. This is known as the *offsetting effect.* A satisfactory theory of behaviour is expected to have some implication for the truth of the offsetting effect.

Becker (1981) seems to suggest that, by extending expected utility analysis to parents' expenditure decision-making, he has been able to establish the offsetting effect. Goldberger is critical of this claim. He argues that the offsetting effect implied by Becker's theoretical model is not the result of the expected utility maximisation assumption but depends on the exogenous assumptions introduced to specify, in our terms, the decision problem being solved by the parent. If the choice situation were defined slightly differently, a different conclusion would be derived. The hypothesis of expected utility maximisation, Goldberger shows, is consistent with both opposing views on the effect of compensatory educational programs. We review Goldberger's analysis in some detail as it explains how formal economic modelling proceeds in practice.[12]

A key to resolving the controversy surrounding the effectiveness of public education programs is to know how parents would respond to a change in the income of their children. To address this query, Becker assumes a representative parent; that is, he supposes that all parents whose children participate in the programs have the same utility function, live in the same environment, and receive the same information. He speaks of 'the parent' rather than parents.

[11] Simon (1986) and Conslik (1996) also mention this example. Another example, relating to Becker's work on marriage market, is found in Lam (1988).
[12] The account given here of Becker's work is based on Goldberger (1989).

Having done so, he introduces several assumptions about the representative parent. The first is that she has an *interdependent* (i.e., non-egoistic) utility function that allows a concern with the consumption patterns of others (Pollak, 2002:10). In particular, it is assumed that her utility derives from her own consumption *C* and her child's income *Y*. Becker's second assumption is that she has a Cobb-Douglas utility function *U*:

$$U = \alpha \log Y + (1-\alpha)\log C. \tag{5.1}$$

The parameter α, which lies between 0 and 1, reflects relative preference for child income as against own consumption. According to the function, the parent's relative preference for her child's income as against her own consumption is independent of *Y* and *C*. The parent receives income *X* which is divided between consumption *C* and investment in child *I*:

$$X = C + I. \tag{5.2}$$

Becker's third assumption relates to the mechanism generating the child's overall income. The child's overall income is supposed to be an *additive* function of the parent's investment *I* and another general component *E*, called "Luck", which represents natural endowments, social status, government support, luck in the market, and so forth. The rate of return on investment *I* is *r*. Let $m = 1 + r$. The child's income *Y* is thus given by

$$Y = mI + E. \tag{5.3}$$

Since the time unit is a generation, *Y* and *X* are technically wealth or permanent income. Consequently, the return factor $m = 1 + r$ can be taken to be larger than unity, say, 1.5 or even more. As a final assumption, the parent is assumed to have full knowledge of her child's luck. She decides at her own consumption and investment in her child by maximising (5.1) subject to (5.2) and (5.3), which yields the optimal level of investment and consumption as:

$$I = \alpha X - (1-\alpha) E/m, \tag{5.4}$$

$$C = (1-\alpha)X + (1-\alpha) E/m. \tag{5.5}$$

Substituting (5.4) back into (5.3) gives the income transmission rule,

$$Y = bX + \alpha E, \qquad b = \alpha m. \tag{5.6}$$

where the parameter b is the "propensity to invest in the child" and α is the "fraction of family income spent on the child" (Goldberger, 1989:506).

The income transition rule (5.6) describes how the parent responds to an increase in her child's luck. Suppose there is a dollar increase in E. According to (5.6), the child income increases only by the fraction of α; the parent partially offsets the increase in E by increasing her own consumption (see (5.4)). Becker takes this implication to argue that "public education and other programs to aid the young may not significantly better them because of compensating decreases in parental expenditures" (Becker, 1981:153).

This conclusion, as shown by Goldberger, is not an inevitable implication of the expected utility maximisation principle. The offsetting result is based, among other things, on the assumption that the child's income is an additive function of parental investment and child's luck. If the child's income were, for instance, a multiplicative function of parental investment and luck, the offsetting result would no longer follow. To illustrate this, Goldberger replaces the additive function (5.3) with the multiplicative function,

$$Y = mIE. \tag{5.7}$$

In this case, the parent divides her income between her own consumption and investment in her child according to

$$I = \alpha X, \tag{5.8}$$

$$C = (1-\alpha)X. \tag{5.9}$$

Evidently, neither optimal investment nor optimal consumption any longer depends on *E*. And the income transmission rule becomes,

$$Y = bXE\,. \tag{5.10}$$

An increase in *E* by the government no longer affects parental investment decision. If *Y* followed hypothesis (5.7) rather than (5.3), public education programs could have strong effects (Goldberger, 1989:507). It is thus wrong to suggest that the expected utility maximisation assumption implies the offsetting effect or its negation. Becker's result is critically based on the specific hypothesis (5.3) about the structure of the environment.

Becker's offsetting result is also based on the choice of a homothetic utility function.[13] Goldberger does not consider this but the choice of a non-homothetic utility function undermines the result too. Consider the simple non-homothetic utility function,

$$U = Y + \ln C\,, \tag{5.11}$$

while retaining the assumption that the child's income is an additive function of parental investment and the child's luck. The optimal level of consumption and investment is given by

$$C = m^{-1} \tag{5.12}$$

$$I = X - m^{-1} \quad . \tag{5.13}$$

The new income transmission rule will be

$$Y = mX - 1 + E\,. \tag{5.14}$$

[13] For definition of homotheticity see Appendix *A*.

As evident from (5.12) and (5.13), the parent's optimal consumption and investment are independent of the child's luck. And so, the model does not entail the offsetting effect.

The subjective expected utility maximisation assumption is, therefore, consistent with both opposing views on the effectiveness of public education programs. It is the exogenous assumptions about the shape of the parent's utility function, the variables entering it, and the mechanisms generating the variables that make a model entail the offsetting result or its negation (Pollack, 2002:9). To predict the effect of the programs using Savage's theory, one ought to know, among other things, whether the parent cares about her child, how she cares, whether her relative preference for her own consumption and investment in her child vary with changes in her child's income, what she thinks of the mechanism generating her child income, how she predicts the effect of her investment on the future wealth of her child, and so forth. But if we knew the answers to these queries, we would already know how she would behave in response to a change in her child's income; the answer to the question concerning the effect of the educational programs are implicit in the answers to these questions. In the end, we might need to introduce an optimisation principle to infer how she actually solves her decision problem but the principle would not need to be the subjective expected utility maximisation principle; satisficing would equally do (Arrow, 1984).[14] Nor is the principle an 'engine of truth' standing above all the other assumptions; it is an assumption like other substantive assumptions entering a model of parent behaviour.

5.2 How Economic Controversies Are Settled

Resolution of economic controversies depends critically on the choice of exogenous assumptions than on the expected utility maximisation principle. In practice, economists have turned to econometric analysis of aggregate data to select a rational choice model. The analysis involves trying various combinations of plausible assumptions to establish a rational choice model that well fits

[14] See Appendix *B* for a definition of satisficing.

aggregate data, and using the model to address behavioural questions of interest. A basic question is whether the econometric approach can fill the theoretical vacuum left by the rational choice theories, and dispense the need for an alternative theory of behaviour. To address this query, we first consider a typical application of the econometric method from the history of economics to bring to the fore some of the assumptions underlying the method. And, on that basis, we will explain why it fails.

5.2.1 The Effect of Economic Events on Votes

An issue of interest in economics concerns the effects of economic events on votes. The literature contains conflicting views on the matter. Kramer (1971), for example, studied data on U.S. voting behaviour, concluding that economic fluctuations have a significant effect on congressional elections, whereas Stigler (1973) concluded that they do not. Against this background, Fair (1978) set himself the task of presenting a theoretical model of voting that is general enough to allow one to define the disagreements in the literature and test them. So, he set up a rational choice model of voting behaviour. Fair considers a two-party political system such as the US, referred to as Democrat and Republican, and focuses on presidential, rather than congressional, elections. Let us define the following notations:

$E(U_{it}^{d})$: voter *i's* expected utility if the Democratic candidate is elected at time *t*.

$E(U_{it}^{r})$: voter *i's* expected utility if the Republican candidate is elected at time *t*.

These expected values are based on the information available up to time *t*. Let V_{it} be a variable that is equal to one if voter *i* votes for the Democratic candidate at time *t* and zero if he votes for the Republican candidate at time *t*. The expected utility theory implies that:[15]

[15] For simplicity the case when the voter is indifferent is not considered here.

$$V_{it} = \begin{cases} 1 & if \quad E(U_{it}^{d}) > E(U_{it}^{r}) \\ 0 & if \quad E(U_{it}^{d}) < E(U_{it}^{r}) \end{cases}. \tag{5.15}$$

Voter *i* votes for the candidate (party) that gives the higher expected utility. Further, let

$$U_{it} = f(\mathbf{Z}_{it}) \tag{5.16}$$

be the voter *i*'s utility function, with $\mathbf{Z}_{it}$ being the variables affecting the voter's utility. The expected utility theory has no implication for $\mathbf{Z}_{it}$ or function *f*. In light of this, Fair interprets the differences in the literature in terms of whether $\mathbf{Z}_{it}$ includes economic factors and, if so, how they affect votes. Fair's assumption is that if economic factors influenced votes, the voter's expected future utility if a party were in power would depend on his forecast of the performance of the economy under the party. He thus embarks on testing whether the voter's expected future utility under a party depends on his forecast of the performance of the economy under the party. This raises several issues about how the voter measures the state of the economy *and* how he forecasts the performance of the economy under a party.

Fair first considers modelling the procedure used by the voter to forecast the performance of the economy under a party. He makes two assumptions about the method (1978:161):

> A_1: The forecast reflects accumulated past experience.
> A_2: The forecast attaches more weight to recent than to remote periods.

This means the voter bases his forecast of the economic performance of a party on the performance of the economy when the party was recently in power. As a result, if economic factors affected voting decisions, the voter's expected future utility under a party would be a function of how well the economy performed when the party was recently in power. Let

$tj1$: last election from t back that party j was in power,
$tj2$: second-to-last election from t back that party j was in power,
ξ_i^j : a vector of variables specific to voter i , assumed to be independent of the variables used to measure the performance of the economy.
M_h: some measure of economic performance of the party in power during the four years prior to election h.

j takes two values d for the Democratic party and r for the Republican party. If party j was in power at time t, then $tj1$ is equal to t.[16] One way to formulate postulates A_1 and A_2 is the following:

$$E(U_{it}^d) = \xi_i^d + \beta_1 \frac{M_{td1}}{(1+\rho)^{t-td1}} + \beta_2 \frac{M_{td2}}{(1+\rho)^{t-td2}} \tag{5.17}$$

$$E(U_{it}^r) = \xi_i^r + \beta_3 \frac{M_{tr1}}{(1+\rho)^{t-tr1}} + \beta_4 \frac{M_{tr2}}{(1+\rho)^{t-tr2}} \tag{5.18}$$

where parameters β_1, β_2, β_3, and β_4 are unknown coefficients and ρ is an unknown discount rate. Equations (5.17) and (5.18) state that voter i's expected future utility under a party is a function of a vector of individual specific variables and the party's performance during the last two times that it was in power. The performance measure is discounted from time t back at rate ρ. For ρ greater than zero, more weight is attached to recent than to remote periods. If desired, the equations can be expanded to include more than just the last two periods each party was in power. Also, M_h can be a function of several variables representing the state of the economy.

In this setting, Fair attempts to settle the disagreement about the effect of economic events on votes by fitting to aggregate voting data various possible models arising from substituting alternative performance measures for M_h in equations (5.17) and (5.18), and determining if any of the models adequately account for the data. To justify the use of aggregate data for estimating the

[16] t is a time trend that takes, for instance, a value 8 in 1916, 9 in 1920, and so on.

individual parameters in (5.17) and (5.18), Fair introduces four extra assumptions about the voters and the economy. Let

$$\psi_i = \xi_i^r - \xi_i^d \text{, and} \tag{5.19}$$

$$q_t = \beta_1 \frac{M_{td1}}{(1+\rho)^{t-td1}} + \beta_2 \frac{M_{td2}}{(1+\rho)^{t-td2}} - \beta_3 \frac{M_{td1}}{(1+\rho)^{t-tr1}} - \beta_4 \frac{M_{tr1}}{(1+\rho)^{t-tr2}}. \tag{5.20}$$

It follows from equations (5.15), (5.17) and (5.18) that voter *i* votes for the Democratic candidate if $q_t > \psi_i$ and votes for the Republican candidate if $q_t < \psi_i$.[17] Having said this, the four assumptions to link the individual and the aggregate levels are as follows.

A$_3$: All voters use the same measure of performance;
A$_4$: The coefficients β_1, β_2, β_3, β_4 and ρ in (5.17) and (5.18) are the same for all voters;
A$_5$: ψ_i in (5.19) is evenly distributed across voters in each election between some numbers $a+\delta_t$ and $b+\delta_t$, where $a<0$ and $b>0$. *a* and *b* are constant but δ_t can vary across elections;[18]
A$_6$: There are an infinite number of voters in each election.

Now, let V_t stand for the percentage of the two-party vote that goes to the Democratic candidate in election *t*. It follows from (5.15), (5.17), (5.18) and assumptions A$_3$ through A$_6$ that:

$$V_t = \alpha_0 + \beta_1^* \frac{M_{td1}}{(1+\rho)^{t-td1}} + \beta_2^* \frac{M_{td2}}{(1+\rho)^{t-td2}} - \beta_3^* \frac{M_{td1}}{(1+\rho)^{t-tr1}} - \beta_4^* \frac{M_{tr1}}{(1+\rho)^{t-tr2}} + v_t \tag{5.21}$$

which makes no reference to the variables ξ_i^d and ξ_i^r (Appendix *C*). Given certain restrictions on the error term v_2, equation (5.21) can be estimated from aggregate

[17] ψ_i is voter's "expected utility bias" in favour of the Republican candidate. It is, in other words, "voter *i*'s expected utility difference between the Republican and Democratic parties before any consideration is given to their past performances.

[18] The key assumption is here that this difference differs across voters in a uniform way (Fair, 1987:162).

data. Fair replaces several measures of performance for M_h to investigate if any of the resulting equations accounts for the U.S presidential election data. The measures include "the growth rate of real per capita GDP over some specified period prior to the election" and "the absolute value of the inflation rate over some specified prior to the election". In the fitted equation, the coefficient estimates for the growth rate and the inflation rate in the year prior to the election appear significant, and of the expected sign. Economic factors as measured by these variables, Fair concludes, do actually enter the voter's utility function.

This study is a typical illustration of how substantive controversies – such as what variables affect votes and how – are settled in economics. In practice, substantive controversies are resolved by searching for a model that best fits aggregate data. Therefore, the answers to substantive economic queries do not come from formal economic theory, which is another name for a rational-choice based model of behaviour. Quite the opposite, one needs the answers to the questions to select an appropriate rational choice model. Rational choice models assume the answers to substantive economic questions rather than addressing them (Fisher, 1989:118; Conslik, 1996:685).

5.3 Why the Econometric Method Fails

For several reasons the econometric approach fails to fill the theoretical vacuum left by the rational choice theory. To begin with, the method requires making the assumption that the laws of the individual and the economy coincide; without this assumption aggregate data cannot be employed to select a rational choice model. Fair takes this coincidence for granted by assuming that all voters have the same utility function, use the same performance measure, share the same rules for forecasting the performance of the economy, and individual characteristics are uniformly distributed in the population. Such assumptions, which are necessary for the laws of the individual to coincide with the laws of the economy, are incredibly strong. Even so, they are not adequate to ensure the coincidence. A full justification of this point demands a proper understanding of the conditions under which the laws of the individual and the economy are the same, which is given in

Chapter 6. Here, it suffices to note that economic variables change status when one moves from the micro level to the aggregate level. The individual takes prices, the rate of economic growth, inflation, and the unemployment level as given but the economy cannot take them as given; it determines them. It is thus wrong to assume that models true of the aggregates are also true of the individuals or vice versa. The fate of rational choice models cannot be settled by analysis of aggregate data. A different type of data is needed.

Another reason for the failure of the econometric method relates to the *nature* of the exogenous assumptions entering a rational choice model. The assumptions convey information about the agent's small world, beliefs, and preferences. As we learnt from our review of the empirical investigations into human decision-making, small worlds, beliefs and preferences are not invariants of human behaviour. Rather, they are *constructed* on the basis of past experiences, goals, needs, and the socio-economic structure of the society. As a result, they vary with accumulation of experiences, arrival of new information, and changes in the economy. This means even if the aggregation difficulties arising from the lack of correspondence between the micro and macro levels did not exist, the method could at best establish the model that was true of the individual during the period from which the data were collected. It could not, logically speaking, establish the model that would be true of the individuals if they received different information, if a different policy regime were in place, or if the institutional structure of the economy were different. Thus, the econometric method is unsuitable for establishing models useful for predicting the effects on individual behaviour of changes in the economy or policy regimes. And so, it fails to fill the theoretical vacuum left by the rational choice theories.[19]

All in all, the marriage of the rational choice theory with econometrics fails to furnish models suitable for predicting the effects of change on behaviour. The key to this objective is the ability to address counterfactual queries such as those stated above. Addressing such queries demands a theory of behaviour that endogenises small worlds, beliefs, and preferences. In other words, it necessitates

[19] The reasoning here is an adaptation of Lucas' critique of econometric policy evaluation (1976).

a theory that explains how a person forms views about the causal structure of the economy, updates his views in light of new information, adapts preferences on the basis of past experiences, and accordingly defines his decision problem. If such a theory is established, there remains no *essential* role for the subjective expected utility theory in predicting and explaining behaviour. The theory, to use Suppes' words, becomes in one sense otiose:

> The psychologist resists accepting them [subjective probability and utility] as basic or primitive concepts of behaviour. Ideally, what he desires is a dynamic theory of the inherent or environmental factors determining the acquisition of a particular set of beliefs or values. If these factors can be identified and their theory developed, the concepts of probability and utility become otiose in one sense (1961:614).

It is appropriate to close this section by making two remarks. One concerns the generality of the analysis. Our argument for the inadequacy of Savage's theory draws on the fact that it gives no explanation of how the agent models his choice situation and defines his decision problem. In this respect, other rational choice theories are the same. They are also concerned with the final stage of decision, choice, and hence fail to serve as a theory of behaviour. The other remark relates to an implication of the analysis. Economists have long argued for the necessity of economic theory to specify explanatory variables in econometric models, the algebraic form of the model, the sign of the model parameters, and even the joint probability distribution of the variables being studied (Fair, 1987:270). And by economic theory, they mainly mean a theory of rational choice or a model based on it (Becker, 1976:5). The preceding analysis makes it evident that the rational choice theories do not provide any information useful for specification of econometric models; they just take them for granted. The so-called theoretical information in economics is simply disparate assumptions that are not derived from any systematic theory, certainly not from the rational choice theories (Peltzman, 1991:206). They are accepted because they intuitively sound plausible (Sims, 2004:282) or are part of a model that fits aggregate data.

6 Expectations

To understand the dynamics of behaviour it is essential to model both the process of preference and expectations (beliefs) formation. Notwithstanding this, economists have treated expectations and preferences differently. Stigler and Becker (1977) famously suggested that economics should not only take preferences as exogenous but also as homogenous across individuals, arguing that differences in actions are best explained in terms of differences in perceived opportunities (Vriend, 1996:279). Ever since, there have been some attempts to study preference formation but Stigler and Becker's view still dominates economics. In sharp contrast, a central position in economics has always been that economic theory cannot take expectations as exogenous (Harsanyi, 1965:450), and a variety of proposals have been set forth to model expectations. An influential proposal is the rational expectations hypothesis. We study some aspects of this hypothesis to further our understanding of the current state of economic theory.

6.1 Adaptive Expectations

The rational expectations (RE) hypothesis emerged as a result of reflection on the shortcomings of the so-called adaptive expectations (AE) hypothesis. According to this hypothesis, the agent considers only the recent values of a variable to form expectations of its future values, and, when the truth of his forecasts transpires, he uses his forecasting error to revise his future forecasts of the variable (Cagan, 1956). The AE hypothesis restricts relevant information on a variable to its recent history. As a consequence, it implies that people do not take note of changes in the economy until the effects of the changes are fed into their forecasting errors and, therefore, make systematic mistakes in perceiving the course of the economy (Bicchieri, 1987:506). Moreover, according to the hypothesis, the effect of interventions on behaviour begins to bear only after previous expectations badly go wrong. And, because of this strictly backward looking feature, the hypothesis logically rules out any immediate effect of policies on expectations and hence behaviour. These implications go against a well-entrenched conviction in economics that people optimally use all available information in making decisions. They realise, it is claimed, the interrelations among economic variables and utilise the information on their movements to form expectations. The AE hypothesis has thus been viewed as an inadequate conjecture about people.

6.2 Rational Expectations

The RE hypothesis is an extreme response to the backward looking feature of the AE hypothesis. In its strong form, it posits that economic agents know the true structural model of the economy and their subjective expectations of the variables representing the economy are the same as the objective expectations entailed by the true model (Pesaran, 1987:165):

> Expectations, since they are informed predictions of future events, are essentially the same as the predictions of relevant economic theory. At the risk of confusing this purely descriptive theory... with a pronouncement as to what firms ought to do, we call such expectations 'rational'. (Muth, 1961:316)

The RE hypothesis stands on several assumptions. An assumption is that the vector of exogenous and endogenous variables of the economy follows a *jointly* stationary stochastic process. Another assumption is that the variables have an objective joint probability distribution in the sense understood in the frequency interpretation of probability. In characterising this assumption, following Knight (1921), new classical economists divide uncertainty into 'reducible' and 'irreducible' components. Reducible uncertainty is defined as risk, which is the uncertainty that is analysable according to the laws of mathematical probability. Irreducible uncertainty is taken to be the 'true' uncertainty, which falls outside the bounds of numerical probability. The RE hypothesis is, by definition, restricted to risky situations (McCann, 1994:63). However, nothing is said about how it can be known whether a given situation is risky or truly uncertain, and so in practice the hypothesis is applied generally. A further assumption is that the agents correctly know the objective probability distribution of the variables describing the economy.[20] Finally, the agents are also assumed to know the true values of all the exogenous and endogenous variables through to the end of the present period.

These assumptions have strong implications for modelling of the economy. Since the agents know the true economic model, their forecasts are always confirmed by the course of events and their views are always consistent with each other. They therefore never have an incentive to revise their view of the economic structure. Moreover, since they maximise their expected utility with respect to the true model, they also never have an incentive to revise their actions. Their actions are always optimal with respect to the environment and with respect to the actions of other fellow agents in the economy. The economy is, therefore, permanently in equilibrium. Disequilibrium, by definition, becomes a vacuous notion, and all supposed disequilibrium phenomena are *a priori* defined out of existence. This last point plays a critical role in solving rational expectations models. These models are solved by requiring the collective outcomes of individual decisions to be an equilibrium state.

In order to better understand the hypothesis it is useful to look closely at how a rational expectations model is built and solved. To this end, we use a perfect foresight version of the *quantity theory* about the relation between money supply and prices.[21] Versions of this model are found in Blanchard and Watson (1982), Sargent (1993), and MacCallum (1983). The account here is based on Sargent (1993), who uses it to discuss the problem of multiple equilibria arising in rational expectations models. This is done in three steps:

Step 1. The economy runs in discrete time, and each individual lives for two periods. The same number of individuals, normalised to one, is born every period. An

[20] The RE hypothesis and subjective expected utility are reconciled through de Finetti's (1937) exchangeability result. Suppose there are repeated trials of some random process; and that individuals are indifferent between receiving a dollar conditional on some sequence of outcomes and receiving a dollar conditional on any other sequence of outcomes of each type; if there exist limiting frequencies of different types of outcomes, and individuals put strictly positive probability on the truth, then each individual's conditional beliefs converge to these limiting relative frequencies (Morris, 1995:232-233).

[21] Agents in a multi-agent economy is said to have *perfect foresight* if the following two conditions hold: (a) people's beliefs are correct and (b) there are no exogenous shock terms impinging on the economy, so that all expectations are correct without error, i.e., ($E_t\,(V_{t+k}) = V_{t+k}$).

individual born at time t is young at time t and old at time $t+1$. Each individual receives an endowment of $2e_1$ when young and $2e_2$ when old. The endowment is non-storable and the only way to save it is to hold money. Let p_t be the price level at time t and $E(p_{t+1})$ the value of p_{t+1} expected as of period t. The decision faced by the individual is to choose his level of nominal balances m_t to carry from time t to time $t+1$ so as to maximise the utility function:

$$\ln(2e_1 - m_t/p_t) + \ln(2e_2 + m_t/E(p_{t+1})). \tag{6.1}$$

The utility function describes how the agent is ready to forfeit m_t / p_t units of goods this period against $m_t / E(p_{t+1})$ units that he expects his real money balances will offer next period. The agent maximises the utility function (6.1) to decide on his nominal balances m_t, subject to the current price level p_t and his expectation of the next period price level, $E(p_{t+1})$. This yields the money demand function

$$m_t/p_t = e_1 - e_2 E(p_{t+1})/p_t. \tag{6.2}$$

Step 2. The laws of the variables entering the model are specified – here the money supply and price level. Suppose the government is supplying money in accordance to the rule

$$M_{t+1} = \alpha M_t. \tag{6.3}$$

It remains to specify the price function that is essential for estimating the expected future price level $E(p_{t+1})$. A trouble is that the expected future price level is among the factors affecting the price level and thus $E(p_{t+1})$ enters the true price function as an argument. The RE hypothesis suggests that the true price function is a function that ensures equilibrium. A method for finding such a function is to conjecture a price function, and check if it leads to equilibrium, which here means if it makes the demand for money m_t equal to its supply M_t.[22] A possible conjecture for the present economy is the following:

$$p_t = \beta M_t. \tag{6.4}$$

Since the agent, by assumption, knows the economic structure, he knows laws (6.3) and (6.4), including the parameters. He uses these laws to forecast the price level next period. It follows that

$$E(p_{t+1}) = \alpha\beta M_t \tag{6.5}$$

Step 3. Finally, (6.4) and (6.5) are substituted into (6.2) and the demand for money m_t is set equal to the money supply M_t. This yields the equilibrium price as,

$$p_t = (e_1 - \alpha e_2)^{-1} M_t. \tag{6.6}$$

The agent holds money if $m_t / E(p_{t+1})$ is greater than m_t / p_t, and stops giving up his endowment $2e_1$ if the two ratios are equal.

The RE hypothesis significantly reduces the complexity of predicting behaviour. According to the hypothesis, a person's maximisation behaviour is solely a function of his environment, preferences, and budget constraint. That is, given his preferences and budget constraint, he behaves in exactly the way that is objectively optimal with respect to the environment. As a result, for predicting behaviour, the economist has no need to study the person's beliefs about the economy or how he has arrived at those beliefs. He only needs to know the person's preferences, budget constraint, and the structure of the economy (Simon, 1990:6). And, issues of human learning and adaptation can be left to psychologists (Sargent, 1993:21). Furthermore, since in economics preferences are assumed to be homogenous across individuals, the RE hypothesis naturally leads to the

[22] This method is known as the method of undetermined coefficients. See Pesaran (1987:80-81) for alternative methods.

representative agent modelling approach, which, if suitable, enormously simplifies the study of macroeconomic phenomena.

6.3 Problems with the RE Hypothesis

The RE hypothesis has been one of the most influential proposals in modern economics, and influenced the views of economists on many aspects of policy analysis and inference from aggregate data. At the same time, like any bold conjecture, it has been the subject of bitter controversies. A full analysis of these controversies is beyond the scope of this chapter. Here, we only look at some of the theoretical debates that are directly related to the role of the hypothesis as a means for specifying people's view of the economy, a role Lucas assigned to the hypothesis (1981:223).

6.3.1 The True Model

A problem with the RE hypothesis concerns the notion of the 'true model'. There are certain situations where it clearly makes sense to speak of a true model. In computer simulations designed to investigate an estimation procedure, the modeller writes down a model, uses it to generate data, and studies whether the procedure can uncover the model from the data if the sample size is allowed to grow arbitrarily large. However, outside such situations, it is not clear what a true model means, particularly in macroeconomics where model construction heavily involves aggregation, idealisation, and simplification. As will be shown in the final chapter, aggregation over interactive heterogeneous units generates relations that are absent at the individual level, and more importantly as one varies the aggregation level, one encounters quite different models. What guides a modeller to decide at a specific level of aggregation are mainly pragmatic considerations, not correspondence between the model and reality, and this casts doubt on the notion of a true macroeconomic model. Moreover, even if the notion of a 'true model' were unproblematic, in macroeconomics the true model would be so complex that would be of no use for prediction or explanation of economic phenomena. These quandaries in making sense of a 'true' macroeconomic model and the difficulties in establishing it reduce the RE hypothesis to the idea that the agent's model of the economy coincides with whatever model the economist accepts to describe the economy (Bullard, 1994). The question then arises as to whose model really reflects the people's view of the economy. A possible response is to search for a model that best fits aggregate data. This, however, takes us back to where we started the search for microfoundations. Many models can fit the data equally well, and the greatest challenge is to determine which model best approximates the economy.

6.3.2 Multiple Equilibria

A step in building a rational expectations model is to conjecture the mechanisms or laws governing the variables representing the economy, such as money supply in the above example. These conjectures are necessary for specifying people's beliefs about the economy. To explain a difficulty with this, it is crucial to bring to the fore a distinction between two types of variables entering a model. First, there are variables whose values do not depend on their own expected value. One such variable is weather that often enters into agricultural models such as wheat production models. The state of weather over the next few years does not depend on people's expectations about future weather. As opposed to such variables, the model may contain variables whose values depend on people's expectations of the future values of the variables. The price of a commodity at time *t* can depend on people's expectations of prices at time *t+1*. This means the way people form expectations about prices is part of the mechanism determining prices. In these cases, the RE hypothesis requires people's expectations to be *consistent* with each other so that the economy is in equilibrium. In the present example, this means that people's expectations of future prices are such that they make the demand and supply of money equal (the market clears). However, this consistency requirement is not enough to ensure a unique solution for rational expectation models with expected endogenous variables. Many expectations formation rules yield consistent expectations, raising the question of which rule is true of the economy. An alternative mechanism for the price level in the above economy is (Sargent, 1993:11):

$$p_t = \beta M_t + \lambda' c \tag{6.7}$$

Like the forecasting rule (6.3), this rule also clears the market. In fact, for every $c > 0$, there is an equilibrium price.[23] Due to this multiplicity, a complete description of the fundamentals of the economy (i.e., tastes, technology, and initial resources endowments) and the condition of belief consistency across individuals are not sufficient for predicting the equilibrium price. It is also essential to know how people converge on a particular expectation formation rule. Contrary to Lucas's initial expectation, the RE hypothesis falls short of specifying people's beliefs about the future of the economy.[24]

6.3.3 A Paradox

The problem of multiple equilibria is well known in the economic literature. It has also been a major research problem in game theory. In addition to this problem, there are other issues with the RE hypothesis that are less known. Recall the hypothesis implies that the vector of exogenous and endogenous variables of the economy follow a jointly *stationary* stochastic process. It also implies that people's subjective expectations of the variables coincide with the expectations implied by the joint objective probability distribution of the variables. Altogether, these assumptions exclude the possibility of discretionary policy interventions. This follows from the fact that if there were some free parameters that could be controlled by public officials there would be, according to the hypothesis, an objective probability distribution for the parameters that were known to the people. In that case, people would already know the likelihood of any variation in the parameters. As a result, they would have taken the information into account when making their future decisions. And so, the likelihood of any change by a policy maker would have already been known to the people and would have already been fed into their behaviour. This means there can be no discretionary policy

[23] The equilibrium price is now determined by $p_t = (e_1 - \alpha e_2)^{-1} M_t + (e_1 / e_2)' c$.

[24] Economists have introduced extra principles to select a unique equilibrium. A proposal is due to MacCallum (1983:144) which seeks to block introduction of 'extraneous' terms such as c into forecasting rules. Such suggestions have turned out to be inadequate. They also lack a behavioural justification (Lucas, 1986).

intervention under the RE hypothesis (Bicchieri, 1987:510; Vercelli, 1991:150). To allow for policy interventions, the assumption that the economy is permanently in a stationary state must be relinquished. This requires abandoning the RE hypothesis.[25]

6.3.4 The Peril of Redundancy

There is another related paradoxical implication of the RE hypothesis that is worth noting. The hypothesis, as just said, implies that the economic environment is stationary and so excludes the possibility of policy interventions. Granting the hypothesis, then, the only practical objective of macroeconomics that remains possible is *ex ante* and *ex post* predictions. Such predictions do not require a structural model built on the optimal rules of individual behaviour. A regression model closely representing the relations among relevant aggregate variables is enough. Therefore, with the impossibility of policy interventions, there is no practical necessity to model expectations and, for that reason, there remains no direct role for the hypothesis in economic modelling. The RE hypothesis implies its own practical redundancy. Sims notes this quandary at the heart of Lucas' program (1982:115-16). He seems to argue that, having assumed stationarity, Muth should have excluded expected variables from the realm of large-scale economic modelling altogether rather than requiring macroeconomic models to be built on an expectation formation mechanism. In a stationary environment, a vector autoregression model tracking the past movements of relevant aggregate variables suffices for the purpose of economic analysis (Sims, 1982:115-16). In an interesting comparison of Lucas and Sims' approaches to macroeconomic modelling, Sargent also acknowledges that the RE hypothesis, taken seriously, can be equally used to "to support Sims' style of more or less uninterpreted vector autoregressive empirical work" (Sargent, 1984:408).

[25] Some new classical economists have acknowledged that the RE hypothesis, literally understood, contradicts with the possibility of policy interventions. Aware of this contradiction, Sargent writes: "In formal work, this contradiction is evaded by regarding analyses of policy interventions as descriptions of different economies, defined on different probability spaces. The mental comparison is among economies identical with respect to private agents' preferences and technologies, but differing in government policy regime" (1984:413). This move raises more questions than solves. Basically, it is not clear how the agents in the economy governed by the existing policy regime come to know the joint distribution of the variables characterising the economy governed by the new regime.

6.3.5 The No-Trade Theorems

The RE hypothesis has also contributed to the emergence of a class of no-trade theorems that are in sharp conflict with observed data (Milgrom et al., 1982). In economics, preferences are taken to be homogenous across individuals. This assumption, joined with the hypothesis, implies a view of the economy as a society of identical individuals. Such an economy provides no place for security markets. Security markets exist because people have diverse information, think of the economy differently, and have heterogeneous preferences (Arrow, 1986:212). Even though rejecting homogenous preferences is enough for eliminating the theorems, it is equally plausible to reject the RE hypothesis to explain the emergence of security markets.

This ends our evaluation of the RE hypothesis. The hypothesis is a bold attempt to specify people's view of their choice situation simply by studying the environment they live in. To achieve this objective, it is assumed that people have already learnt the structure of the economy, adapted their optimal rules of behaviour, and the economy is in equilibrium (Lucas, 1986). These suppositions are strong. Nevertheless, even when they are supplemented with complete knowledge of the fundamentals of the economy, they are still inadequate for prediction of economic outcomes, due to the multiple equilibria problem. The marriage of the subjective expected utility theory with the RE hypothesis fails to provide a predictive economic theory. Predicting whether as a result of an intervention the economy converges to equilibrium and the equilibrium at which it settles down calls for a theory of how people structure their choice situation, re-define it as a result of a policy change, and adapt their behaviour as a result of subsequent experiences. Until an adequate theory explaining how people learn about the economy and adapt is found, macroeconomic theory cannot hope to produce the policy predictions that are its ultimate goal (Bicchieri, 1987:512).

7 Conclusion

New classical economists have proposed two hypotheses to derive generalisations of individual behaviour: the rational choice hypothesis and the rational expectations hypothesis. The claim is that these hypotheses furnish the basic elements of a theory that specifies the variables relevant to explaining economic phenomena, draws a distinction between exogenous and endogenous aggregate variables, specifies the algebraic form of the functions linking the aggregates, suggests their joint probability distribution, and more importantly characterizes the conditions under which an observed regularity at the aggregate level remains invariant.

This chapter studied the contributions of these rationality hypotheses to development of a theory of economic behaviour. Rational choice theories were examined using Savage's theory of subjective expected utility and two types of issues were distinguished. One was whether the postulates of the theory were true. The other was whether the theory, regardless of the truth of its assumptions, was adequate for explaining and predicting behaviour, particularly in dynamic, evolving situations.

On the first issue, it was argued that Savage's postulates were predicated on two more basic assumptions that preferences are fixed, and ready-made, and that there always exists a description of the world that allows complete disentanglement of values from beliefs. Drawing on the lessons of experimental psychology, we argued that, like beliefs and small worlds, preferences are not read off from a master list but constructed. As a result, since different constructions of beliefs, small worlds and preferences can systematically lead to different choices, prediction and explanation of behaviour in a dynamic situation demands a theory that explains the *process* of preference, belief, and small world formation. We also argued that a description of the world, allowing complete disentanglement of beliefs and values, was hard to find, and even if it existed, it would be so cumbersome to be of any use in guiding decisions.

On the second issue, our central point was that these theories take the structure of the small world, likelihood judgements, and preferences as given, and only state how an ideal agent solves an already well-structured decision problem. As a result, in predicting behaviour, a very large list of substantive assumptions is needed to specify the agent's view of his choice situation and the decision problem he is trying to solve. Yet, the nature of these assumptions is such that they involve the answer to the very same question a theory of behaviour is expected to answer. In fact, by varying the exogenous assumptions entering a rational choice model any side of any substantive controversy can be rationalized. For this reason, the rational choice models answer no substantive economic question; they only repackage what has already been stated in the assumptions. In practice, economists have tried to select a rational choice model based on econometric analysis of aggregate data. But the econometric approach is unsuitable for resolving questions of individual behaviour.

Moreover, for evaluating the outcomes of novel policy interventions, one needs to predict how the agents would react to the policy, and this requires predicting how in response to the policy they modify their view of their choice situation and redefine their decision problem. These queries entirely fall outside the scope of the rational choice theories, which take as given the structure of the choice situation. Contrary to common belief, the critical difficulty with the rational choice theories is not that they are false. It is that they in principle have very little to contribute to economic theorising.

Finally, the chapter studied the RE hypothesis, which is an attempt to specify people's view of the economy without studying how they learn about it. Economic decisions usually involve expectations of endogenous variables, such as prices. In such cases, the hypothesis is reduced to the condition of belief consistency across individuals. However, there are always many ways in which people's beliefs can be consistent with each other. The hypothesis falls short of specifying people's view of the economy.

These remarks demonstrate that understanding economic behaviour requires a different type of theory of behaviour. It requires a theory that explains how people form preferences, learn about the economy, model their choice situation, define their decision problem, and redefine it as new information arrive. In a nutshell, economics needs a *learning-based* (adaptive) theory of behaviour, not a choice-based theory.

Appendices

Appendix A: Homothetic Utility Function

A monotone preference relation $\geq$ on a choice set $\mathbf{X} \subseteq \mathbf{R}_+^L$ is called *homothetic* just in case $\mathbf{x} \geq \mathbf{y} \Leftrightarrow \alpha\mathbf{x} \geq \alpha\mathbf{y}$ for all $\alpha > 0$. Homothetic preferences can be represented by a monotonic transformation of a homogenous of degree 1 utility function. Informally, homothetic preferences mean that the agent always spends a fixed proportion of his or her income on each good.

Appendix B: Satisficing

Satisficing is a choice procedure. Following Rubinstein (1998:12), let *A* be some 'grand' set of options (or the set of all possible options), *O* an ordering of the set *A*, and $\mathbf{S} \subseteq \mathbf{A}$ the set of satisfactory alternatives. For any choice problem *C*, satisficing involves sequentially examining the alternatives in *A* according to the ordering *O*, until an alternative is found that is a member of *S*.

Appendix C: Fair's Voting Equation

As in the text, let V_{it} be a variable that is equal to one if voter *i* votes for the Democratic candidate in period *t* and zero if he votes for the Republican candidate. Also, let

$$\psi_i = \xi_i^r - \xi_i^d, \tag{C1}$$

$$q_t = \beta_1 \frac{M_{td1}}{(1+\rho)^{t-td1}} + \beta_2 \frac{M_{td2}}{(1+\rho)^{t-td2}} - \beta_3 \frac{M_{td1}}{(1+\rho)^{t-tr1}} - \beta_4 \frac{M_{tr1}}{(1+\rho)^{t-tr2}}. \tag{C2}$$

The expected utility theory implies that:

$$V_{it} = \begin{cases} 1 & if \quad q_t > \psi_i \\ 0 & if \quad q_t < \psi_i \end{cases},$$

which means the voter votes for the Democratic candidate if $\psi_i < q_t$. Now, recall the aggregation assumptions (A5) and (A6), restated here as:

A₅: ψ_i is evenly distributed across voters in each election between some numbers $a + \delta_t$ and $b + \delta_t$, where $a < 0$ and $b > 0$. *a* and *b* are constant but δ_t can vary across elections.

A₆: There are an infinite number of voters in each election.

These assumptions imply that ψ is uniformly distributed between $a+\delta_t$ and $b+\delta_t$, where the subscription is now dropped from ψ_i. The probability density function for ψ, denoted by $f_t(\psi)$ is

$$f_t(\psi)=\begin{cases}\dfrac{1}{b-a} & for\, a+\delta_t<\psi<b+\delta_t \\ 0 & otherwise\end{cases} \tag{C3}$$

and the cumulative distribution function for ψ, denoted as $F_t(\psi)$, is

$$F_t(\psi)=\begin{cases}0 & \psi<a+\delta_t \\ \dfrac{\psi-a-\delta_t}{b-a} & a+\delta_t<\psi<b+\delta_t\,. \\ 1 & \psi>b+\delta_t\end{cases} \tag{C4}$$

Because of δ_t, the probability density and distribution functions are different for each election. Let V_t denote the percentage of the vote that goes to the Democratic candidate in election *t*. Since a person votes for the Democrat candidate if $\psi_i<q_t$, the probability that he votes for the Democrat candidate is $p(\psi<q_t)$. The proportion of voters voting for the Democrat candidate in election *t* is $np(\psi<q_t)/n=p(\psi<q_t)$, which means V_t is equal to the probability that ψ is less than or equal to q_t. Since the probability density function of ψ is given by (C3), V_t is equal to $F_t(q_t)$. Using (C4), V_t can be stated as:

$$V_t=\frac{-a}{b-a}+\frac{q_t}{b-a}-\frac{\delta_t}{b-a}\,. \tag{C5}$$

Substituting q_t in (C5) yields

$$V_t=\alpha_0+\beta_1^*\frac{M_{td1}}{(1+\rho)^{t-td1}}+\beta_2^*\frac{M_{td2}}{(1+\rho)^{t-td2}}-\beta_3^*\frac{M_{td1}}{(1+\rho)^{t-tr1}}-\beta_4^*\frac{M_{tr1}}{(1+\rho)^{t-tr2}}+v_t \tag{C6}$$

where

$\alpha_0=-a/(b-a)$; $\beta_3^*=\beta_3/(b-a)$;

$\beta_1^*=\beta_1/(b-a)$; $\beta_4^*=\beta_4/(b-a)$;

$\beta_2^*=\beta_2/(b-a)$; $v_t=-\delta_t/(b-a)$.

Chapter 3

Homo Economicus as an Intuitive Statistician (1)

Model Free Learning

1 Introduction

> This is our key bounded rationality assumption: we back away from the rational expectations assumption, replacing it with the assumption that, in forecasting prices, firms act like econometricians (Evans et al., 2001:28).

The preceding chapter argued that the subjective expected utility theory is simply a method for solving an already well-structured decision problem. However, prediction of behaviour, particularly in dynamic situations, requires a theory that explains how the agent models his choice situation and defines his decision problem. Therefore, even if true, the subjective expected utility theory is inadequate as a theory of economic behaviour. New classical economics have set forth the rational expectations (RE) hypothesis as a way of specifying the agent's view of his choice situation (or the economy in general). The hypothesis identifies the agent's subjective expectations with the mathematical expectations implied by the true economic model, suggesting that he maximises his expected utility with respect to the true model. Accordingly, the new classical paradigm defines economics as the enterprise to derive observable economic phenomena from two basic hypotheses: (1) people are (subjective) expected utility maximisers and (2) they maximise their expected utility with respect to the true economic model.

Attempts to overcome the theoretical shortcomings of the RE hypothesis described in the last chapter have resulted in the re-emergence of the bounded rationality project, originally proposed by Herbert Simon (1955 and 1956). While there has been a burst of interest in the topic over the last two decades or more, there is no consensus yet on the definition of bounded rationality or what the critical questions of the project are (Rubinstein, 1998). The goal of the project is to replace the behavioural assumptions of economics with more realistic assumptions and investigate the implications of the changes for our understanding of the economy (Conslik, 1996). So, depending on what behavioural assumptions of economics are withdrawn and what assumptions are retained, various notions of bounded rationality can be defined. Most studies of bounded rationality in new classical economics retain the principle of subjective expected utility maximization but replace the RE hypothesis with the assumption that the agent constructs a model from the available economic data, which may not coincide with the true model. Thus, in new classical economics, the project of bounded rationality is a program to derive observable economic phenomena from the general principles that: (1) the agents are subjective expected utility maximisers and (2) they maximise their expected utility with respect to models constructed from the available economic data.[1]

From this perspective, the primary issue of the bounded rationality program is to theorise how the agent learns about the economy and models his choice situation. There are several proposals on offer. The conjecture that has received most attention is that the *homo economicus* is an *intuitive* statistician; i.e., he intuitively models the economy like a statistician (Arthur, 1996:4). Thomas Sargent, a leading economist from the new classical camp, summarises this view of the bounded rationality program as follows:

> I interpret a proposal to build models with 'boundedly rational' agents as a call to retreat from the second piece of rational expectations (mutual consistency of perceptions) by expelling rational agents from our model environments and replacing them with 'artificially intelligent' agents who behave like econometricians. These 'econometricians' theorise, estimate, and adapt in attempting to learn about probability distributions which, under rational expectations, they already know (Sargent, 1993:3).

This conjecture will be called the *intuitive statistician* (IS) hypothesis of bounded rationality. A pioneering work on this view of bounded rationality is Bray (1982), who considers an economy in which the agents know the correct model up to a small number of parameters and use the least squares method to estimate the unknown parameters. Letting the agents live indefinitely, she investigates if they ever learn the true parameters, which is essential for forming rational expectations. The significance of this question lies in the fact that the learning problem facing the agents in Bray's model economy is not identical with ordinary parameter

[1] There is a fundamental difference between the considerations that led Herbert Simon to propose his bounded rationality program and those that led new classical economists to study it. Herbert Simon initiated his project as an alternative to neoclassical economics. Whereas new classical economists have began studying bounded rationality to provide adequate foundations for the rational expectations hypothesis, and to extend the new classical tools and notions to phenomena traditionally unaccountable within the paradigm. Besides the differences in motivations, Simon rejects both elements of the new classical economics.

estimation. As the agents learn about the economy, they modify their expectations and behaviour, which in turn alter the relations being learnt. It is not then possible to use the textbook convergence theorems on the long-run behaviour of the least squares estimator to argue that the agents will asymptotically learn the truth. The question addressed by Bray is different. Her objective is to examine the conditions under which her model economy converges to rational expectations equilibrium, even though feedback from learning can change the relations being learnt. Since Bray's publication, a sizeable number of similar studies have emerged. Bray (1983), Honkapohja (1995), Marrimon (1997), Williamson (1997), Salmon and Kirman (1995), Evans and Honkapohja (2001), Sargent (1993), and Sobel (2000) contain original contributions as well as surveys of the literature on learning in economics.[2]

The relevance of these theoretical studies is unclear for several reasons. These studies usually assume that the agents already know the correct unestimated model of the economy, without any explanation of how the model was learnt in the first place (Sargent, 1993:166; Sobel, 2000:256).[3] The assumption that the agents know the correct model is crucial, since starting with a wrong model can make learning of rational expectations impossible (Nyarko, 1991). Therefore, the convergence results established are contingent on the model economies being studied; they do not generally hold. Furthermore, the results are invariably of an asymptotic nature. But what is needed for evaluating policies are short run predictions of how agents would revise their view of the economy in response to a policy change, redefine their choice situation, and modify their behaviour. As Keynes put it, in the long run we are all dead. Finally, the dynamics of the economy in these studies come exclusively from people's adjustments of their behaviour. However, the economic structure can change for reasons other than feedback from learning, entirely altering the inference problem facing the agent.

[2] The proposal that human mind acts like an intuitive statistician has a long history in cognitive psychology. An interesting discussion of the proposal is found in Cheng and Holyoak (1995), who focus on how people, like statisticians, learn about the causal structure of their environment. The hypothesis also occupies a central place in Shanks (1995)'s monograph on the psychology of learning.

[3] In Bray (1982), agents are assumed to know the supply curve and must only form price expectations to plug into it.

The possibility of convergence to rational expectations equilibrium is not directly a concern of this thesis. The concern is to investigate if the IS hypothesis helps us understand and predict how the agent models his choice situation and defines his decision problem. A positive response to this query presumes that there is a 'tight enough' theory of statistical (scientific) inference, describing how statisticians learn about the world, turn economic data into a model of the economy, and revise the model in the face of new information (Sargent, 1993:23).[4] Otherwise, the IS hypothesis would not be of much help in predicting how the statistician (and thus the agent) models his choice situation. And *a fortiori*, no general conclusion could be derived from the hypothesis about the conditions under which an economy converges to equilibrium.

Therefore, a major concern of this thesis is whether there is a 'tight enough' theory of statistical inference. To clarify the query, it is useful to start with a preliminary conjecture about how a statistician models a choice situation. In statistics, the environment is perceived through a collection of measurable quantities (features), which are conceived as realizations of some random variables with an unknown joint probability distribution. The statistician first uses the data on these quantities to estimate their joint probability distribution. He next uses the estimate of the joint distribution to uncover the causal relations among the variables. If the resulting model is inadequate, the initial set of variables is modified, and the two phases of inference repeated.

This description of the processes of statistical inference is imprecise. Nevertheless, it helps us in separating issues relating to inference about probabilities from issues relating to inference about causes, and provides a framework for defining certain important questions about the possibility of establishing a precise theory of statistical learning. This chapter and the one to follow examine some basic issues relating to learning of the joint probability distribution of a set of variables describing a choice situation. The fifth chapter investigates if there can be a theory that tells us how to move form the joint probability distribution of a set of variables to the causal structure linking the variables.

[4] Sargent (1993) raises this question, hoping that it has a positive answer.

Several approaches to statistical inference are on offer. The diversity arises partly from philosophical disagreements about the nature of probability and partly from alternative methodologies that, given an interpretation of probability, can be adopted to solve inference issues. The debates about the nature of probability are not crucially related to the issue of whether there exists a 'tight enough' theory of statistical inference, and will not be taken up here. Instead, two general methodological approaches to statistical modelling are studied, based on the frequency and subjective interpretations of probability respectively. An analysis of these approaches provides an adequate ground for judging if there can be a 'tight enough' theory of statistical inference, which is essential for assessing the bounded rationality program as defined here.

The current chapter investigates the possibility of a 'tight enough' theory of statistical learning by looking at nonparametric statistics, which is a branch of statistics that avoids restrictive non-sample, probabilistic, assumptions, and seeks to leave model discovery to the data.[5] We use this framework to investigate two queries. The first query is whether it is possible with a reasonably sized sample to obtain a good approximation of the joint probability distribution of several variables using nonparametric estimators, or whether substantial non-sample information is required to achieve this. The second query is whether there exist inferential procedures that receive observations on a set of variables and yield the best estimate of the underlying joint probability distribution, which is possible given the data. If not, statistical model discovery cannot be left entirely to the data, which raises the question of where statistical models come from. Both issues are clearly important for the bounded rationality program.

The rest of this chapter is organised as follows: Section 2 defines the notion of a statistical model, model specification, and some key problems in statistical inference. Section 3 describes the basic idea of nonparametric inference. Section 4 states the IS hypothesis within the framework of nonparametric statistics, relating

[5] The terms parametric and nonparametric, as will be explained shortly, are used to distinguish those inference problems in which the regression or density function is known up to a finite number of parameters from those in which the algebraic form of the function is unknown, and thus the inference problem involves more than ordinary parameter estimation (Manski, 1991:34).

the statement to the literature in economics. Section 5 studies whether it is possible with a reasonably sized sample to obtain a good approximation of the joint probability distribution of several variables using nonparametric estimators. Section 6 examines whether there exist inferential procedures that receive observations on a set of variables and yield the best estimate of the underlying joint probability distribution, which is possible given the data. Section 7 concludes the chapter by stating some of the implications of the analysis for the bounded rationality project.

2 Statistical Model Specification

As said earlier, in statistics, the environment (choice situation) is perceived in terms of a collection of measurable quantities, some of which are known and some of which are not known. The quantities are considered as realizations of random variables with some unknown joint probability distribution. The goal of statistical inference is to determine the values of the unknown quantities from the known quantities, which in theory requires modelling the joint distribution of the random variables. So, an appropriate point of departure for our study is to disentangle problems that arise in modelling the joint distribution of a set of variables, give a precise definition of a statistical model, and highlight the basic issues that a theory of statistical learning has to explain.[6]

A problem in model building, which in a sense precedes any statistical inference, concerns the choice of variables that characterise the environment. Two forms of variable selection should be separated. Sometimes the objective in building a model is to generate accurate *ex ante* and *ex post* predictions of a response variable *Y*. In that case, variable selection requires specifying some variables that are systematically related to *Y*, and there is no need for them to be the causes of *Y*. Alternatively, if the goal is to use the model to analyse the effect of changes in the environment on *Y*, variable selection requires finding the causes of *Y*. In either case, variable selection poses difficult questions that must, at least tentatively, be

[6] The discussion in this section builds on the works of Granger (1990; 1999), Lindley (1982), and Spanos (1986, 1999, and 2001). The definition of a statistical model to follow is adapted from Spanos.

solved before being able to construct a useful model. The problems in mind concern defining the appropriate form of the variables, finding the right method of measurement, deciding on the correct level of aggregation, and so fourth. The emphasis in this thesis is on the variable selection problem in the second sense. A solution to this problem calls for a theory of causal inference, which is taken up in the fifth chapter. For now, we assume that the relevant variables are known, and concentrate on issues relating to learning probabilities.[7] We proceed by defining various issues that arise in modelling the joint distribution of a set of variables.

Let us start with the simplest case where there is only one variable of interest. Specifically, let Z_t denote the variable of interest and $\mathbf{D} = \{z_1, z_2, ..., z_{T-1}\}$ be the past values of Z_t. The aim is to predict the future values of Z_t from the known values in *D*. This requires estimating the joint probability distribution of $\mathbf{Z} = \{Z_1, Z_2, ..., Z_T\}$, which we denote it by $p(Z_1, Z_2, ..., Z_T, \Theta)$ or simply $p(\mathbf{Z}, \Theta)$, where Θ is a parameter space defining the distribution. However, the problem of inferring $p(\mathbf{Z}, \Theta)$ from the data *alone* is ill-posed, since it has no unique solution regardless of the size of the sample *D*.[8] To show this, note that using sequential conditioning, the joint distribution $p(\mathbf{Z}, \Theta)$ can be decomposed into a product of univariate marginal and conditional distributions:

$$p(\mathbf{Z}, \Theta) = p(Z_1 / \Theta_1) \prod_{t=2}^{T} p(Z_t / z_{t-1}, ..., z_1, \Theta_t), \quad \text{for all } \mathbf{z} \in R_{\mathbf{Z}}^{T}. \tag{2.1}$$

For each sample size *T*, the conditional distribution $p(Z_T / z_{T-1}, ..., z_1, \Theta_T)$ involves *T*-1 conditioning variables. Therefore, with each increase in the sample size, the conditional distribution for Z_T changes, making it impossible to infer $p(\mathbf{Z}, \Theta)$ from the data no matter how large the sample grows. Spanos terms this phenomenon the *increasing conditioning set* problem (1999:266).

[7] Granger (1999, chapter 1) touches on some of the difficulties arising at this stage of specification analysis.

[8] The philosophical point to be made in this section may be well known but the aim is to precisely define a statistical model, which is essential for the analysis to follow.

Furthermore, the notion of conditional density is defined only for specific values of the conditioning variables. Thus, for each $\mathbf{z} \in R_{\mathbf{Z}}^{T}$, estimating $p(\mathbf{Z}, \Theta)$ involves estimating one marginal and *T-1* different conditional distributions. This is impossible since the number of distributions (or in other words, parameters) to be estimated always exceeds the sample size. Spanos calls this phenomenon the *stochastic conditioning* or *heterogeneity* problem (1999:267).[9] It is therefore necessary to introduce certain simplifying assumptions to make any inference about the target distribution $p(\mathbf{Z}, \Theta)$.

To explain the kind of assumptions necessary for inference from data, note that the increasing conditioning set problem arises because Z_t is allowed to depend on the whole past history of the stochastic process. This suggests that the problem can be circumvented by restricting the dependence of Z_t on its past. To illustrate, one possibility is to assume that Z_t is *completely* independent of its past. Complete independence reduces the joint distribution $p(\mathbf{Z}, \Theta)$ into a product of univariate distributions:

$$p(\mathbf{Z}, \Theta) = \prod_{t=1}^{T} p(Z_t / \Theta_t), \qquad \text{for all } \mathbf{z} \in R_{\mathbf{Z}}^{T}. \tag{2.2}$$

Another possibility is to assume that Z_t conditional on its immediate past Z_{t-1} is independent of the rest of the history of the process. This independence assumption, called the first order Markov condition, simplifies (2.1) into,

$$p(\mathbf{Z} / \Theta) = p(Z_1 / \Theta_1) \prod_{t=2}^{T} p(Z_t / z_{t-1}, \Theta_t), \qquad \text{for all } \mathbf{z} \in R_{\mathbf{Z}}^{T}. \tag{2.3}$$

In any case, inference about $p(\mathbf{Z}, \Theta)$ requires some independence assumption (restriction) to sever the tie between the conditional distribution $p(Z_t / z_{t-1}, \dots, z_1, \Theta_t)$ and the sample size.

[9] For a concrete example see Spanos (1999:263-267).

The stochastic conditioning problem arises because the conditional densities $p(Z_t / z_{t-1},...,z_1,\Theta_t)$ are allowed to vary for each possible $\{z_{t-1},...,z_1\} \in R^{T-1}$. The only way to deal with the problem is to impose some homogeneity restriction across the conditional densities $p(Z_t / z_{t-1},...,z_1,\Theta_t)$ defined over all possible values $\mathbf{z} \in R_{\mathbf{Z}}^T$. The strongest form of homogeneity is *complete* homogeneity, which takes the conditional densities $p(Z_t / z_{t-1},...,z_1,\Theta_t)$ defined over all $\mathbf{z} \in R_{\mathbf{Z}}^T$ to be the same. Complete homogeneity renders the indices in Θ_t, which distinguish different densities $p(Z_t / z_{t-1},...,z_1,\Theta_t)$, redundant, simplifying (2.2) to:

$$p(\mathbf{Z},\Theta) = \prod_{t=1}^{T} p(Z_t /\Theta) \ , \qquad \text{for all } \mathbf{z} \in R_{\mathbf{Z}}^T . \qquad (2.4)$$

A set of random variables $\{Z_1, Z_2,...,Z_T\}$, which is completely independent and homogeneous, is called a random sample, or an independently and identically distributed (IID) sample. An alternative concept of homogeneity, which will be used later in the thesis, is *strict stationarity.* The stochastic process $\{Z_t, t \in T\}$ is said to be strictly stationary if

$$p(Z_{t_1}, Z_{t_2},...,Z_{t_n};\theta) = p(Z_{t_1+\tau}, Z_{t_2+\tau},...,Z_{t_n+\tau};\theta), \qquad \text{for any } \tau,\ (t_i+\tau) \in T \quad (2.5)$$

i.e., the joint distribution remains unchanged when each point 1,2,...,T is shifted by a constant τ. When n is equal 1, strict stationarity is reduced to complete homogeneity.

These two types of assumptions, although necessary, are still insufficient to transform the problem of inferring $p(\mathbf{Z},\Theta)$ from data into a well-posed problem, i.e., a problem with a unique solution given the data. With a finite sample, it is also necessary to restrict *a priori* the class of density functions to which $p(Z_t,\Theta)$

may belong to a class F smaller than the class of all possible density functions.[10] The proposed distribution family must be small enough to warrant a unique solution. The distributional hypothesis allows restating (2.4) as,

$$p(\mathbf{Z},\Theta)=\prod_{t=1}^{T} f(Z_t / \theta), \qquad \text{for all } \mathbf{z}\in R_{\mathbf{Z}}^{T}. \tag{2.6}$$

The independence, homogeneity and distributional assumptions reduce the problem of inferring $p(\mathbf{Z},\Theta)$ from data into finding a distribution $f(Z_t / \theta)$ from the restricted distribution family F that best fits the data. If the non-sample assumptions are appropriate, and if the sample size is adequately large, then $f(Z_t / \theta)$ can be reliably estimated form the data.

In light of this analysis, we may define a statistical model as a set of three assumptions drawn from the three categories of independence (I), homogeneity (H), and distribution (D) (Spanos, 2000:239). A more precise of definition calls for some further remarks about these assumptions:

A key remark is that these assumptions are *basic*. That is, once we decide on the form of the assumptions for a vector of observables **Z**, no additional assumption is needed to specify the marginal and conditional distributions of the variables in **Z**, the algebraic form of the regression function of any of the variables on the others, or the distribution of the error terms. All these are determined by the three assumptions made about **Z**. As a simple illustration, consider a bivariate random variable $\mathbf{Z}_t=(X_t,Y_t)$, with data being $\mathbf{D}=\{(x_t,y_t)\}_{t=1}^{N}$. Further, suppose $\mathbf{Z}_t$ is randomly distributed and has a bivariate normal distribution, giving the model:

Bivariate Normal Model

A1: Data Distribution: $\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}\begin{pmatrix} \sigma_Y^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_X^2 \end{pmatrix}\right)$

A2: Independence: $(\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_N)$ is C-Independent

A3: Homogeneity: $(\mathbf{Z}_1,\mathbf{Z}_2,\ldots,\mathbf{Z}_N)$ is C-Homogeneous.

[10] The reason is that even estimating a univariate distribution from a random sample involves estimating infinitely many parameters, which is impossible with a finite sample.

The model fully defines the marginal distribution of *X*, the conditional distribution of *X* given *Y*, the marginal distribution of *Y*, and the conditional distribution of *Y* given *X*. It also determines the algebraic form of the regression function of *Y* on *X*, and *X* on *Y*. If *Y* is the response variable, the model implies (Spanos, 1986, ch.22):

$$\begin{aligned}
&X \sim N(\mu_x, \sigma_x^2)\\
&(Y / X = x) \sim N(\mu^*, \sigma^2)\\
&\mu^* = E(Y / X = x) = \beta_0 + \beta_1 x\\
&\beta_0 = \mu_y - \beta_1 \mu_x;\ \beta_1 = \sigma_{xy} / \sigma_x^2;\ \sigma^2 = \sigma_y^2 - \sigma_{xy}^2 / \sigma_x^2
\end{aligned}$$

A second key remark is that these assumptions cannot be combined arbitrarily. An assumption from one of these categories can restrict possible choices from the other categories. For example, the choice of a first order Markov condition for $\mathbf{Z} = (X_t, Y_t)$ and a bivariate normal distribution assumption are not compatible. The assumption necessitates a multivariate distribution. A final remark is that all these assumptions are of a probabilistic nature; all have to do with the distribution of the observables.

A statistical model can thus be redefined as a set of *internally consistent* probabilistic hypotheses drawn from the three categories of independence, homogeneity and distribution (Spanos, 2001). From this perspective, statistical model specification involves positing *a priori* appropriate independence, homogeneity and distribution assumptions to make inference about the target model (distribution) possible.

To recapitulate the points so far, any inference from data necessarily calls for three types of assumptions – a model. In theory, once these assumptions are introduced, the inference problem is reduced to parameter estimation, for which there are usually routine procedures. So, the most fundamental and challenging aspect of inference (learning) from data consists in model specification. And, as a result, the most immediate task facing a theory of statistical inference (learning) is to explain where the models come from, and how to go about selecting the three basic assumptions in any inference problem.

3 Nonparametric Statistical Inference

Two possible responses to these queries are found in the statistical literature. This chapter analyses a response found in *nonparametric* statistics. The concern in this branch of statistics has mostly been with estimating a density (regression) function from a random sample, and less attention has been paid to inference from non-random samples, which requires deciding on an appropriate independence and homogeneity assumption. We begin by assuming a random sample to spell out the core idea of nonparametric inference, and then explain how the idea can be extended to inference from non- random samples. Having done so, we define the IS hypothesis within the framework of nonparametric statistics, linking the definition to the economic literature on learning.

3.1 The Basic Idea

To explain the basic idea of nonparametric inference, it is convenient to start with a simple univariate case. Suppose $\mathbf{D} = \{x_i\}_{i=1}^{N}$ is a random sample from an unknown distribution with density function $f(x)$ and that the concern is to use the data to estimate $f(x)$. This requires restricting *a priori* the class of density functions to which $f(x)$ belongs to a class smaller than the class of all possible density functions. In ordinary (parametric) statistics, inference begins by assuming that $f(x)$ belongs to a particular distribution family defined by a small number of parameters, say, the exponential family. Nonparametric inference avoids starting with such a restrictive distribution assumption. Instead, it only assumes that $f(x)$ belongs to the general class of *smooth* functions. Smoothness basically means that, for each x in a 'small' neighbourhood of point x_0, $f(x)$ is almost the same as $f(x_0)$ and, therefore, a small shift away from x_0 to x does not greatly alter $f(x_0)$.[11] The restriction of $f(x)$ to the family of smooth functions allows us to estimate $f(.)$ at each point x_0 by averaging over the observations

[11] Although the above intuitive notion of smoothness is adequate for the purpose of this chapter, there is not yet a complete understanding of the abstract idea of "smoothness", which is usually defined in terms of "the number of derivatives". For a critical discussion see Marron, 1996.

falling in a 'small' neighbourhood around it. The degree (strength) of smoothing is determined by the size of the neighbourhood over which averaging takes place. A larger neighbourhood size implies a greater degree of smoothing, and hence a smaller class of functions to which $f(x)$ is *a priori* thought to belong.

In addition, nonparametric inference ties the strength of smoothing, or equivalently the neighbourhood size over which something takes place, to the size N of the sample. As the sample size grows, the size of the neighbourhood is correspondingly reduced so as to enable the data to reveal the details of $f(x)$. In the limit, when the sample size approaches infinity, the neighbourhood size is forced to zero so that the shape of the density function is fully determined by the data alone. In this way, nonparametric inference aims to do away with the need for pre-specifying the functional form of the density function, and bases that decision on the data alone. If successful, nonparametric inference turns model building (here, finding the right distribution assumption) into an integral part of the process of inference from data, and evades mis-specification.[12]

The reason for naming this approach 'nonparametric' should now be clear. It is called nonparametric because it avoids beginning with the assumption that $f(x)$ belongs to a distribution family defined by a finite number of parameters. Since the approach leaves the determination of the functional form of $f(x)$ to the data, nonparametric procedures have also been called 'model free' or 'distribution-free' procedures. The terms 'nonparametric', 'model-free', and 'distribution-free' are used interchangeably in what follows.

3.2 The Naïve Estimator

The nonparametric literature has flourished over the last three decades, producing a remarkable list of procedures for implementing model-free inference. Here, to set the stage for our discussion and to give a brief glimpse of the field, we review a well-known group of procedures for local averaging that has evolved from

[12] Yatchew, (1998) offers a readable review of nonparametric inference, directed at economists.

attempts to improve on an estimation method called the *simple* or *naïve* estimator (Silverman, 1986:12).

It follows from the definition of a probability density that if variable X has density $f(x)$, then (Bishop, 1995:51-2):

$$f(x_0) = \lim_{h \to 0}\left[\frac{P(x_0 - h < X < x_0 + h)}{2h}\right]. \tag{3.1}$$

Suppose we are given N real observations $\{x_i\}_{i=1}^{N}$ from an unknown density $f(x)$.[13] For any given h, the naïve estimator estimates the density function $f(.)$ at point x_0 by replacing the probability $P(x_0 - h < X < x_0 + h)$ with the proportion of the observations falling in the interval $(x_0 - h, x_0 + h)$. That is:

$$\hat{f}(x_0) = \frac{\sum_{i=1}^{N} I[X_i \in (x_0 - h, x_0 + h)]}{2Nh}, \tag{3.2}$$

where $I(.)$ is the indicator function and parameter h controls the neighbourhood size for averaging. When the support of $f(x)$ is densely populated with data and h is sufficiently small, estimator (3.2) is likely to generate a reliable estimate of the density function.

The naïve estimator has several drawbacks. To begin with, it assigns equal weights to all the observations in the interval $(x_0 - h, x_0 + h)$ and so allows them to contribute *equally* to the estimate $\hat{f}(x_0)$. However, it is more plausible to assume that $f(x)$ is more similar to $f(x_0)$ for points which are closer to x_0 than those further away. A more accurate estimate of $f(x)$ at point x_0 should thus be obtained by giving greater weight to data points closer to x_0. Moreover, the estimator takes the width of the interval $(x_0 - h, x_0 + h)$ to be fixed across the

[13] The exposition of nonparametric estimators in this section draws on Härdle (1990), Härdle (1993), most notably Silverman (1986), and Scott (1992).

entire sample space. Consequently, it has the tendency to miss the details of the density function in the main part of the distribution where the data are plentiful and create noise in the tail area where the data are sparse. This suggests that the estimator can be improved by a procedure that adjusts the width of the smoothing interval to match the local density of the data.

3.3 Kernel-based Estimators

These considerations have led to development of numerous nonparametric estimators that outperform the naïve estimator. Let's restate the naive estimator employing a weight function *w*:

$$\hat{f}(x_0) = \frac{1}{Nh}\sum_{i=1}^{N} w\left(\frac{x_0 - X_i}{h}\right) \tag{3.3}$$

$w(z) = ½$ if $|z| < 1$ and 0 otherwise.

(3.3) makes it explicit that the naïve estimator assigns equal weight to every point in $(x_0 - h, x_0 + h)$. One way to improve on (3.3) is to replace $w(z)$ with a function that assigns weights to points in $(x_0 - h, x_0 + h)$ so that points closer to x_0 receive higher weights while those farther receive lower weights. A convenient class of such functions, termed *kernel* functions, is the family of unimodal functions centred at zero that decline in either direction at a rate controlled by a scale parameter. A common kernel function is the normal density function $K(z) = (2\pi)^{-1/2}\exp(2^{-1}z)$, where $z \in [-1/2, 1/2]$. In general, let *K* be a bounded function that integrates to one and is symmetric around zero. Substituting $K(z)$ for $w(z)$ in (3.3) yields the general class of kernel estimators, defined by

$$\hat{f}(x_0) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_0 - X_i}{h}\right), \tag{3.4}$$

where the scale parameter *h* is called the *bandwidth, smoothing* parameter, or *window width.* A large *h* places a greater weight on observations far apart from x_0

whereas a small h allows only observations very close to x_0 to influence the estimate. If the kernel K is a probability density function, the estimate $\hat{f}(x)$ is also a probability density function. Estimator (3.4) improves on estimator (3.3) but still uses a fixed bandwidth across the x-region. The so-called adaptive kernel estimator improves on (3.4) by varying h in accordance with the local density of the data. To decide on the window width at each data point, "an initial (fixed bandwidth) density estimate is computed to get an idea of the density at the data points." This pilot estimate is then used to adjust "the size of the bandwidth over the data points when computing a new kernel estimate" (Silverman, 1986:100-10).[14]

The definition of the univariate kernel density estimator (3.4) is easily generalised to multivariate cases. Let **Z** be a vector of variables with p elements. The p-variate kernel estimator with kernel K and bandwidth h is defined by

$$f\hat{(}\mathbf{z}) = \frac{1}{Nh^p} \sum\nolimits_{i=1}^{N} K\left(\frac{(\mathbf{z}_0 - \mathbf{Z}_i)}{h}\right). \tag{3.5}$$

K can be any radially symmetric unimodal p-variate probability density function such as the standard p-variate normal density function. A common method of performing multivariate nonparametric density estimation is the product kernel method that replaces p-dimensional kernel K in (3.5) with a product of one-dimensional kernels. In the bivariate case, where $\mathbf{Z} = (X, Y)$, the bivariate product kernel estimator is given by

$$\hat{f}(x,y) = \frac{1}{Nh^2} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right).^{15} \tag{3.6}$$

An important aspect of a multivariate probability distribution is the regression function of each of the variables under study on the remaining variables. This describes how the mean value of the variable in question, conditioned on the

[14] See Silverman (1986: 100-10) for a formal description of the adaptive estimator.

[15] It is important to note that although the expression (3.6) uses kernel independence, this does not imply the independence of the variables (See Appendix *A*).

values of the rest of the variables, varies. The theoretical regression function of Y on X is given by

$$r(x) = E(Y \mid X = x) = \int y f(y \mid x) dy = \frac{\int y f(x, y) dy}{\int f(x, y) dy} \tag{3.7}$$

Substituting the density estimate (3.6) into (3.7) yields the kernel regression estimator (Scott, 1992:220):

$$\hat{f}(x_0) = \frac{\frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x_0 - x_i}{h})}{\frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x_0 - x_i}{h})} y_i . \tag{3.8}$$

Since (3.8) is *linear in the observations* $\{y_i\}$, it can simply be written as (Scott, 1992:220):

$$\hat{f}(x_i) = \mathbf{W}(h)\mathbf{y}, \tag{3.9}$$

where $\mathbf{W}(h)$ is known as the smoother matrix and $\mathbf{y}$ is the vector of observed response values. $\mathbf{W}(h)$ is an $n \times n$ matrix whose elements w_{ij} denote the weight assigned to point x_j in estimating the target function at point x_i. As is evident from (3.7), the issues arising in estimation of a density or regression function are theoretically the same.

Nonparametric estimator (3.5) (or, 3.8) has been shown to be consistent: under very general conditions, as the sample size approaches infinity, the estimator approximates the target density function arbitrarily closely regardless of the form of the function. A simple proof for the consistency of estimator (3.8) is found in Yatchew (1998). The nonparametric approach then theoretically provides a way of learning a density (regression) function from random data, without having to posit *a priori* a parametric distribution family.

In practice, data is not usually known to be random, making the choice of independence and homogeneity assumptions as crucial as the choice of a distribution assumption. An issue for nonparametric inference is how to generalise the methods of model-free inference to non-random samples. Unlike learning the distribution of a random sample, there is no algorithmic procedure for learning of independence and homogeneity assumptions. The only way to extend nonparametric inference to the choice of these assumptions is to follow a hypothetic-deductive method. To be precise, one has to begin with an independence or homogeneity conjecture, deduce its implications, and nonparametrically test them against the data. Consider, for instance, the first order Markov condition which implies that $p(Z_t / z_{t-1}, z_{t-2}) = p(Z_t / z_{t-1})$. One can proceed by hypothetically assuming that the vectors (Z_t, Z_{t-1}, Z_{t-2}) and (Z_t, Z_{t-1}) are randomly distributed to nonparametrically estimate the probabilities $p(Z_t / z_{t-1}, z_{t-2})$ and $p(Z_t / z_{t-1})$. The estimates can then be used to check if the equality holds. In theory, this proposal extends the nonparametric approach to non-sample data. But the manoeuvre, as will be seen, encounters insurmountable practical problems.

4 The Homo Economics as a Nonparametric Statistician

Although the flourishing of nonparametric statistics is relatively recent compared to ordinary (parametric) statistics, there have been a good number of attempts by economists to model the *homo economicus* as a nonparametric statistician. Historically, Bray's work (1982) can be viewed as an early proposal to view the agent as a nonparametric statistician. She studies an economy in which the agents know the true (supply) curve $p_t = a + bE(p_{t+1}) + u_t$ but must form expectations $E(p_{t+1})$ to plug into it. She conjectures that they form expectations $E(p_{t+1})$ by taking the average of past prices, which is equivalent to learning with the naïve estimator. Commenting on Bray's pioneering work, Lucas suggests that 'learning by averaging' seems to be a plausible conjecture about human learning (1986:236). Thomas Sargent also considers using histogram and kernel estimators for modelling learning behaviour (1993:106-107).

More recently, Chen and White (1998) have criticised early works on learning in economics such as Bray and Savin (1986), which assume that the agents already know the correct unestimated model of the economy without any explanation as to how the model was learnt in the first place. To eliminate this shortcoming, Chen and White propose that agents are nonparametric statisticians who utilise an *online* version of the kernel regression estimator (3.8) to learn about the economy. To explain what this means, note that the estimators described earlier, including estimator (3.8), are all defined from the whole data, meaning that the estimate must be recomputed from the whole sample for every newly arriving observation. In learning situations of interest in economics, data arrives as an ongoing sequence $\{(x_1, y_1), (x_2, y_2), \ldots\}$. It is thus more plausible that the agent works with an estimator that at any time t can be represented as a function of the estimator at time $t-1$ and the new pair of observations (x_t, y_t). Interestingly, estimator (3.8) can be reformulated to achieve this (Härdle, 1990:66):

$$\hat{f}_{N+1}(x_0) = \hat{f}_N(x_0) + (hN)^{-1} K_{N+1h}(x - x_{N+1})(y_{N+1} - \hat{f}_N(x_0)), \tag{4.1}$$

which dispenses with the need for re-computing the estimate from the whole sample each time. With this proposal, the person uses the data available at time t to obtain $\hat{f}_N(x)$ and uses the estimate to make predictions necessary for his future decisions. As new data comes in, he uses rule (3.10) to update the estimate.[16] Chen and White establish the necessary and sufficient conditions under which regression estimator (3.10) asymptotically converges to the true regression function in spite of the fact that feedback from learning may alter the relation being learnt.[17]

[16] N stands for the sample size at time t.

[17] In addition to these proposals, a sizeable number of studies of learning in economics and game theory utilise neural network inference procedures (e.g., Salmon, 1995). Neural network initially viewed as an independent field aiming to tackle complex learning tasks not usually considered in statistics. Yet it soon emerged that the procedures are essentially nothing but variants of nonparametric inference procedures and are subject to similar strengths and limits (Friedman, 1994; Cheng, et al., 1994; Ripley, 1993). They cannot solve any learning problem that theoretically falls beyond the reach of nonparametric inference.

This chapter follows these economists in viewing the economy as a society of nonparametric statisticians, and investigates if the conjecture helps shed light on some critical issues in theoretical economics. Specifically, we investigate if agents in such a society can learn the probabilistic features of their environment from ordinarily available data samples, whether it is possible to predict what the agents think given the data generated by the economy, and finally whether the conjecture helps us understand how the agents revise their view of the economy in the face of a new policy.

5 Intrinsic Limitations of Model Free Inference

Nonparametric estimators, as stated, can be shown to be asymptotically consistent, in that they uncover the target function as the number of observations approaches infinity. The asymptotic results teach us how learning is in principle possible and provide some general insights into the working of nonparametric estimators and what must be done as the sample size increases to ensure an accurate estimate (White, 1992:121). In reality, however, we only have access to a finite and usually small number of observations, and because the economy also changes over time, remote past data is often uninformative. For economics, the relevant question is not thus whether there are model-free estimators that can asymptotically discover the truth or whether the opinions in a society of nonparametric statisticians asymptotically converge to truth. The relevant question is whether it is possible with a 'reasonably-sized' sample to learn a 'good' approximation of a relatively complex target function using nonparametric methods. This section argues that accurate approximation of 'complex' functions using nonparametric techniques is practically impossible. Even a 'crude' model-free approximation of a function relating several variables requires a gigantically large sample that is rarely available in practice. The argument is inspired by a critique of the claims surrounding the theory of neural networks given in Geman et al. (1992).

5.1 The Bias-variance Decomposition

Essential for investigating the limitations of nonparametric methods with 'reasonably sized' samples is a precise definition of what is meant by a 'good' or 'accurate' estimate. This can be achieved by considering nonparametric estimation of a simple regression function. Suppose we are given a random data set $\{(x_i, y_i)\}_{i=1}^{N}$ and are interested in estimating the regression function $f(x)$ in

$$y = f(x) + \varepsilon, \tag{5.1}$$

where ε has mean zero and is independent of *X*. An objective in searching for an estimate of $f(x)$ is to predict the value of *Y* when only *x* is known. A possible way to define the accuracy of an estimate is then in terms of the accuracy of its predictions. A popular measure of predictive accuracy is the mean squared prediction error (*MPE*):

$$MPE = E[y - \hat{f}(x)]^2, \tag{5.2}$$

which provides a measure of the accuracy of the estimate $\hat{f}(x)$ when *X* takes value *x* and *Y* takes value *y*. The expectation $E(.)$ is taken with respect to the joint probability distribution of *Y* and *X*. The error (5.2) can be decomposed into two distinct elements (White, 1992:97-98):

$$MPE = E[(y - f(x)]^2 + E[\hat{f}(x) - f(x)]^2. \tag{5.3}$$

The first term on the right hand side is the variance of *Y* at point x, which is independent of the estimate and hence plays no role in evaluating accuracy. The second term is the mean squared distance between the estimate and the regression function at point *x*, and provides a natural measure of approximation accuracy. The term is known as the *mean squared estimation (MSE)* error:

$$MSE = E[\hat{f}(x) - f(x)]^2, \tag{5.4}$$

where the expectation is taken with respect to $p(x)$. From this viewpoint, a 'good' approximation refers to an estimate that yields a 'negligible' *MSE* error. Since the estimate $\hat{f}(x)$ depends on the data, it can be viewed as a realisation of a random variable defined over all samples *D* of fixed size *N* that can possibly be drawn from the system. This means we can define the mean and variance of the estimate. Letting $E[\hat{f}(x)]$ be the mean of $\hat{f}(x)$ taken over all hypothetical samples *D* of fixed size *N*, the *MSE* error can then be decomposed into two distinct components (Geman et al. 1992:10):

$$\begin{aligned}
&E[(\hat{f}(x) - f(x))^2] \\
&\quad = E\{[(\hat{f}(x) - E[\hat{f}(x)]) + (E[\hat{f}(x)] - \hat{f}(x))]^2\} \\
&\quad = E[(\hat{f}(x) - E[f(x)])^2] + E[(E[\hat{f}(x)] - f(x))^2] \\
&\quad + 2E[(\hat{f}(x) - E[\hat{f}(x)]) \times (E[\hat{f}(x)] - f(x))] \\
&\quad = E[(\hat{f}(x) - E[\hat{f}(x)])^2] + (E[\hat{f}(x)] - f(x))^2 \\
&\quad + 2[E[\hat{f}(x)] - E[\hat{f}(x)])] \times [E[\hat{f}(x)] - f(x)] \\
&\quad = E[(\hat{f}(x) - E[\hat{f}(x)])^2] + (E[\hat{f}(x)] - f(x))^2 \qquad (5.5)
\end{aligned}$$

The first term on the right hand side is the *variance* of the estimate at point *x*, measuring the dispersion of $\hat{f}(x)$ around its mean. The second term is the *squared bias* of the estimate at point *x*, giving the squared distance between the mean estimate value $E[\hat{f}(x)]$ and the regression function at point *x*. Since the variance and bias components contribute to the *MSE* error, both must approach to zero for a good approximation or, in other words, accurate learning to occur. Therefore, the question posed earlier is in fact whether it is possible to make both bias and variance 'small', with 'reasonably' sized samples in 'interesting' inference problems, using nonparametric procedures such as kernel regression estimators (Geman et al., 1992:44).

5.2 The Bias-Variance Tradeoff

The estimate (estimator) $\hat{f}(x)$ depends on three factors: (a) the estimator family (say, the kernel family), (b) the smoothing parameter (or parameters), and (c) the data. By altering any of these elements it is possible to vary the estimate, and hence control the *MSE* error. Enough for the current purpose is to consider the effect of varying the smoothing level and the data (sample size). We begin by investigating the effect of varying the level of smoothing on the squared bias and variance components of the *MSE* error.

Increasing smoothing reduces the variance part of the error. In the extreme case, if each neighbourhood (bandwidth) is chosen to cover the whole x-region, the kernel estimate becomes equivalent to the average of the response values everywhere. In that case, the variance part of the *MSE* error is at its lowest possible value, namely zero. However, when each bandwidth is so chosen to cover the whole x-region, the estimator always yields a straight line, which is most likely quite different from the target function. In that case, the response value y corresponding to each x will be significantly different from the estimate, leading to a substantial bias (Hastie, et al., 1990:17). In general, an attempt at eliminating variance by increasing smoothing can cause an increase in the bias component that may be greater than the reduction in *MSE* error obtained by reducing the variance. Consequently, decreasing variance by increasing smoothing does not necessarily reduce the overall error; it may in fact increase it.

Conversely, decreasing smoothing reduces the squared bias component of the *MSE* error. In the extreme case, if each neighbourhood (bandwidth) is chosen to contain only one observation, the kernel estimator interpolates the data. In that case, the squared bias term achieves its lowest possible value at the data points and, if the target function is smooth, is also small in the close neighbourhoods of the points. But the reduction in the bias term can sharply increase the variance of the estimator, since the estimate at each point x would most likely be quite different from its average value (Bishop, 1996:336; Hastie, et al., 1990:17). As a general rule, then, for a fixed sample, an attempt at reducing the squared bias part by decreasing smoothing could increase the variance part of the error, thus increasing the overall value of the error.

These considerations about the effect of varying smoothing, which can be made formally precise in the case of each family of nonparametric estimators, point to a trade-off between the squared bias and variance components of the *MSE* error. For a fixed sample, the squared bias component can be reduced at the expense of increasing the variance factor and the variance factor can be reduced at the expense of increasing the bias component (Silverman, 1986:35). Geman el al. (1992) term this tradeoff the bias-variance dilemma.

Since this dilemma plays a central role in the analysis to follow, it is worth illustrating it with a simple example, which we adopt from Wahba and Wold (1975). Suppose $x \in [0,3]$ and y is related to x by

$$y = f(x) + \varepsilon, \tag{5.6}$$

where $f(x) = 4.26(e^{-x} - 4e^{-2x} + 3e^{-3x})$ and ε is distributed as *N*(0,0.2). We generate 100 data points from the model to investigate the effect of varying smoothing on the performance of a kernel regressor. If the bandwidth is so chosen to cover the whole *x*-region (e.g., if it is set at 6) as in Figure 5.1, the resulting estimate is a straight line, significantly different from the target function. Alternatively, if the bandwidth is reduced to 0.01 as in Figure 5.2, the estimator interpolates the data, and again the estimate drastically differs from the true function. However, when the bandwidth is set to an intermediate value of 0.7 as in Figure 5.3, both the variance and bias of the estimate are reduced, and the estimator closely approximates the target function.

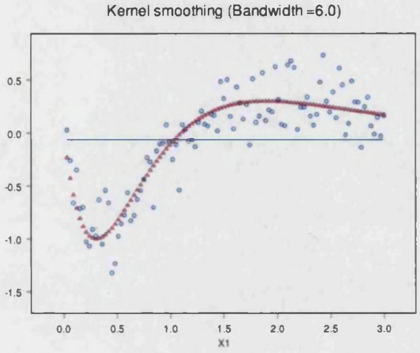


Figure 5.1

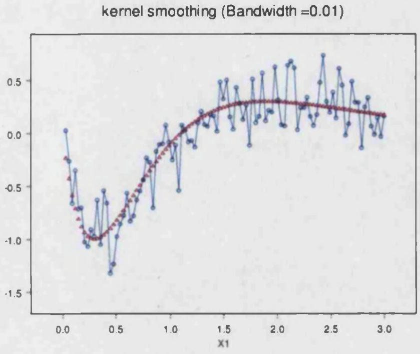


Figure 5.2

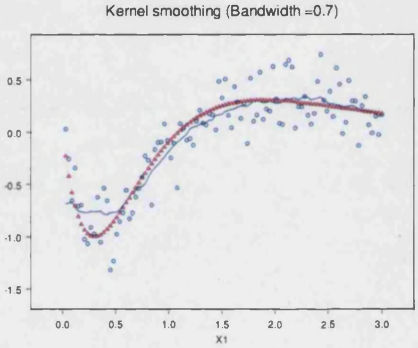


Figure 5.3

The purple line shows the (true) regression function whereas the blue line shows the estimate.

An immediate implication of the bias and variance dilemma is that, given a data set, smoothing cannot be reduced arbitrarily. Quite the opposite, for a fixed sample size, there is a unique (set of) smoothing parameter value (values) that ensures an optimal tradeoff between the squared bias and variance in the sense of minimising *MSE* error (Friedman, 1994:32). This optimal value fixes the class of functions that the estimator can approximate given the data, and hence fixes the minimum bias possible. If the optimal neighbourhood size (bandwidth) relative to the data is, for instance, the whole *x*-region, the estimator will only be able to approximate straight lines. In that case, if the target function is considerably different from a straight line, the estimator will produce a highly biased estimate of the function. With a finite sample, there is in a sense no difference between parametric and nonparametric estimators (White, 1992:117); they both search through a proper subset of the class of all possible functions.

The bias-variance dilemma can only be resolved by increasing the sample size. As the sample size increases, and the input variable space (*x*-region) is increasingly densely populated with data everywhere, smoothing can be reduced without increasing variance. And, as smoothing is reduced, the estimator becomes able to search over an increasingly larger class of functions, thus reducing the chance of bias. To illustrate the point, let us return to the above example. This time, we hold the level of smoothing fixed but vary the sample size. If we simulate a sample of 100 observations from the model, and fit a model using a kernel regressor with bandwidth 0.03, the result is a highly variable curve significantly different from the target function (Figure 5.4). If the sample size is 1,000, the same level of smoothing yields a much smoother curve, with lower bias and variance (Figure, 5.5). When the sample size is 10,000, we get an estimate that closely matches the regression function (Figure 5.6). By increasing the sample size, it is possible to reduce bias and variance simultaneously.

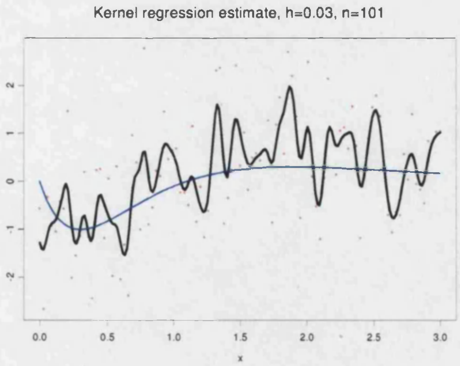


Figure 5.4

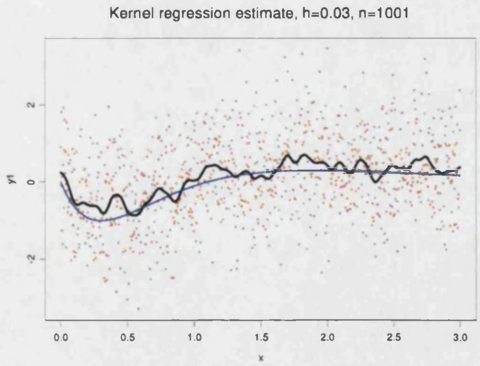


Figure 5.5

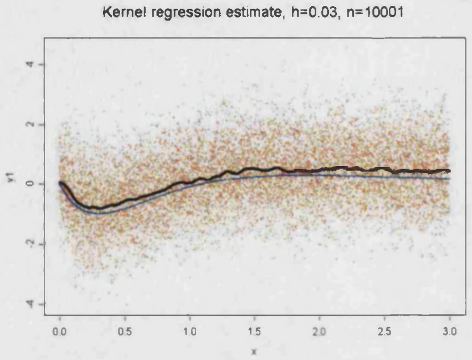


Figure 5.6

The blue line shows the (true) regression function whereas the black line shows the estimate.

Therefore, the key to driving both the bias and variance components of the *MSE* error towards zero using local averaging (fitting) estimators is to densely populate the *x*-region (input variable space) with data. If this turns out to be impossible in interesting inference situations, because of the bias-variance tradeoff, local averaging (fitting) estimators can only search through a proper and most likely small subset of the class of all possible functions. In that case, they may not be able to produce an accurate approximation of the target function.

5.3 The Curse of Dimensionality

Although it may be possible to densely populate low dimensional input variable spaces (i.e., one or two predictors) with ordinarily available samples, this is practically impossible in high dimensional spaces due to the *curse of dimensionality* problem (Bellman, 1961). Recall that the basic idea of local averaging (or fitting) is to divide the input variable space (***x***-region) into a number of cells and take the average of the responses in each cell as the estimate of the regression (or density) function in that cell. The curse of dimensionality refers to the fact that the number of cells increases exponentially with the dimension of the input variable space (i.e., the number of regressors). In general, if *d* indicates the dimension of *X*, ($\mathbf{x} \in R^d$), and each regressor co-ordinate is divided into *M* divisions, the total number of cells will be M^d. Since each cell must contain some data points to make any inference, the number of data points required for local averaging also grows exponentially with the dimension of the input variable space. For example, if M is taken to be 10, and ten observations are required for densely populating each cell, a sample of 10×10^2 observations will be needed to densely populate a two dimensional input variable space (two regressors). On the same ground, a sample of 10×10^{10} observations will be required to equally populate a ten dimensional input variable space (ten regressors). The curse of dimensionality thus makes it impossible to adequately densely populate high dimensional input variable spaces with ordinarily available samples.

To provide more insight into the problem, suppose we have 10,000 data points uniformly distributed over the ten-dimensional unit cube $[0,1]^{10}$. A bandwidth of diameter 0.2 in each regressor co-ordinate results in a volume of $0.2^{10} \approx 1.02\times10^{-7}$ for each cell, and the expected number of observations in each cell is approximately 1×10^{-3}. Obviously, no local averaging is possible with this number of data points. Alternatively, if we increase the neighborhood size to include at least ten observations, the bandwidth must cover at least 0.5 of each co-ordinate. In that case, averaging is carried out over at least half of the range along each co-ordinate and is no longer local. The general lesson is that in high dimensional spaces, if the neighborhood is 'local' (i.e., small), it is almost surely empty. If the neighborhood is not empty, it is not 'local'.[18]

What is more, to drive both elements of the *MSE* error of a local averaging (fitting) estimator towards zero, which is necessary for the estimator to arbitrarily closely approximate the target function, it is necessary to increasingly divide the input variable space into smaller and smaller cells, and, in parallel, the number of data points in each cell must increasingly grow larger and larger. In the limit, the number of cells M and the number of data points in each cell must approach infinity to ensure a good approximation. As a consequence, densely populating of an even low

[18] The example in this paragraph is adapted from Härdle (1990:258). For further discussion of the implications of the curse of dimensionality see Bellman (1961:94); Friedman (1991), (1994), Friedman et al. (1981:817), Scott (1995), Bishop (1995), Härdle (1990), Silverman, 1986:129.

dimensional input variable space (say, with four or five regressors) demands an astronomically large sample that is impossible to achieve in practice, or at least in situations of interest in economics.

Taken together, the bias-variance tradeoff and the curse of dimensionality imply that an astronomically large sample, which is impossible to achieve in practice, is required to arbitrarily closely approximate the target function even for moderate numbers of regressors (say, 4 or 5). Ordinarily available samples in situations of economic interest do not even allow for a crude approximation of a high dimensional function using local averaging techniques. This intrinsic limitation of model-free inference reveals that even with an unusually large sample the agent is not able to accurately learn the probabilistic relations characterizing his choice situation from data alone. Learning the probabilistic relations of a choice situation calls for substantive probabilistic non-sample information.

5.4 Defeating the Curse of Dimensionality

The impossibility of local averaging (or fitting) in high dimensional input spaces have prompted search for nonparametric inference methods that build an approximation of a high dimensional function that takes the form of expansions in low dimensional (univariate) functions. If it were possible to approximate a complex high dimensional function with a sum or product of low dimensional (univariate) functions, nonparametric inference would theoretically only involve estimation of low dimensional functions. In that case, the curse of dimensionality would raise no intrinsic issue for a data-driven method of inference. And the argument for the impossibility of model-free learning of high dimensional functions would break down at a closer scrutiny. To explain that this is not really the case, and to draw some further important methodological conclusions about the boundaries of model-free learning, we look at the method of project pursuit regression developed by Friedman et al. (1981). The method is directly aimed at extending the idea of nonparametric inference to high dimensional data.[19]

[19] Friedman (1994) and Hastie et al. (1994) review some of nonparametric multivariate approximation methods.

In multivariate regression analysis the objective is to model the conditional expectation of response variable Y given predictor variables $\mathbf{X} = \{X_1,...X_p\}$ on the basis of a sample $\{y_i, x_{1i},..., x_{pi}\}_1^N$. The data are assumed to have come from a system described by

$$y = f(x_1,..., x_p) + \varepsilon \tag{5.7}$$

The projection pursuit regression (PPR) estimator models the conditional expectation of Y given $\boldsymbol{X}$, $f(\mathbf{x})$, as a sum of general functions of linear combinations of the predictors, i.e.,

$$\hat{f}(\mathbf{x}) = \alpha_0 + \sum\nolimits_{m=1}^{M} g_m(z_m), \qquad z_m = \sum\nolimits_{i=1}^{p} \alpha_{mi} x_i , \tag{5.8}$$

where the univariate variable z_m denotes a projection of the vector $\boldsymbol{X}$ onto a one-dimensional space, and g_m is a univariate smooth function, called basis function. The PPR estimator constructs an approximation $\hat{f}(\mathbf{x})$ in an iterative manner. It begins by setting α_0 equal to $\overline{y}$, the average of the observed responses, and computes the residuals $r_{1i} = y_i - \bar{y}$. Next, it assigns some initial values to projection parameters α_{1i} to define a univariate variable $z_1 = \sum_{i=1}^{p} \alpha_{1i} x_i$ and regresses r_{1i} on z_{1i} using some univariate nonparametric estimator. It updates the parameters α_{1i} by minimising the squared residuals sum $\Delta = \sum (r_{1i} - \hat{g}(z_{1i}))^2$ over all possible choices of α_{1i}, inserts the optimal values of α_{1i} into $z_1 = \sum_{i=1}^{p} \alpha_{1i} x_i$, and re-estimates $\hat{g}_1(z_1)$. Again, it uses the new estimate to update α_{1i} and repeats the process until no further reduction of the sum of residuals can be achieved. It then adds the final estimate $\hat{g}_1(z_1)$ to $\overline{y}$ and computes the new residuals $r_{2i} = [y_i - (\bar{y} + \hat{g}_1(z_1))]$. These steps are repeated to obtain a second basis function $\hat{g}_2(z_2)$, and the process of constructing new basis functions is continued until no further reduction can be achieved in the residuals.

It has been shown that if the number of basis functions M in (5.8) is let to grow to infinity, the function can approximate arbitrarily closely any continuous function (Diaconis et al., 1984). This means that as long as the target function is continuous and the number of basis functions M is let to grow arbitrarily large, the PPR estimator can approximate it arbitrarily closely.[20]

This consistency result is theoretically reassuring but is of not much help in practice. The number of basis functions M in a projection pursuit regression approximation plays the same role as the smoothing parameter in the kernel estimators. If M is taken to be small, the estimator can only search through a small subset of continuous functions, which may neither include the target function nor a good approximation thereof, and will therefore be biased. If M is taken to be large, the estimate interpolates the data and will be highly variable. Again, the bias-variance tradeoff restricts the number of basis functions that can be included in a projection pursuit estimate given a data set, thus limiting the class of functions that the estimator can approximate in practice. Consequently, the sample size must be adequately large to include an adequately large number of basis functions so as to ensure a good approximation.

Moreover, as pointed out by Huber (1985), there are relatively simple functions that cannot be approximated by a sum of a finite number of additive basis functions. An example is $f(x_1, x_2) = e^{x_1 x_2}$ (Huber, 1985). An assumption behind the use of the projection pursuit method is that a good approximation of the target function can be obtained by an estimate containing only a small number of basis functions. There is no reason to think that this assumption is valid if nothing is known about the target function.

5.5 The Loss of Interpretability

There is another aspect of the attempt to extend nonparametric inference to high-dimensional data that is worth noting. In practice, as explained, any extension of nonparametric inference to high dimensional input spaces takes the form of

[20] See Ripley (1996) for a statement of the proof.

expansions in low dimensional functions (Barron et al., 1991:80). And, a good estimate may require a large number of basis functions. In that case, the estimate is a model like (5.8) with a large number of univariate functions g_m. Such a model gives no clear description of how each regressor X_i *separately* relates to the response variable *Y*; each regressor X_i relates to *Y* in a very complex way (Hastie, et al., 1994:67). As a result, even if it were known that $X_1,...,X_p$ are causes of *Y* and have no latent common causes with *Y*, it would still be impossible to use the model to trace the distinct effect of each X_i on *Y*. The model is only useful for *ex ante* and *ex post* predictions; it is not suited for analysis of actions and policies or understanding of the system. A similar remark is true of the outcome of other nonparametric multivariate approximation methods, including the neural network approach (Warner, 1996). The price to pay for extending nonparametric inference to high dimensional data is the loss of interpretability (Friedman, 1994:9).

A general lesson learnt from this consideration is that establishing an interpretable model suitable for evaluating actions and policies calls for substantive probabilistic information. One has to begin with a parametric model to ensure interpretability. If no substantive probabilistic assumption is made at the outset, the outcome is a black box model that lacks interpretability and is only suitable for *ex ante* and *ex post* predictions. There is therefore a tradeoff between the interpretability of a model and the amount of *probabilistic* information used to obtain it.

We have so far explained some of the limitations of nonparametric inference from *random* data. It is appropriate to close it by looking at the possibility of generalising nonparametric inference to any sample regardless of whether it is random or not. Any such attempt, as stated earlier, requires hypothetically assuming that the data are random, and nonparametrically estimating the joint distributions of various subsets of the variables to assess alternative independence and homogeneity assumptions. However, since accurate estimation of the joint distribution of several variables with ordinary samples is not practically possible, successful nonparametric evaluation of these assumptions is not practically possible either. Alternative methods are needed for selecting independence and homogeneity assumptions.

6 Model Selection

The analysis of the bias-variance tradeoff demonstrates that, for any data set, there is an optimal smoothing parameter value that minimises the *MSE* error. The optimal value fixes the class of functions over which the estimator can search, and hence determines the best approximation of the target function possible given the data. A crucial issue in nonparametric inference, therefore, concerns the choice of the smoothing parameter value that is optimal given the data in hand. We refer to this issue as the smoothing parameter or nonparametric model selection problem. In nonparametric statistics, the assumption is that nothing is known about the target function apart from smoothness. This implies that one has to look at the data or, more precisely, assess the predictive accuracy (error) of possible models to select a model. This is indeed the approach pursued in nonparametric statistics. Broadly speaking, a number of models with different smoothing parameters are fitted to the data, the predictive error of each model is estimated, and the model with minimum prediction error is chosen (Moody, 1994:149). The remainder of this chapter first describes the rich variety of methods for estimating prediction error in order to explain the possibility of defining alternative predictive model selectors. It then shows how different model selectors often choose different models in practice. Having done this, it investigates if there are any adequate grounds for choosing a model selector as optimal. It finally spells out some intrinsic limitations in estimating prediction error. And the implications of the discussion for nonparametric inference are spelled out.

6.1 Alternative Model Selectors

A model selector is consisted of a discrepancy (distance) function and an estimation strategy. The discrepancy function is to measure the distance between the predicted value of the response variable and its actual value or, in short, to define prediction error. The estimation strategy is to estimate the accuracy of the model with respect to the population. To explain some of the approaches to prediction error estimation, we continue working with the squared Euclidean distance $[y_i^* - \hat{f}_h(x_i)]^2$, where $\hat{f}_h(x_i)$ is the response value predicted by the model for a new observation at x_i, and y_i^* is the actual response value. For the purpose of this section, we propose to measure the prediction error rate of a model using the *average mean squared prediction error* (APE):

$$APE(h) = N^{-1}\sum_{i=1}^{n} E(y_i^* - \hat{f}_h(x_i))^2 \qquad 6.1)$$

The error, as made it explicit, depends on the smoothing parameter *h*. A problem is that future data are not known, and except for the strategy of "wait and see" any attempt at estimating error (6.1) involves exploiting exiting data. However, the same data cannot be used for both obtaining a model and estimating its predictive accuracy. An attempt to do so amounts to estimating *APE* using the average squared residuals (*ASR*):

$$ASR(h) = N^{-1}\sum_{i=1}^{n}\{y_i - \hat{f}_h(x_i)\}^2. \qquad (6.2)$$

Following a technique explained in Eubank (1988), the expected value of *ASR* can be decomposed into (see Appendix *B*):

$$E(ASR(h)) = \delta^2 + f(x)'(I - \mathrm{W}(h))^2 f(x) + N^{-1}\delta^2 tr[\mathrm{W}(h)^2] - 2N^{-1}\delta^2 tr[\mathrm{W}(h)] \qquad (6.3)$$

W(*h*) is the smoother matrix with bandwidth *h*, *f*(x) is the regression function, δ^2 is the variance of *Y* (given *x*), and *tr*[W(*h*)] is the trace of the smoother matrix.[21] However, applying the same technique to the average mean squared prediction error yields

[21] The prime in $f(x)'$ stands for transpose.

$$APE(h) = \delta^2 + f(x)'(I - \mathbf{W}(h))^2 f(x) + N^{-1}\delta^2 tr[\mathbf{W}(h)^2]. \qquad (6.4)$$

A comparison of (6.3) and (6.4) shows that *ASR* on average underestimates the mean prediction error *APE* by factor $2N^{-1}\sigma^2 tr[\mathbf{W}(h)]$. For this reason, statisticians call this estimate of the prediction error the *apparent rate* of error or the *substitution* error (Efron, 1983). In fact, substitution error (6.2) can arbitrarily be reduced by selecting a sufficiently small smoothing parameter value so that the model interpolates the data. However, this would not necessarily lead to a model that minimises the *MSE* error, which is essential for minimising prediction error. The literature provides three avenues for obtaining an unbiased estimation of prediction error.

A strategy is to split the data in two sets, a *training* set and a *test* set. The training set is used to obtain a model and the test set is used to evaluate its performance. By using different data for model construction and evaluation, the data splitting strategy evades the problem with the apparent rate of error. It, nevertheless, has several drawbacks. First, by leaving part of the data aside as a test set, the strategy fails to make optimal use of the data in estimating a model. In the nonparametric setting, a smaller sample necessitates a greater degree of smoothing, which reduces the class of functions over which an estimator can possibly search. Consequently, the procedure is likely to lead to the choice of a highly biased model. To be precise, the strategy estimates the predicting error of a model built from (say) half of the data but the primary concern is to estimate the predictive accuracy of a model that can be constructed from the whole data (Zucchini, 2000:19). Secondly, when the sample size is small, as is usually in practice, splitting the data leads to a small test set, which may also be inadequate for achieving a reliable estimate of the model prediction error (Faraway, 1998:335). Finally, the strategy involves an arbitrary decision in dividing the data into a training set and a test set, which could affect estimation of the prediction error and hence model selection (Glymour, et al., 1995:37).

Another strategy attempts to overcome the inefficiency of the simple data splitting method by utilising resampling techniques to create a test set. Cross validation is the oldest resampling technique used for estimating prediction error, attributed to Stone (1974). The method, in its most common form, involves leaving a data point (x_i, y_i) aside at a time as a test set, fitting a model to the remaining *N-1* data points, and using the model to predict the omitted observation. The process is repeated for all the *N* observations, and the average of the errors is taken as the estimate of the model prediction error. Let $\hat{f}_h^{-i}(.)$ be the model estimated from sample ***D*** excluding data point (x_i, y_i), and $\hat{f}_h^{-i}(x_i)$ the response value predicted by $\hat{f}_h^{-i}(.)$ at point x_i. The cross-validation estimate of prediction error is given by

$$CV(h) = \frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}_h^{-i}(x_i)]^2 \; . \tag{6.5}$$

This technique is called "leave-one-out" cross validation, since each time only one data point is left out. Alternative cross-validation methods can be defined by holding out a different number of observations (say, 5) each time. A cross validation based model selector chooses a smoothing parameter *h* that minimises error estimate (6.5) or a similar one.

This re-sampling strategy yields an unbiased estimate of the mean prediction error. The reason for the unbiasedness of an estimator such as (6.5) can intuitively be understood by noting that

$$E[y_i - \hat{f}_h^{-i}(x_i)]^2 = \sigma^2 + E[f(x_i) - \hat{f}_h^{-i}(x_i)]^2, \tag{6.6}$$

and

$$E[y_i^* - \hat{f}_h(x_i)]^2 = \sigma^2 + E[f(x_i) - \hat{f}_h(x_i)]^2. \tag{6.7}$$

As the sample size increases, the estimate $\hat{f}_h^{-i}(x_i)$ is expected to become closer to the estimate $\hat{f}_h(x_i)$, which is based on the full data, i.e., $\hat{f}_h(x_i) \approx \hat{f}_h^{-i}(x_i)$. And, as a result, the mean value of *CV*(*h*) becomes increasingly close to the mean prediction error, i.e., $E(CV(h)) \approx APE(h)$. This means that *CV*(*h*) is an approximately unbiased estimator of the mean prediction error (Hastie, et al., 1990:43). Hall (1983) establishes that a sequence of smoothing parameters produced by the cross validation procedure (6.5) leads to consistent density estimation. A sequence of smoothing parameters minimising *CV*(*h*) is therefore expected to minimise the mean prediction error.

Although the description of cross validation techniques gives a glimpse of resampling approach, for the discussion to follow, we also need to mention an alternative resampling procedure, called bootstrapping, which exploits a different strategy for constructing a test set. In its simplest form, the bootstrap method takes the original data set in place of the *unknown* distribution, considers each observation in the data set as equally probable, and draws *N* new observations from the set with *replacement.* The new sample is called the *bootstrap* sample. It fits the model to the bootstrap sample and estimates its prediction error by applying it to the original data set. The technique generates *B* bootstrap samples, estimates the model on each, and applies each fitted model to the *original* data to obtain *B* estimates of the model's prediction error. The average of these estimates is taken as the model's prediction error. A bootstrap model selector suggests choosing the smoothing parameter that minimises the average prediction error (Efron et al., 1993:247-254). Appendix *C* formally defines some bootstrap error estimators which will be mentioned in the text.

Statisticians have also pursued a third avenue to obtain an unbiased estimate of mean prediction error. As said earlier, the mean average squared residuals *E*(*ASR*) differs from the average mean prediction error *APE* by factor $2n^{-1}\delta^2 tr[\mathrm{W}(h)]$. If this term could somehow be estimated, it would be possible to transform *ASR* into an unbiased estimate of *APE* by adding an estimate of the term to *ASR*. In that case, the expected value of the augmented *ASR* would be the same as *APE,* and there would remain no need for computationally intensive resampling procedures. One would be able to estimate *APE* by correcting *ASR* with a term that cancels the bias term out. This possibility is the drive behind an ongoing search for estimates of prediction error that take the form

$$E(ASR(h)) + 2N^{-1}\delta^2 tr[\mathrm{W}(h)]. \tag{6.8}$$

To further illustrate the variety of ways of estimating prediction error, it is worth looking at one of the model selectors that proceed by minimising an estimate of (6.8). Note that the cross validation criterion can also be written as:

$$CV(h) = \frac{1}{N}\sum_{i=1}^{N}\{y_i - \hat{f}_h^{-i}(x_i)\}^2 = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{y_i - \hat{f}_h(x_i)}{1-w_{ii}}\right\}^2 \tag{6.9}$$

where w_{ii} are the diagonal elements of the smoother matrix W (Eubank, 1988:30). Thus, the leave-one-out cross validation estimator corrects the bias of *ASR* by multiplying it with function $(1-w_{ii})^{-2}$. Craven and Whaba (1979) suggests an approximation to (6.9) by replacing the diagonal elements w_{ii} with their average, namely $tr(\mathrm{W})/N$, calling it *generalised cross validation* (*GCV*). That is, they replace (6.9) with

$$GCV(h) = \frac{1}{N}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_h(x_i)}{1-tr(\mathrm{W}(h))/N}\right\}^2 \tag{6.10}$$

as an estimate of the mean prediction error. While *CV* corrects the bias of *ASR* by multiplying it with function $(1-w_{ii})^2$, *GCV* corrects it by $\{1-tr(\mathrm{W}(h))/N\}^{-2}$. If we take a first order Taylor expansion of this (correcting) function and ignore its reminder, we obtain $1+2N^{-1}tr(\mathrm{W}(h)$. Using this approximation, *GCV* can be restated as

$$GCV(h) \approx ASR(h) + 2N^{-1}tr[\mathrm{W}(h)]ASR(h) \tag{6.11}$$

As shown in Härdle (1990:155), the expected value of the second term in (6.11) is approximately the same as the second term in (6.8) and asymptotically cancels out the bias term in *ASR*. Following a different route, Eubank (1988:35-6) sketches an alternative proof for the consistency of *GCV* as an estimator of *APE*. Now, an important point is that there is nothing unique about the correcting function $\{(N - tr(\mathrm{W}))/N\}^{-2}$. Any function, with the same first-order Taylor expansion as $1 + 2N^{-1}tr(\mathrm{W}(h)$, can equally correct the bias term in *ASR*. This possibility opens the way for producing alternative unbiased model selectors and is behind most known selectors such as Akaike's information criterion (Akaike, 1972), finite prediction error (Akaike, 1974), Shibata's model selector (1981) and Rice's bandwidth selector (1984). All these selectors are based on an estimate of prediction error that corrects *ASR* by a function whose first order Taylor expansion is $1 + 2N^{-1}tr(\mathrm{W}(h)$ (Härdle, 1990:167).[22]

This section has described some of the general strategies for error estimation, stressing how by varying the estimation strategies alternative model selectors can be invented. It is also important to bear in mind that the estimation strategies can be combined with other discrepancy functions than the squared Euclidean function to generate alternative model selectors. One can use, for instance, the Kullback-Leibler discrepancy. In general, alternative model selection techniques can be invented by slightly varying the discrepancy function or the estimation strategy (Amemiya, 1980:325).

6.2 Which Model Selector Should Be Used?

The existence of alternative model selectors raises the question of which selector to choose in practice. If these methods picked up the same model (smoothing parameter value), one could arbitrarily select any of the methods. But since they use different estimation strategies, when applied to ordinarily available samples, they often suggest different models. Consider the dataset plotted in the figures below, which consists of 100 observations simulated from the model used earlier to illustrate the bias-variance tradeoff. Two selection criteria have been applied to the data to find the optimal bandwidth in kernel regression of *Y* on *X* – the leave-one-out cross validation and generalised cross validation method. The former

[22] For a discussion of these selectors see Härdle (1990), chapter 5

suggests the optimal bandwidth to be 0.25, producing the model in Figure 6.1 while the latter suggests it to be 0.07, producing the model in Figure 6.2. These models are evidently different.

Figure 6.1

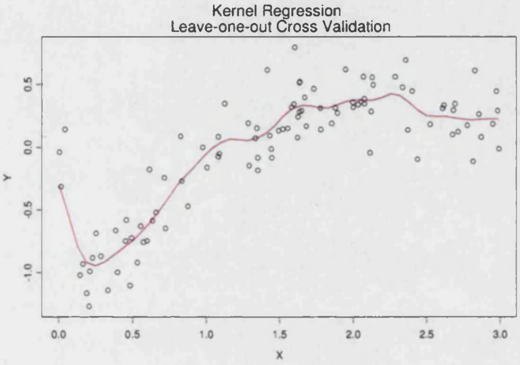
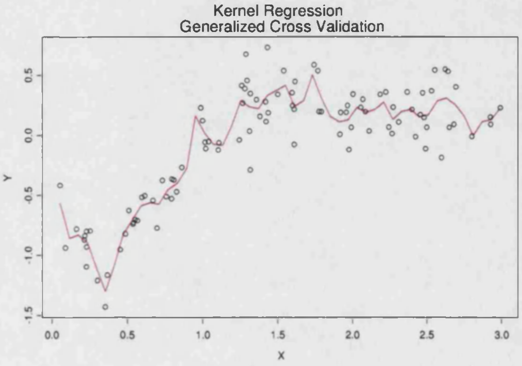


Figure 6.2



Similar findings have been observed in numerous extensive studies of the small sample behaviour of the selectors.[23] So, in small samples the selectors can lead to conflicting models, which raises the question of which selector to choose in practice. Of the two elements defining a selector, the error estimator is more critical than the discrepancy function. An error estimator is required to be consistent, unbiased, and efficient. Consistency is to ensure that the estimator is asymptotically able to correctly estimate the error; unbiasedness is to ensure that the estimates, on average, coincide with the object of inference; and efficiency (i.e., minimum variance) is to ensure that there is no other unbiased consistent estimator yielding a more precise estimate. One thus has to search for a selector that is based on a consistent, unbiased, and efficient error estimator.

Most error estimators described above have been shown to be consistent and asymptotically equivalent (Efron, 1983:328). This means consistency alone cannot help select an optimal error estimator. It is also necessary to consider the finite properties of the estimators, i.e., unbiasedness and efficiency. The problem is that there is no theoretical result as to which type of error estimator is both unbiased and most efficient. Consequently, statisticians have turned to simulation

[23] See Breiman (1992), Breiman *et al* (1992), Efron (1983), Efron (1986), Efron *et al* (1993), Efron *et al* (1997), and Härdle *et al* (1988).

experiments to study the finite-sample behaviour of the estimators. However, the studies have revealed that the estimators are either unbiased but highly variable or biased and less variable.

In a series of simulation experiments, Efron (1983) investigated the finite sample behaviour of the leave-one-out cross validation method, several variants of the bootstrap method, and some other error rate estimators not mentioned here. The studies revealed that the leave-one-out cross validation estimator was of low bias but suffered from a high degree of variability across different samples of fixed size. Other estimators in the study showed either high bias and less variability or high variability and low bias. Comparing the bias and variance of the estimators, Efron observed that a bootstrap estimator, called the 0.632 estimator, though biased, was comparatively less variable.[24] He recommended it for model selection. Likewise, Breiman et al. (1992) compared the finite sample properties of leave-one-out cross validation, *k*-fold cross validation, and a variant of the bootstrap method in a number of subset (variable) selection experiments. The simulations showed that the leave-one-out cross validation had low bias but suffered from high variability while five-fold cross validation method suffered from a large bias and less variability. Comparing the results, the authors suggested using the ten-fold cross validation method. Finally, Efron et al. (1997) report a number of simulation studies that support a bootstrap estimator different than the 0.632 estimator. If a lesson can be learnt from these studies it is that no error estimator outperforms others in all respects. Either they are unbiased but highly variable or they are biased and less variable. A judgement is needed about the relative importance of unbiasedness and efficiency to pick out an estimator.

Beside this, the real problem with the use of simulation studies is that their results cannot be generalised automatically. The fact that an estimator outperforms others in a series of simulations does not imply that it always outperforms others. In experiment with different models a different estimator may outperform the rivals. To give a historically interesting example, as said earlier, in a series of studies Efron (1983) found that the 0.632 bootstrap estimator outperformed several other

[24] For a definition of this error estimator see Appendix *C*.

methods, and so recommended it for estimating prediction error. Not long after, Breiman et al. (1984) noted that the estimator badly fails in predicting the error rate of highly overfit models, such as a one nearest neighbour classifier (estimator), where the apparent rate of error is zero.[25] For example, if Y takes either 0 or 1 with probability ½, independently of (useless) predictor vector X, then, the true error rate for any classifier equals 0.50. Yet, the 0.632 estimator predicts the expected error rate of a one nearest neighbour classifier to be $0.632 \times 0.5 = 0.316$. In this case, both the leave-one-out cross validation estimator and the simple bootstrap estimator correctly predict the error rate of ½. Similar counterexamples have been found for hold-out error rate estimators. For example, in a no-information dataset, where the assignment of cases to each class is completely random (e.g., Fisher's iris data set), the best an estimator can predict is to predict majority.[26] But if the number of cases for each class in the dataset happened to be equal, the leave-one-out cross validation method would wrongly predict 0% predictive accuracy for a majority prediction rule (Kohavi, 1995). The hold-out methods including the cross validation techniques work only if leaving part of the data aside as a test set does not destroy the structure of the data. The validity of simulation results is then confined to the type of models (and data) considered and cannot be generalised automatically (White, 1992:110). All in all, the question of which selector to choose in practice currently has no theoretical answer. In the end, the choice of a selector is to some extent left to the modeller's judgement (Leamer, 1982:217):

> In this paper I have compared several simple criteria on the basis of which we can select one regression equation among many other candidates. ... the general picture that has emerged from this paper is that all of the criteria considered are based on a somewhat arbitrary assumption which cannot be fully justified, and that by slightly varying the loss function and the decision strategy one can indefinitely go on inventing new criteria. This is what one would expect, for there is no simple solution to a complex problem (Amemiya, 1980:352).

[25] A K-nearest neighbour classifier considers K nearest neighbours and assigns the class by majority vote.

[26] Fisher's iris dataset, a well-known database in the pattern recognition literature, contains 3 classes of 50 instances each, where each class refers to a type of iris plant (Fisher, 1936).

In conclusion, the predictive model selectors do not provide an entirely *objective* (data-driven) solution to the model selection problem. They only provide an *automatic* solution in the sense that, given the choice of a discrepancy function and an error rate estimator, the method fixes the model that minimises the error, as measured by the estimator (Green et al., 1994:24). The idea of designing an inference procedure that receives data and yields the model that given the data best approximates the true model has no satisfactory foundation. Any nonparametric inference is founded at a deep level on a decision about the optimal level of smoothing that cannot be fully justified by the data:

> The absence of theoretical guidance on setting the bandwidth, and more generally on defining nearness, leaves the empirical researchers with enormous discretion. This discretion gives applied nonparametric regression analysis a subjective flavor (Manski, 1991:44).

6.3 Extrapolation Error

The difficulty in choosing an optimal selector is not the only factor that limits the power of the model selection strategy available in nonparametric statistics. Even if it were possible to locate an optimal selector, there would still be no satisfactory, and entirely data-driven, solution to nonparametric model selection. An explanation of this point calls for making a distinction between two notions of prediction error: *in-sample* and *extra-sample* prediction error. In sample prediction error (accuracy) refers to the predictive performance of a model at the locations in the input variable space from which the data have been drawn. We refer to these locations in the input variable space as the *sample region*. On the other hand, extra-sample prediction error refers to the predictive performance of a model over the locations in the input variable space for which no data are available. Given this distinction, an important point to note is that in-sample predictive accuracy is not necessarily the evidence for extra-sample predictive accuracy. The reason is that two models can be exactly alike over the sample region but behave considerably differently outside the region. In that case, if the data produced by one of the models belonged to the points in the input variable space where the models are alike, the other model would equally predict the data despite the fact that it yields quite wrong predictions elsewhere.[27] Consider the following two models:

(I) $y = f(x) + e_1$, $f(x) = 1/2 + 1/2\tanh(x-2)$, $e_1 \sim N(0,0.2)$

(II) $y = g(x) + e_2, g(x) = 0.05 - 0.2x + 0.3x^2 - 0.002x^4$, $e_2 \sim N(0,0.2)$

where X takes values in interval [0,6.5].[28] As shown in Figure 6.3, these models are alike over interval [0,4] but fall apart over interval [4,6.5]. Suppose model (I) was true. If the data happened to fall in the first interval, where both models are alike, model (II) would equally predict the data well. But the predictive performance of the model over this interval would give a wrong indication of how it would perform over the second interval. Therefore, an estimate of in-sample prediction error (accuracy) cannot be taken as an estimate of extra-sample prediction error (accuracy). On the same ground, the fact that a model minimises in-sample prediction error is not a guarantee that it also performs well outside the sample region. To decide how to extrapolate beyond the data, it is essential to have an estimate of the extra-sample accuracy of the models considered. An estimate of in-sample accuracy is neither necessary nor sufficient.

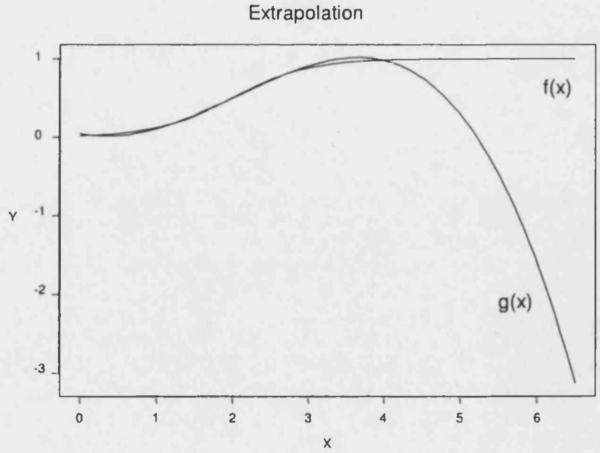


Figure 6.3

[27] This is another way of stating Goodman's riddle of induction (Goodman, 1955). See Howson (2000) for an exposition of the riddle.

[28] This model has been constructed based on a similar example found in Forster, 2000.

Now, any prediction error estimator is an estimator of in-sample prediction error. Consider cross validation or bootstrap estimators. These estimators essentially work by correcting the optimism of the apparent rate of error (Efron and Tibshirani, 1993:249), which is a measure of how a model predicts the same data used to obtain it.[29] The correction is to remove the effect of noise in the data so as to enable the error estimator to correctly estimate the prediction error of future observations drawn at the same locations in the input space from which the data have been drawn. The resulting selectors, therefore, are only able to locate a model that minimises in-sample prediction error. In other words, they can only tell what sort of model will yield accurate prediction if we draw a 'similar' sample, where by similarity we mean a sample drawn at the same points in the input space from which the original sample was drawn. The selectors are silent about the model that is true of the population, and consequently give no guidance as to how to generalise beyond the sample region (Browne, 2000:8).

This conclusion has an important implication for nonparametric inference. In general, if it were possible to densely populate the input region everywhere with data, every prediction would involve only in-sample prediction and, as a result, a cross validation estimate of a model's error, for instance, would provide an estimate of the predictive accuracy of the model with respect to the population. In that case, the distinction between in-sample and extra-sample prediction error would be irrelevant. However, the discussion of the curse of dimensionality makes it clear that in 'interesting' inference situations, the input space is almost everywhere empty, which means nonparametric prediction in 'interesting' inference situations almost always involves extrapolation (extra-sample prediction). Since the model selection criteria are silent about the predictive performance of a model outside the sample region, nonparametric extrapolation is almost always arbitrary (Geman et al, 1992:44). As a consequence, in interesting inference situations such as modelling a choice situation that involve a relatively large number of variables, reliable prediction with ordinarily sample sizes necessarily calls for substantive prior background information. That is to say, one

[29] This point is evident from the reformulation of the leave-one-out cross validation given in (6.9).

needs to posit *a priori* a parametric model and be sure that the model is correctly specified:

> One can usually be confident that the regression of interest is continuous. Hence one can usually trust nonparametric estimates to be consistent. On the other hand, these estimates are often imprecise in practice. Moreover, they cannot be extrapolated off the support of x. Parametric modelling permits more precise estimation and makes extrapolation possible. The problem, of course, is that an assumed parametric model may be misspecified (Manski, 1991:44).

Overall, the difficulty in locating the 'best' error rate estimator is not the only trouble with the predictive approach to model selection in nonparametric inference. The more serious problem is that in interesting nonparametric inference settings, such as modelling a complex choice situation, prediction almost always involves extrapolation about which data are absolutely silent. Nonparametric extrapolation in interesting cases is inevitably arbitrary. Nonparametric models are only reliable for in-sample prediction.

7 Conclusion

This chapter began with the remark that economics is in need of a theory that explains how the agent learns about the economy, defines his choice situation, and redefines it in response to policy interventions. A basic unifying hypothesis in new classical economics has been that he behaves like an econometrician (the IS hypothesis). Theoretical economists hope this conjecture helps them predict the model that the agent builds of his choice situation based on available economic data. The objective is to combine the information with information about the agent's preferences and budget constraint to specify the decision problem he is trying to solve, which is essential for predicting his behaviour. The success of this hypothesis first and foremost depends on the existence of a 'tight enough' theory of statistical learning that describes how, given a data set, the statistician constructs a model of the mechanism generating the data.

To discuss some aspects of this issue, we began by showing that any statistical inference necessitates three types of assumptions, which define a model.

Therefore, the central concern of a theory of statistical inference should be model specification. We then looked at nonparametric statistics, which suggests starting with a general and highly flexible model (distribution assumption) and leaving to the data to determine the precise form of the model. We first described how nonparametric statistics provides a way of learning a density or regression function from a random sample without having to posit a rigid model in advance. We then explained how the idea might be extended to non-random data. If nonparametric inference could be accomplished in practice, we would have, at least in the case of random samples, inference procedures that receive data and yield the best approximation of the underlying distribution given the data.

As seen, in order for a nonparametric estimator to deliver a good approximation of a function both the variance and squared bias component of the *MSE* error of the estimator must approach zero. Because of the bias-variance dilemma, this is only possible by densely populating the input variable space. But, due to the curse of dimensionality problem, it is practically impossible to densely populate input variable spaces in interesting inference situations, where the number of input variables considered exceeds three or four. In such situations, local averaging (fitting) inference demands an astronomically large sample that is usually impossible to achieve in practice. With a reasonably sized sample, a good approximation of the relations among several variables using local averaging-based techniques is impossible.

This impossibility also rules out the practical possibility of extending nonparametric inference to non-random data. The extension requires estimating the joint probability distributions of various subsets of the variables under study in order to assess the appropriateness of alternative independence and homogeneity assumptions. Since nonparametric density estimation of high dimensional data is practically impossible, the choice of appropriate independence and homogeneity assumptions cannot be left to nonparametric methods either. Modelling the probabilistic relations among a set of variables characterising a choice situation requires beginning with substantive probabilistic assumptions, which essentially means one has to work within the framework of parametric inference. In that case, learning will only be possible if the model is correctly specified.

The bias-variance trade-off implies that, given a data set, there is an optimal value for the smoothing parameter of a nonparametric estimator that fixes the class of functions that it can approximate. The only avenue available to nonparametric statistics for specifying the optimal smoothing parameter value is to consider the predictive performance of various models arising from alternative smoothing parameters. We have seen that there are competing procedures for nonparametric model selection. Although asymptotically equivalent, the methods pick up different smoothing parameters in practice, leading to different models. There is, however, no general theoretical consideration that can help to choose a model selector. The performance of the methods depends on the target function. For some functions, cross validation-based techniques may, for instance, work better and, for some others, other methods may work better. As a consequence, in a purely nonparametric inference situation, there is an element of arbitrariness in the choice of a model selector and hence a model.

Even after choosing a model selector, the problem of smoothing parameter selection is not entirely resolved, as these selectors often have local minima. Moreover, the estimators underlying the selectors only measure in-sample prediction error. What they at best tell us is how to simplify the model so as to avoid overfitting. They do not tell us how to extrapolate beyond the sample. Taken together, these considerations rule out the possibility of inventing procedures that receive data and yield the best possible approximation of the underlying model given the data. Data only speak in the light of background information, and, as the information differs, they speak differently.

There are also numerous nonparametric estimators. Besides the family of kernel estimators, one may also mention nearest neighbours estimators, spline regression methods, neural network methods, local polynomials, and many others. When the concern is with arbitrarily large samples, this multiplicity may pose no problem, since all these methods are consistent. However, when the sample is small, they often produce different estimates, raising the question of which method to choose in practice (Breiman, 1992). Modelling learning also demands some decision about the agent's choice of an estimator.

Finally, due to the curse of dimensionality, in high dimensional input variable spaces, nonparametric models take the form of expansions in low dimensional functions. In such models, the relation between the dependent and independent variables are entirely blurred, which seriously limit their use in analysing how the dependent variable would vary if the independent variables were changed by intervention. Consequently, the models are not useful for analysis of actions and policies. Analysis of actions and policies requires an interpretable model, which necessitates working with a parametric model from the start.

These limitations of nonparametric inference define the boundaries of any theory of learning that fails to take seriously the role of non-sample information. It is a combination of background information and sample data that enables a person to come up with an intelligible model of a choice situation. A theory of learning ought to explain how non-sample information is obtained, how the information interacts with sample data, and how the interaction leads to a specific model. Some of these issues will be discussed in the next chapter that concentrates on the theory of Bayesian inference.

Appendices

Appendix A: Product kernel independence

Notice that although the estimator

$$\hat{f}(x,y) = \frac{1}{Nh^2}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) K\left(\frac{y-y_i}{h}\right) \qquad \text{(A1)}$$

uses kernel independence, this does not imply that the variables X and Y are independently distributed. If the variables were assumed to be independent, the kernel estimator would have the form

$$\hat{f}(x,y) = \left(\frac{1}{Nh_x}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h_x}\right)\right) \times \left(\frac{1}{Nh_y}\sum^{N} K\left(\frac{y-y_i}{h_y}\right)\right). \qquad \text{(A2)}$$

Appendix B: Decomposition (Eubank, 1989)

This result is based on a lemma established in Searle (1971, Chapter 2) and mentioned in Eubank (1988: 402). Let $\mathbf{Z}$ be a $n \times 1$ vector with mean $\mathbf{m}$ and variance-covariance matrix Σ. Suppose W is a symmetric $n \times n$ matrix. Then

$$E(\mathbf{Z}'W\mathbf{Z}) = \mathbf{m}'W\mathbf{m} + tr(W\Sigma) \qquad \text{(B1)}$$

where $tr(W\Sigma)$ is the trace of $W\Sigma$.

Now let $\{(y_i, x_i), \dots, (y_n, x_n)\}$ be a vector of observations from

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon$$

where $\mathbf{y}$ is the vector of responses, $f(\mathbf{x})$ the vector of unknown means, and ε the vector of zero mean, uncorrelated random errors with common variance σ^2. Further let $\hat{f}_h(\mathbf{x})$ be a linear estimator of $f(\mathbf{x})$. The mean average squared residuals for $\hat{f}_h(\mathbf{x})$ is given by

$$E(ASR(h)) = n^{-1}\sum_{i=1}^{n} E\{y_i - \hat{f}_h(\mathbf{x}_i)\}^2, \qquad \text{(B2)}$$

which can be rewritten as

$$E(ASR(h)) = n^{-1}E(\mathbf{y} - W(h)\mathbf{y})^2 \qquad \text{(B3)}$$

$$= n^{-1}E[(\mathbf{y} - W(h)\mathbf{y})(\mathbf{y} - W(h)\mathbf{y})]$$

$$= n^{-1}E[\mathbf{y}'(\mathbf{I} - W(h))(\mathbf{I} - W(h))\mathbf{y}]$$

$$= n^{-1}E[\mathbf{y}'(\mathbf{I} - W(h))^2\mathbf{y}] \qquad \text{(B4)}$$

where $W(h)$ is the smoother matrix and $\mathbf{y}$ the vector of responses. Let $\Sigma = \sigma^2\mathbf{I}$, and note that $W(h)$ is symmetric. Applying (B1) to (B4) yields the result.

$$= n^{-1}f(\mathbf{x})'(\mathbf{I} - W(h))^2 f(\mathbf{x}) + n^{-1}\sigma^2 tr[(\mathbf{I} - W(h))^2]$$

$$= n^{-1}f(\mathbf{x})'(\mathbf{I} - W(h))^2 f(\mathbf{x}) + \sigma^2 + n^{-1}\sigma^2 tr[(W(h))^2] - 2n^{-1}\sigma^2 tr[W(h)] \qquad \text{(B5)}$$

Now consider applying the technique to average mean prediction error

$$APE(h) = n^{-1}\sum_{i=1}^{n} E(y_i^* - \hat{f}_h(x_i))^2 \qquad \text{(B6)}$$

which can be restated as

$$= \sigma^2 + n^{-1}\sum_{i=1}^{n} E(f(x_i) - \hat{f}_h(x_i))^2$$

$$= \sigma^2 + n^{-1}E[f(x) - \hat{f}_h(x))'(f(x) - \hat{f}_h(x))]$$

$$= \sigma^2 + n^{-1}E[f(x) - W(h)\mathbf{y})'(f(x) - W(h)\mathbf{y})]$$

$$= \sigma^2 + n^{-1}E[f(x) - W(h)(f(x) + \varepsilon))'(f(x) - W(h)(f(x) + \varepsilon)]$$

$$= \sigma^2 + n^{-1}f(x)(\mathbf{I} - W(h))^2(f(x) + n^{-1}\sigma^2 tr[(\mathbf{I} - W(h))^2] \qquad \text{(B7)}$$

which is the same as the equation (6.4) in the text.

Appendix C: Bootstrap Estimates of Prediction Error (Efron et al. 1993)

Without any loss of generality, let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be an IID sample from bivariate distribution π, $\hat{f}$ be an estimate of the regression function f, and $\hat{f}(x_i)$, the predicted value of Y at point x_i.

Let $\Delta = [y_i, \hat{f}(x_i)]$ denote a measure of error between the response y_i and prediction $\hat{f}(x_i)$. In regression, $\Delta = [y_i, \hat{f}(x_i)]$ is often chosen to be $[y_i - \hat{f}(x_i)]^2$.

Let denote the prediction error for $\hat{f}$ by

$$Perr(D,f) = E^*_\pi\{\Delta[y^*, \hat{f}(x^*)]\}, \tag{C1}$$

where the expectation is taken over a new observation (x^*, y^*) from distribution π. The apparent error rate is

$$Aerr(D,\hat{f}) = \frac{1}{N}\sum_1^N \Delta[y_i, \hat{f}(x_i)]. \tag{C2}$$

Let $D^b = \{(x_j^b, y_j^b)\}_{j=1}^N$ be a bootstrap sample. The simplest bootstrap error estimator generates B bootstrap samples, estimates the model on each, and then applies it to the *original sample* to give B estimates of prediction error:

$$err(D^b,\hat{f}) = \frac{1}{N}\sum_1^N \Delta[y_i, \hat{f}_b(x_i)] \tag{C3}$$

In this expression, $\hat{f}_b(x_i)$ is the predicted outcome at x_i based on the model $\hat{f}_b$ estimated from bootstrap data set D^b. The overall prediction error estimate is the average of these B estimates:

$$\overline{Perr}_{boot} = \frac{1}{B}\sum_{b=1}^B \sum_1^N \Delta[y_i, \hat{f}_b(x_i)]/N \tag{C4}$$

$\overline{Perr}_{boot}$ is not a good estimate, since the training and test samples overlap, causing an underestimation of prediction error $Perr(D,\hat{f})$.

A second way to employ the bootstrap technique is to estimate the bias (or optimism) of the apparent error rate $Aerr(D,\hat{f})$ as an estimate of prediction error $Perr(D,\hat{f})$ and obtain an estimate of the error by adding the bias term to $Aerr(D,\hat{f})$. Let denote the optimism by

$$\omega(\hat{f}) = Perr(D,\hat{f}) - Aerr(D,\hat{f}) \tag{C5}$$

The bootstrap estimate of $\omega(\hat{f})$ is given by

$$\hat{\omega}(\hat{f}) = \frac{1}{B.N}\{\sum_{b=1}^B \sum_1^N \Delta[y_i, \hat{f}_b(x_i)] - \sum_{b=1}^B \sum_1^N \Delta[y_{bi}, \hat{f}_b(x_{bi})]\}. \tag{C6}$$

An alternative bootstrap estimate of the prediction error is the apparent error plus the downward bias in the apparent error given by

$$\overline{Perr}_{boot}(D,\hat{f}) = err(D,\hat{f}) + \hat{\omega}(\hat{f}). \tag{C7}$$

For each data point (x_i, y_i) the bootstrap samples can be divided into those that contain (x_i, y_i) and those that do not. The prediction error for the data point (x_i, y_i) will likely be smaller for a bootstrap sample containing the point. It can be shown that the percentage of points belonging to both the original sample and the bootstrap sample is approximately 63.2%. A possible way to remedy this problem can be to take as test samples only those data point that do not belong to D^b. That is,

$$\overline{Perr}_{boot3}(D,\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B_i}\sum_{i\in I_i}\Delta[y_i - \hat{f}(x_i)] \tag{C8}$$

where I_i is the set of indices of the bootstrap sample D^b that do not contain (x_i, y_i), and B_i is the number of such bootstrap samples. However, since the samples used to obtain $\overline{Perr}_{boot3}$ have no common elements with the test samples, they are likely to give rise to a pessimistic estimate of prediction error. On the other hand, 63.2% of the samples containing (x_i, y_i) are, as said, likely to lead to an optimistic estimate of prediction error. The so-called bootstrap 0.632 estimator is defined by the weighted average of the apparent error estimate $Aerr(D,\hat{f})$ and the error estimate $\overline{Perr}_{boot3}$:

$$\overline{Perr}_{.632} = 0.368\times Aerr(D,\hat{f}) + 0.632\times\overline{Perr}_{boot3} \tag{C9}$$

Chapter 4

Homo Economicus as an Intuitive Statistician (2)

Bayesian Diagnostic Learning

1 Introduction

> Learning takes place through Bayesian updating of the individual prior beliefs. ... However, since the use of Bayesian updating is a consequence of expected utility maximisation, assumption (2) [Bayesian updating] is already a consequence of assumption (1) [subjective expected utility maximisation] (Kalai, et al., 1993:102).

> "...hypothesis and model generation is far more important to problem solving that is hypothesis testing and that it is very much the statistician's business to be involved with model generation and regeneration" (Box, 1994:218).

The last chapter began the study of the bounded rationality program that viewed the economy as a society of intuitive statisticians – the intuitive statistician hypothesis. A basic question for the program is whether there is a 'tight enough' theory of statistical inference. It has so far been argued that there can be no entirely data-driven procedure that receives a finite sample of data and yields the model that, given the data, best approximates the mechanism generating the data. To build an interpretable model of a choice situation, it is necessary to start with substantive probabilistic information or, more precisely, a parametric probability model. In order to advance our analysis, this chapter studies the intuitive statistician hypothesis within the framework of the Bayesian theory.

There is now a sizeable literature on learning in economics that models the agent as a Bayesian statistician. These studies consider economies of Bayesian statisticians, who know the true economic model except for a small number of parameters, and use Bayes' theorem to learn the parameters from the data generated by the economy. The papers investigate the conditions under which the opinions of these intuitive statisticians converge on the true parameter values. As noted earlier, since feedback from learning can shift the structure, the issue in question is not ordinary parameter estimation but involves estimation of shifting parameters. Various convergence theorems of probability theory have been employed to demonstrate that if the agents do not entertain extreme priors excluding the true parameter values, they eventually learn the parameters with probability one. This result is often claimed to justify the use of the solution concepts of rational expectations equilibria in solving economic models and Nash equilibria in game theory. Good reviews of, and original contributions to, the literature on Bayesian learning are found in Blume and Easley (1995), Bray and Kreps (1986), Cyert and DeGroot (1974), Kiefer and Nyarko (1995), Nyarko (1997), Nyarko (2000).

As explained in the last chapter, the relevance of such studies to the study of the economy is unclear. The studies assume that the agents know the true model except for a finite number of parameters, providing no explanation of how the model has been learnt in the first place. This is a critical issue because starting with a mis-specified model can make learning of rational expectations impossible (Nyarko, 1991). Moreover, the results are invariably of an asymptotic nature, and thus do not bear on real inference situations where the samples are usually small. In fact, the economic structure can shift for reasons other than learning, rendering past data irrelevant. A theory of human learning should first and foremost explain how a person builds a model of his choice situation from ordinarily available samples. This chapter, therefore, departs from the dominant trend in the studies of Bayesian learning by exclusively focusing on the issue of parametric model formulation and problems arising in learning from small samples.

This chapter starts by arguing that the Bayesian theory is solely concerned with coherent (consistent) analysis of uncertainty regarding a closed set of specified possibilities (events, hypotheses, models), which are assumed to be adequate as a description of the (inference) situation at hand. Coherent analysis constitutes only one phase out of several in the whole process of statistical learning. A vital activity preceding coherent analysis is the initial generation of models. Another critical activity following coherent analysis is appraising the empirical adequacy of the models (Smith, 1986:250). In practice, these phases of learning are iterated in a cyclical manner. New data cast doubt on the adequacy of the current models, calling for generation of new models. Construction of the new models necessitates forming a new coherent system of beliefs, which in turn raises the question whether the new models include a model that captures the salient features of the data. A satisfactory account of statistical learning should explain how models are built, how they are assessed, and how they are modified. This chapter aims to generalise the framework of Bayesian inference by introducing some additional proposals to shed light on those aspects of inference such as model formulation that are usually left unexplained in Bayesian statistics. Having done so, it spells out the implications of the broader theory of Bayesian inference for the bounded rationality project.

The chapter is organised as follows: Section 2 touches on some issues in the foundations of the Bayesian theory. Section 3 explains how the idea of coherence is translated into a theory of statistical inference. Section 4 outlines various phases of the process of inference from data, and places coherent analysis within a wider view of inference. Section 5 begins spelling out an exploratory theory of model specification by explaining the process of initial model formulation. Section 6 provides a framework for Bayesian empirical model assessment, and completes the exploratory theory of model specification. Section 7 takes up Bayesian model selection. Section 8 discusses some objections levelled against the theory of Bayesian model specification. Section 9 spells out the implications of the broad view of Bayesian learning for the bounded rationality project.

2 Foundational Issues

In economics, a choice situation (the environment) is viewed through the perspective of a collection of measurable quantities. These quantities are of two kinds: those whose numerical values are known and those that are not known. The general problem facing the modeller is to infer the unknown quantities from the known ones. Knowledge of the known quantities usually fails to determine uniquely that of the unknown quantities, and given the known quantities, there remains uncertainty about the values of the unknowns. The hallmark of the Bayesian position is that our uncertainty attitudes towards these unknowns should accord with the laws of probability. A foundational question is whether the Bayesian theory prescribes how these uncertainties should be updated as some of the unknowns become known. To address this query, it is first essential to understand the reasons why subjective uncertainties ought to accord with the laws of probability.

The second chapter studied some aspects of the decision theoretic approach that took personal probability to be part of a theory of coherent preferences in the face of uncertainty. Since, in this chapter, the concern is not directly with decision-making but with learning from data, it is convenient to consider another approach to establishing the probability axioms, which makes no formal reference to preference considerations. The approach is based on the so-called Dutch book (DB) theorem, which aims to justify the probability axioms as rationality constraints by showing that partial beliefs are 'coherent' if and only if they conform to the axioms. This section studies the assumptions underlying the theorem to explore if they impose any restriction on how learning from experience should take place. The exposition is mainly built on Skyrms (1986) and Howson's various writings.[1] The theorem stands on three basic assumptions:

The first assumption consists of two related components: One is that you (the agent) have a degree of belief in any hypothesis *H* you may ever consider. The other is that the strength of your belief in *H* is reflected in the price that you are ready to pay in a bet on or against it.[2] It is, therefore, considered as possible to measure your degree of belief in *H* in terms of the price you are ready to pay in some appropriate bet on or against it. Several definitions are in order to elaborate on this point. A bet on a statement *H* is an arrangement between you (the bettor) and the bookie whereby you pay the bookie amount *$d* to receive amount *$c* if *H* is true and receive nothing if *H* is false.[3] The total amount involved, (*d*+*c*), is called the *stake*, the ratio *d/c* the *odds*, and the ratio *d/d+c* the *betting quotient.*[4] Finally, the price that you are ready to pay for a bet in which the stake and whether you bet on or against *H* is decided by your opponent is considered to be *fair* in your eyes. Given these preliminaries, the first assumption identifies your degree of belief in *H* with the betting quotient in a bet on or against *H* whose price you consider as fair (Howson, 2000:126). Following de Finetti (1980), a bet is

[1] Classic sources for the theorem are F.P. Ramsey, (1926 [1980]) and B. de Finetti, (1937 [1980]).

[2] This assumption can be weakened. All that is needed is that if you have a degree of belief in *H*, it is reflected in the price you are ready to pay for a bet on or against *H*.

[3] The dollar sign is omitted in what follows.

[4] "[B]etting quotients are ... just odds normalized so that they lie within the half-open interval [0,1); this is extended to the closed-unit interval [0,1] by allowing the odds to take the 'value' ∞" (Howson, 2000:125).

sometimes defined differently. Let *s* stand for the stake in the above bet and *p* for the betting quotient. The bet can be restated as an arrangement in which you agree to pay *ps* in order to receive *(1-p)s* if *H* turns out true and nothing otherwise. Your fair betting quotient *p* then represents your degree of belief in *H*.

A corollary of the definition of a bet, which plays a vital role in the DB theorem, is that the sum of a collection of bets on some propositions, under certain conditions, determines a bet on another proposition. Note that a bet on a statement *H* admits only two possibilities – *H* is true or *H* is false – and specifies a unique payoff in each case. The sum of a collection of bets then equals a new bet if it admits only two possibilities and specifies a unique payoff in each case. As a simple illustration, consider the case involving two mutually exclusive propositions *P* and *Q*. Let B_1 be a bet on *P* with stake *s* and betting quotient *p*, and B_2 be a bet on *Q* with stake *r* and betting quotient *q*. If the stakes *s* and *r* are equal, then, the sum of these bets is equivalent to a bet on *PvQ* with stake $s^* = s = r$ and betting quotient $p^* = p + q$ (Skyrms, 1986:176). Table 4.1 shows this:

Table 4.1

Equivalent Bets (1)

		Bet 1	**Bet II**	**Sum of Bet I and II**	**Bet III**
P	***Q***	On *P*	On Q		On PVQ
T	F	(1-p)s	-qr	(1-p)s-qr	(1-p*)s*
F	T	-ps	(1-q)r	(1-q)r-ps	(1-p*)s*
F	F	-ps	-qr	-(ps+qr)	-p*s*

The second assumption underlying the DB theorem is that the value of the sum of a set of bets is the total value of the bets and, therefore, if a set of bets is regarded as individually fair, they are also considered as collectively fair. To give an example, if in the above situation the betting quotients *p* and *q* are viewed as fair, the betting quotient *p** for the third bet is also viewed as fair. Schick (1986) was the first to note the significance and independence of this assumption in establishing the DB theorem, calling it the value additivity assumption. The principle, he argues, presumes that the value that people assign to a bet is independent of whether other bets are in effect. But people are usually risk averse. If they have already committed themselves to a bet, the highest price that they would pay for a new bet is less than it otherwise would be (1986:114). In such cases, people are hedging against the possibility of losing both bets, and there is nothing irrational about this behaviour. The value additivity principle cannot, therefore, be taken for granted.

The literature provides several considerations in support of the value additivity assumption. Skyrms (1986:179) defines a fair bet as a bet with expected utility zero, and seeks to derive the assumption from this definition. This move seems to assume that your belief distribution obeys the probability calculus, which undermines the appeal of the DB theorem as an independent approach to establishing the probability axioms.[5] Howson (2000) defends a view of the probability calculus as an extension of deductive logic to partial beliefs. In this setting, he envisages a parallel between the value additivity assumption and the closure principle applied in deductive logic. Just as the closure principle is taken for granted in deductive logic to define the truth-value of a compound sentence in terms of the truth-value of its components, it is equally 'natural', Howson suggests, to take value additivity for granted to determine the value of a compound bet from the value of its components (2000:129).[6] This suggestion is plausible but applies only when the concern, as in deductive logic, is solely with bets that are *simultaneously* made. The proposal does not counter Schick's worries in sequential betting scenarios. The validity of the principle, if valid at all, is confined to static betting scenarios.

An implication of the value additivity principle plays a crucial role in establishing the DB theorem. Note that a fair bet can informally be interpreted as a bet that

[5] For a further discussion of this point see Howson (1995:4-5)

[6] In a new manuscript, Colin Howson (2004) substantially reformulates the argument for the probability axioms as consistency constraints on partial beliefs, effectively rejecting the traditional formulation embodied in the DB theorem. In this new setting the value additivity assumption is introduced "as a constraint on the solution assignment of fair betting quotients" (2004:18). The new formulation more vividly supports the conclusions drawn in the text about the scope of the Bayesian theory.

confers zero advantage to either side. Since any sum of zeros is zero, the net advantage of a collection of fair bets is also zero. Given the value additivity assumption, if you consider a collection of bets as individually fair but the net advantage of the bets is nonzero, then the *only* explanation is that you are evaluating a bet (or equivalent bets) at two different rates, regarding both as fair.[7] In that case, it is possible for a cunning bookie to invite you to accept a set of bets that all are individually fair in your eyes but, taken together, lead you to a sure loss. The trick for the bookie is simply to sell you the bet at your higher fair price and buy back an equivalent bet or an equivalent set of bets at your lower fair price. A collection of bets that guarantees a loss no mater what the outcome of the events upon which the wagers are made is called a Dutch book (Skyrms, 1986:185).

The third assumption is a coherence (rationality) condition. Some statements of the DB theorem identify the condition with a simple behavioural criterion – essentially that a rational agent ought to avoid a combination of decisions that leads to a sure loss (Dawid, 2002:3). For several reasons, discussed in Christensen (1991), this criterion fails to support the laws of probability as rationality constraints on partial beliefs. In a nutshell, there are situations where a person accepts a combination of bets which leads to a sure loss but does not *actually* hold any beliefs violating the probability axioms. The person may recognise, for instance, that a collection of bets offered to him by a friend leads to a sure loss but accepts them to avoid harming her confidence. Such a decision is not usually considered as *irrational.* On the other hand, a person may have beliefs that breach the laws of probability or even logic but consciously refuses to participate in any decision which entails a sure loss. In such cases, even though he actually escapes a sure loss, there still seems to be something amiss about his beliefs. If the concern is to establish the probability axioms as rationality constraints on partial beliefs, the coherence condition must do more than pointing to some dire practical consequences; it must directly be concerned with relations among beliefs (Christensen, 1991:238).

[7] It is here assumed that the sum of the bets is equivalent to an additional bet, which is not generally the case.

Another notion of coherence appears in Ramsey's brief allusion to the DB theorem in his seminal work "Truth and Probability" (1926 [1980]). There, he regards the theory of probability as "an extension to partial beliefs of formal logic, the logic of consistency" (1980:41).[8] From this perspective, the underlying notion of coherence is logical consistency. Logical consistency directly deals with the internal structure of a belief system, and can well support a justification of the probability axioms as rationality constraints on partial beliefs. In what follows, we therefore build our analysis around this notion of coherence, which has also increasingly been adopted in the recent philosophical literature on the Bayesian theory.

These assumptions state all that is needed for proving the DB theorem. The proof starts by establishing that if your fair betting quotients for a collection of bets violate the probability axioms, a Dutch book can be made against you. In other words, there exists a finite series of bets that you consider as individually fair but collectively lead to a sure loss. The converse of this result is also shown to be true. If your fair betting quotients conform to the probability axioms, no Dutch book can be made against you. Given the value additivity assumption, the susceptibility to a Dutch book is the evidence that you are rating two equivalent bets at two different prices, considering both as fair. This means you believe in a pair of contradictory propositions that a bet is simultaneously fair and unfair. Since conformity with the probability calculus is both necessary and sufficient to avoid a Dutch book, the only way to avoid such a contradiction is to arrange your fair betting quotients or, in other words, your partial beliefs, in accordance with the probability axioms. And since logical consistency is a rational desideratum, the laws of probability become rationality constraints on partial beliefs.[9]

[8] As Howson (2004:5-6) points out, Ramsey set forth this view of the laws of probability within the theoretical framework of axiomatic utility, not the theory of logic. Recent defenders of epistemic probability have made every effort to entirely disentangle the proof of the probability axioms from formal utility considerations (See Howson, 2004).

[9] Classic statements of the DB theorem only establish finite additivity. Williamson (1999) has extended the theorem to countable additivity.

The notion of conditional probability plays a key role in understanding whether the Bayesian theory furnishes a model of learning from experience. It is therefore useful to review the DB argument for the quotient rule

$$P(H / E) = P(H \& E) / P(E),$$

which relates conditional probability to non-conditional probabilities. A key element in the argument is a definition of conditional bet, rooted in de Finetti's writings. He defines a bet on *H* conditional on *E* as a bet on *H* that proceeds if *E* turns out to be true and is called off if *E* is false (1980:69). Thus, the conditional probability $P(H / E)$ is taken to stand for the price at which you will buy or sell a bet that pays $1 if *H* is true, with the understanding that the purchase is called off if *E* turns out to be false. Another element is the fact that the sum of a bet on *H&E* and a bet against *E*, when the loss of the first bet is the wining of the second bet, is equivalent to a bet on *H* conditional on *E* (Skyrms, 1986:189). To be precise, let *q* be your fair betting quotient for a bet on *H&E* with stake *r*, and *r* be your fair betting quotient for a bet against *E* with stake *q*. The sum of these bets is equivalent to a bet on *H* conditional on *E* with fair betting quotient q/r and stake *r*, as shown in Table 4.2:

Table 4.2

Equivalent Bets (2)

		Bet 1		**Bet II**	**Sum of Bet I and II**	**Bet III**
E	*H*	On *H&E*	¬*E*	Against *E*		On *H* given *E*
T	T	(1-q)r	F	-(1-r)q	r-q	(1-q/r)r
T	F	-qr	F	-(1-r)q	-q	-(q/r)r
F	T	-qr	T	Qr	0	0
F	F	-qr	T	Qr	0	0

The ratio q/r corresponds to the ratio of the fair betting quotients for bet *H&E* over bet *E*. This suggests that if your fair betting quotient *p* for the conditional bet differed from q/r, there could be a Dutch book made against you. Since the conditional bet is called off if *E* turns out false, the trick to construct such a collection of bets is to introduce an additional bet on *E* with a suitable stake. Specifically, consider a bet on *H&E* and a bet against *E* with betting quotients and stakes as given above. Further consider a bet against H conditional on E with betting quotient p and stake r, as well as a bet on E with stake q-pr. Taken together, these bets lead to a net loss (gain) of $r(pr$-$q)$ regardless of whether H is true or false. Assuming that r is greater than zero, the net loss (gain) will be zero only if p equals the ratio q/r. This happens only if the fair betting quotient for the conditional bet is equal to the ratio of the fair betting quotients of $H\&E$ over E. Like the basic probability axioms, the quotient rule also becomes a theorem of the probability calculus.[10]

The quotient rule has a number of implications including Bayes' theorem,

$$P(H/E) \propto P(E/H)P(H).$$

This theorem is usually thought to express a fundamental model of learning from experience. Savage remarks that, by entailing Bayes' theorem, the theory of coherent preferences gives a natural interpretation of (or at least one important sense) of the phrase 'learning from experience' (1967:596-7). The theorem, he says, "prescribes, presumably compellingly, exactly how a set of beliefs should change in the light of what is observed" (1967:602). A similar view is also commonly held in economics. Kiefer and Nyarko (1995:40) argue that economics needs no assumption beyond the subjective expected utility maximisation assumption to model learning behaviour, since, by implying Bayes' theorem, the assumption yields a rational model of learning. Any additional assumption about how people learn about the economy is clamed to be *ad hoc*.

This interpretation of the role of the theorem is unwarranted, as Ian Hacking argued long ago (1967:315). The theorem is a consequence of the quotient rule, which only says how conditional probabilities ought to be related to non-conditional probabilities where all the probabilities involved refer to the time before the conditioning event is learnt. So, like the quotient rule, the theorem is just a coherence constraint. In more detail, given $P(E/H)$, the theorem constrains the compatible pairs of $P(H)$ and $P(H/E)$; given $P(H)$, it defines

[10] This argument for the quotient rule is based on the argument given in Howson et al. (1993).

the mapping from $P(E/H)$ to $P(H/E)$; given $P(H/E)$ and $P(E/H)$ it fixes $P(H)$; and given $P(./H)$ and $P(H)$ it defines the mapping from E to $P(H/E)$ (Smith, 1986:98). The theorem is silent about where one has to begin. Though it is common to begin with $P(H)$ and $P(E/H)$ and use the theorem to infer $P(H/E)$, one can also start by fixing $P(H/E)$ and use the theorem to search for a pair of $P(H)$ and $P(E/H)$ that is compatible with it. As far as the justification behind the theorem is concerned, both routes are equally permissible (Lindley, 1983:7). Consequently, the theorem is silent about how a set of beliefs should change in the light of what is observed.

Savage's interpretation of Bayes' theorem supposes an extra assumption that the probability of H after having learnt E is the same as the probability of H on the supposition that E were true (Hacking, 1967:317). This assumption is nowadays known as the Bayesian conditionalization rule (BCR). Precisely speaking, the rule states that if your degree of belief in H conditional on E is $P(H/E)$, and you learn E for sure and nothing else, your new degree of belief in H, denoted by $Q(H)$, ought to be the same as $P(H/E)$,

$$Q(H) = P(H/E).$$

Due to the necessity of this assumption, the question posed at the outset becomes whether the rationality considerations behind the probability axioms lend any support to the BCR. A response is found in Teller (1973), who argues that if you violate the rule there will be a finite series of bets that you consider as individually fair but collectively result in a loss no matter the outcomes. This has been taken to support the BCR just as the DB arguments support the probability axioms. We analyse Teller's argument to show why it fails and to hint at why there can be no justification for the rule anyway. To this end, we draw on a simple statement of the argument given in Howson (1997).

Suppose your updating strategy differs from the BCR. This means, upon learning E, you assign to H either a probability less than $P(H/E)$ or a probability greater than $P(H/E)$. Consider the first case where $Q(H) < P(H/E)$. Further, suppose in your opinion $P(H/E)=x$, $P(E)=y$, and $Q(H)=z$. In this case, a bookie can ensure a net gain by adopting the following betting strategy. He first sells you a conditional bet on *H* given *E*

B_1: [\$1 if *H*, \$0 otherwise],

and a bet on *E*

B_2: [\$(x-z) if *E*, \$0 otherwise]

at your fair prices. Later the truth of *E* becomes known. If *E* is false, the conditional bet is called off, and you end up losing $(x-z)y$. If *E* is true, he buys from you a third bet on *H*

B_3: [\$1 if *H*, \$0 otherwise]

at your fair price. But then, regardless of whether *H* is true or false, you will end up losing $(x-z)y$. If your updating strategy were to assign a new probability to *H* greater than $P(H/E)$, i.e., $Q(H)>P(H/E)$, the trick for the bookie would be to buy from you a bet on *H* given *E* at your lower fair price and later sell you back a bet on *H* at your higher fair price. In either scenario, your net loss would be zero if your new probability for *H* were equal to its old probability conditional on *E*. A rational person, it is concluded, must update his probability function in accordance with the BCR. Since the bookie needs to be aware of your updating strategy at the outset to devise a collection of bets that guarantees a sure loss, the Teller type argument is referred to as the *Dutch strategy* (DS) argument.

Although Teller's argument *prima facie* appears similar to the DB argument for the quotient rule, there are fundamental differences between them which are detrimental to the justificatory power of the DS argument. To begin with, in the argument for the quotient rule, assuming that you violate it, the bookie only has to know your current partial beliefs to make a Dutch book against you. The susceptibility to a Dutch book originates solely from the internal structure of your beliefs and, as a result, points to an undesirable feature of your belief system. In contrast, in devising a DS argument, the bookie needs to know not only your fair betting quotients (partial beliefs) but also the direction in which you intend to depart from the BCR. If you do not reveal in advance your updating strategy, he cannot make a Dutch strategy against you. The susceptibility to a Dutch strategy thus arises from a conjunction of your partial beliefs with a decision to pre-announce your updating strategy. The susceptibility to the sure loss does not automatically indicate a defect in your belief system. You can avoid it simply by refusing to pre-announce your updating strategy. And there is nothing irrational about it.

Second, the success of the DB argument for the quotient rule depends on the validity of the value additivity principle. If the principle is not granted, susceptibility to a Dutch book will have other explanations including the failure of value additivity, and cannot be taken as an indication of belief inconsistency. The postulate, as we saw, is not a logical principle. The only support for it is that whenever a number of bets are made *simultaneously,* it seems plausible to require that the value of a bet equivalent to the sum of the bets be the sum of the values of the individual bets. Like the DB argument, the DS argument also requires the assumption of value additivity to interpret susceptibility to a sure loss as an indication of belief inconsistency. However, the concern in the DS argument is with decisions made over time. In a dynamic decision making scenario, there is no reason that an individual should not take note of his or her earlier commitments, and for this reason value additivity cannot be taken for granted. As a result, vulnerability to a Dutch strategy cannot be taken as an indication of belief inconsistency. The susceptibility can in fact arise from the failure of value additivity.

Third, the bets involved in the DB argument are made simultaneously, the underlying beliefs all belong to a single point in time, and the coherence requirement used is known to be a rational ideal. The possibility of devising a DS argument, in contrast, hinges on the bookie having the opportunity to sell to or buy from you bets that are fair in your eyes at different times. This means the beliefs underpinning a Dutch strategy belong to different moments of time. So, even if the validity of the value additivity assumption is not challenged, the most that the possibility of a Dutch strategy can reveal is temporal inconsistency. But temporal consistency is not a rationality requirement. If consistency over time were a rationality requirement, the very idea of rational belief updating would be self-contradictory. In consequence, the DS argument has no implication for how to shift from one belief system to another in the light of what is observed (Christensen, 1991:264).

These criticisms show why there can be no argument similar to the DB argument for the BCR. However, they do not establish that there can be no justification whatsoever for the rule. The Bayesian literature in fact offers several alternative attempts to justify the rule as well as a generalisation of it by Richard Jeffrey (1968).[11] An analysis of these endeavours is beyond the scope of this chapter. Nevertheless, some general considerations indicate why they are also bound to fail. Note that the rule applies only when the probability of the conditioning event E shifts to unity.[12] In that case, the law of total probability implies that $Q(H) = Q(H \mid E)$,[13] which means the rule holds if and only if

$$Q(H / E) = P(H / E) .$$

This equality, termed the *invariance* condition, implies that in order for the rule to hold the new information must have no effect on the conditional probabilities in the domain of one's probability function. That is, having learnt E, the old and new conditional probabilities must agree with each other (Diaconis, et al., 1985:36). Any attempt at establishing the BCR as a general updating rule requires showing that the invariance condition must hold under any circumstances. However, there are certain cases where new information not only shifts the probability of the conditioning event but also justifiably demands reassessing some of the conditional probabilities in the domain of one's probability function. Howson (1997) provides a case in which by learning E one is logically forced to change

[11] Williams (1980) derives the BCR from the minimum information principle.

[12] This can be seen by applying the rule to E itself.

[13] $Q(H) = Q(E)Q(H / E) + Q(\neg E)Q(H / \neg E)$.

some of the conditional probabilities in the domain of one's probability function. Another case, closer to statistical practice, occurs when new observations cast doubts on the adequacy of the set of models considered, calling for construction of new models. When a new model is added or the models in the present set are modified, one should inevitably revise the probability of each model given the conditioning event (proposition) *E*. Since such legitimate belief shifts, arising from introduction of new models, cannot be ruled out *a priori*, there can be no prospect for establishing the invariance condition as a general rationality requirement. And, therefore, there can never be an argument establishing the BCR as a general rationality constraint.

Justificatory issues aside, the BCR is subject to sever limitations that undermine its role as a general learning model. The rule only applies to situations where the new information shifts the probability of the conditioning event (proposition) to unity. In reality, new information is usually vague, imprecise, and fraught with errors, and rarely shifts the probability of an event to unity (Jeffrey, 1968:171). In most real cases, the rule does not then apply anyway. The rule also requires both $p(E)$ and $p(H \& E)$ to be specified prior to learning of *E* and hence does not apply to unanticipated information (Diaconis et al., 1985). Finally, the rule does not apply to situations where a zero probability event occurs. All in all, the circumstances in which the rule applies are extremely limited.[14]

With these remarks, we end our brief study of some of the issues regarding the Bayesian theory, which have a direct bearing on the possibility of a theory of statistical learning. The main issue is whether the considerations behind the justification of the probability axioms impose any constraint on how to shift from a coherent system of beliefs to a new coherent system of beliefs in the light of new information. It has been seen that the answer is in the negative. The only claim of the Bayesian theory left standing is that, for the sake of consistency, one's likelihood judgements at each moment of time ought to accord with the laws of probability. This is a substantive requirement but does not prescribe how to shift from a coherent system of beliefs to a new coherent belief system. The

[14] A lucid discussion of these issues is found in Diaconis et al. (1985).

Bayesian theory, in itself, is not a theory of learning from experience. Contrary to some economists, the subjective expected utility maximisation assumption does not bring with it a rational theory of learning from experience.

3 The Orthodox View of Bayesian Inference

Coherent analysis has a place in a theory of statistical inference but there is much more to a theory of statistical inference than coherent analysis. As a step towards explaining the key issues that a theory of parametric inference must address and to define some necessary notions, we first give a brief account of the orthodox theory of Bayesian inference, which views inference from data *solely* in terms of prior to posterior analysis. Suppose we want to model the relation of random variable Y with X. According to the orthodox account, the modeller somehow knows the set of models W that can be true of the relation of Y with X:

$W = \{$All possible models that could possibly be true of the observables X and $Y\}$.

The assumption that W is known reduces the problem of inference from data to that of inferring the member of W that is most likely given the data. The Bayesian approach requires the modeller to express his uncertainty about the models in terms of a probability distribution that captures the confidence he has in each model prior to seeing the data. Let $D = \{x_t, y_t\}_{t=1}^{N}$ denote the data on X and Y, and let W contain only two models:

$$M_1 \quad : Y \sim N(\beta_1 X, \delta_1^2), \qquad \beta_1, \delta_1^2 \in \theta_1$$

$$M_2 : \quad Y \sim N(\alpha_1 + \beta_2 X, \delta_2{}^2), \qquad \alpha_1, \beta_2, \delta_2^2 \in \theta_2$$

Inferring the model, which is most likely given the data, requires estimating the parameters in each model. The hallmark of the Bayesian approach is to regard the parameters as random quantities, requiring the modeller to express his uncertainty towards them in the form of a (joint) probability distribution. Thus, a Bayesian model consists of at least two components, a data model $f(./\theta)$ and a (joint) prior density $\pi(\theta)$:

$$
\begin{array}{llll}
M_1: & Y \sim N(\beta_1 X, \delta_1^2) & \pi(\theta_1) & \beta_1, \delta_1^2 \in \theta_1 \\
M_2: & Y \sim N(\alpha_1 + \beta_2 X, \delta_2{}^2) & \pi(\theta_2) & \alpha_1, \beta_2, \delta_2^2 \in \theta_2
\end{array}
$$

$\pi(\theta_i)$ is the prior probability distribution for the parameters in M_i, representing the analyst's belief regarding the parameters prior to seeing the data.[15] The parameters in the prior density $\pi(\theta_i)$ are called hyperparameters as opposed to those in the data model $f(./\theta_i)$. Bayes' theorem combines the information in the prior density with the information in the data to derive the distribution of the parameters θ_i of each model, namely

$$
\begin{aligned}
& p(\theta_i / D, M_i) \\
& = p(D/\theta_i, M_i)\pi(\theta_i / M_i) \Big/ \int p(D/\theta_i, M_i)\pi(\theta_i / M_i) d\theta_i .
\end{aligned} \tag{3.1}
$$

In (3.1), $p(\theta_i / D, M_i)$ stands for the posterior distribution of θ_i and $p(D/\theta_i, M_i)$ for the likelihood function under model M_i. Assuming M_i is true, the posterior distribution $p(\theta_i / D, M_i)$ expresses all the information required for making inference about θ_i. A point estimate of θ_i, for instance, is obtained by computing the posterior mean

$$
\bar{\theta}_i = E(\theta_i / D, M_i) = \int \theta_i p(\theta_i / D, M_i) d\theta_i . \tag{3.2}
$$

Prediction is also obtained using posterior distribution $p(\theta_i / D, M_i)$. Suppose y_{t+1} is a future observation, independently drawn from the same distribution that has generated the data. The predictive distribution of y_{t+1} is given by

[15] Hierarchical models have further distributional assumptions relating to the distribution of hyperparameters.

$$p(y_{t+1}/x,D,M_i)=\int p(y_{t+1}/\theta_i,M_i)p(\theta_i/D,M_i)d\theta_i \tag{3.3}$$

This distribution, termed the *posterior predictive* distribution, summarises the information concerning the likely value of a new observation given the information in the data model, the prior and the data. If the posterior distribution $p(\theta_i/D,M_i)$ in (3.3) is replaced with the prior density $p(\theta_i/M_i)$, one obtains the *prior predictive* distribution

$$p(y_{t+1}/x,M_i)=\int p(y_{t+1}/\theta_i,M_i)p(\theta_i/M_i)d\theta_i\,, \tag{3.4}$$

which summarises one's information about an observation y_{t+1} before having seen any data.

As in parameter estimation, the orthodox theory treats the problem of model selection within the framework of prior to posterior analysis. It uses Bayes' theorem to derive the probability of each model given the data:

$$p(M_i/D)=\frac{p(D/M_i)p(M_i)}{p(D/M_1)p(M_1)+p(D/M_2)p(M_2)}\,, \tag{3.5}$$

where $p(D/M_i)$ is the marginal probability distribution of the data under model M_i, obtained by integrating over the model parameter space

$$p(D/M_i)=\int p(D/\theta_i,M_i)p(\theta_i/M_i)d\theta_i \tag{3.6}$$

with $p(D/\theta_i,M_i)$ being the likelihood of θ_i under model M_i. The theory then suggests choosing the model that scores the highest posterior probability. Also, the degree to which the data confirm M_1 over M_2 is measured by the posterior odds for M_1 against M_2, i.e., the ratio of their posterior probabilities. By equation (3.5), this is:

$$\frac{p(\mathrm{M}_1 / D)}{p(\mathrm{M}_2 / D)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(D / M_1)}{p(D / M_2)}. \tag{3.7}$$

The first ratio on the right-hand side of (3.7) is the prior odds ratio and the second is the Bayes factor. The numerator and the denominator of the Bayes factor are respectively the marginal likelihood of models M_1 and M_2. If the posterior odds ratio is above one, the data is said to support M_1 over M_2 and vice versa. If the posterior odds ratio equals unity, the data is said to give equal support to both models.[16] When the models are *a priori* considered equally likely, the posterior odds ratio is reduced to the Bayes factor.

To sum up, from the perspective of the orthodox theory, the agent knows the models possibly true of his choice situation or the economy in general. He uses data to estimate the models, and selects the model with the highest posterior probability. When new data come in, he re-estimates the models, computes their probabilities conditional on the data, and again searches for the model with the highest posterior probability.

4 Bayesian Statistical Inference: A Wider View

The orthodox Bayesian theory gives an incomplete description of the process of inference and of what a statistician does when confronting a real dataset. The starting point in this theory is the presumption that the candidate models are known. So, the central task of inference is defined as that of finding the model that is most probable given the data. This assumption is both theoretically and empirically indefensible. Candidate models are not known in advance, and the most important aspect of inference from data consists of model specification (formulation).

A number of activities precede model formulation. They include initial examination of the data, choosing appropriate transformations of the data,

[16] A review of Bayesian model selection is found in Kass and Raftery (1995).

producing descriptive statistics, finding outliers (Box and Snell, 1981; Gilchrist, 1984; Chatfield, 1995; and Leamer 1978). There are, therefore, at least two important phases of inference before the orthodox Bayesian theory, which interprets inference in terms of prior to posterior analysis, becomes relevant. The first is initial examination of data and the other is formulation of an initial model.[17]

The objective in initial model formulation is to specify a model that can serve as an informed basis to search for a model that can accurately account for the data. A Bayesian model is made of at least two components: a data model and a (joint) prior probability density for the model parameters. A data model, as stated in the last chapter, consists of a set of internally consistent hypotheses of independence, homogeneity, and distribution. The initial specification of a Bayesian model thus involves postulating appropriate assumptions of independence, homogeneity, and data distribution, *as well as* specifying a joint prior density for the data model parameters. Since *initial* specification of the basic assumptions concerns creating the objects (models) to which uncertainty applies, it is by definition a non-Bayesian matter. Any attempt at explaining initial model formulation necessitates stepping out of the framework of prior to posterior analysis (Hill, 1990:57). Once a model has been formulated, the next phase of inference is estimation (model fitting), where coherent analysis begins to become relevant.

Initial model formulation is a complex activity involving many decisions. There is no assurance a model generated in the early stage of research can account for the data and yield accurate predictions. An imperative question before making any use of the model is whether it is empirically adequate. In assessing the merits of candidate models $M_1, M_2, ..., M_K$, the orthodox theory requires specifying a prior distribution over the models and computing the posterior probability of each model using Bayes' theorem:

[17] Prior to these activities, a step in modelling consists of specifying the variables characterising the system. As in the previous chapter, for the time being, it is assumed that the variables are already known.

$$P(M_i / D) = \frac{P(D / M_i)P(M_i)}{\sum P(D / M_i)P(M_i)}$$

This approach only allows the comparison of relative probabilities (Lindley, 1982:81), which is not indicative of empirical adequacy. The high probability of a model can be the result of the choice of a particular prior for the parameters. As in Lindley's paradox, it is possible by adopting flat priors to arbitrarily increase the posterior probability of a model, and this can happen even if the sample size is very large (Gelfand et al. 1992:151; See Appendix *A* for a statement of the paradox). Moreover, the posterior probability of a model is always *conditional* on the set of candidate models considered (Box, 1980:427). When the set of candidate models contains only a single model, by Bayes' theorem the model automatically receives posterior probability one, and as the number of models in the set grows, the probability of the initial model can decrease and in fact approach zero (Box, 1983:73). The high probability of a model may thus be due to the analyst's failure to include among the candidates the true model or a close approximate thereof, rather than the adequacy of the model. Only if the set of candidate models is known to be wide enough to contain an adequate model, a connection can be made between the high posterior probability of a model and its empirical adequacy. Any attempt at ensuring this, though, calls for investigating the compatibility of each model with the data (Anscombe, 1961:34). This cannot be done using Bayes' theorem. Model assessment also necessitates a type of analysis different from prior to posterior analysis. It calls for a method that deals with the relation between a model and the data, not with apportioning of uncertainty across models (Barnard, 1962:42-3; Mallows, 1970:77).

The process of empirical adequacy assessment may reveal the failure of the initial model, calling for model re-specification. This involves varying the model assumptions one at a time, monitoring the effect of the variation, and continuing the process until an adequate model is obtained. Since in *re-specification analysis* the concern is with improving the adequacy of a single model, the analysis cannot be cast in terms of prior to posterior analysis. Re-specification analysis is also a non-Bayesian issue.

The process of initial model formulation, empirical model assessment, and re-specification analysis may produce several models fitting the data. For practical purposes, it may be needed to choose a model from among the candidates. It is here that coherent analysis can once again become relevant. Finally, the steps from initial examination of data to model selection are not a one-off process. New data may reveal the inadequacy of the final model. In real life, statistical inference is an iterative process. The statistician formulates a set of models, estimates them, assesses their adequacy, modifies them if necessary, chooses a model and derives the predictions required for decisions. As new data arrive, he reassesses the adequacy of the models, expands or modifies the set of candidate models, and derives new predictions, waiting for future data to shed light on the adequacy of the models. It is thus plausible to think of parametric statistical inference as a process with the following key phases:

(a) Data description
(b) Initial model formulation (or specification)
(c) Model fitting (or estimation)
(d) Model assessment (or criticism)
(e) Model re-specification
(f) Model selection
(g) Iteration

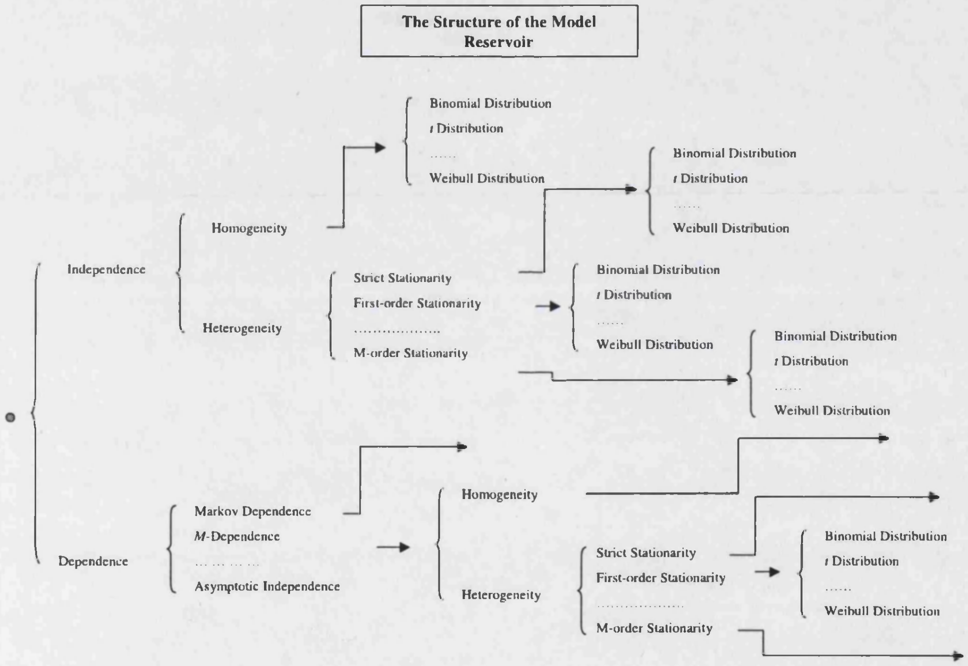
The Bayesian theory is only relevant to model estimation and model selection. It leaves out other central aspects of inference, namely, initial model formulation, empirical model assessment, and re-specification analysis. If the theory is intended to be a satisfactory account of statistical inference, it must be broadened to cover these critical aspects of inference. The rest of this chapter joins together various pieces from the literature to define a broader view of Bayesian inference, which goes some way towards explaining those aspects of inference not covered by the orthodox Bayesian theory.

5 Initial Bayesian Model Formulation

Is there a theory of (initial) model specification? Fisher is said to be the first who raised the issue of model specification in his seminal paper (1922) "On the mathematical foundations of theoretical statistics." In this paper, he divides the problems of statistics into three types: (i) problems of *specification*, (ii) problems of *estimation*, and (iii) problems of *distribution*. Fisher's discussion of specification problems in the paper is confined to a single paragraph, dominated by the first sentence: "As regards problems of specification, these are entirely a matter for the practical statistician ..." (1922: 314). The statement suggests "that in his view there can be no theory of modelling, no general modelling strategy, but that instead each problem must be considered entirely on its own merits" (Lehmann, 1990:160). Fisher's view of model specification has continued to dominate the statistics community, and has been endorsed by most statisticians, including Savage (1971), Mallow (1970), Dawid (1982), and Poirier (1988). However, a look at the modern statistical literature suggests that the view of model specification as an art with no general strategies is unduly pessimistic. Modern statistics provides a great deal of teachings that are highly relevant to establishing an exploratory theory of statistical model formulation. This chapter pieces together various elements of an exploratory theory of Bayesian modelling that takes us some way towards understanding how a statistician proceeds to build a model. The theory addresses three aspects of the model building process: 'initial model formulation', 'empirical model assessment' and 're-specification analysis.' The current section outlines a framework for initial model formulation by drawing on proposals found in D'Agostino (1986), Lehmann (1990), Rubin (1984), Spanos (1986,1999, and 2001).

5.1 Initial Data Model Specification

A theory of initial model formulation requires a clear definition of the problem and a method to solve it. To provide a satisfactory definition, we can divide the issue of Bayesian model formulation into specification of a data model and a prior distribution. We first consider the initial specification of a data model. As argued in the last chapter, when the concern is to establish an interpretable model of several variables, it is necessary to start with a parametric model, which raises the question of where the models come from. An interesting response to this question is found in Lehmann (1990): A line of research in mathematical statistics has been to define alternative notions of independence, homogeneity, and probability distribution families. The research has resulted in a rich variety of independence and homogeneity hypotheses, as well as a large list of univariate, bivariate, and multivariate probability distribution families. Consistent combination of these independence and homogeneity hypotheses with the distribution families produces a large collection of primitive data models, which can be used as building blocks to create numerous and in a sense countless mixture models. In this way, theoretical statistics provides a rich *reservoir* of models, to use Lehmann's apt term (1990:161). Figure 5.1 schematically shows the structure of the model reservoir.[18]



So, in response to the question where the models come from, Lehmann suggests that they come from the reservoir of statistics. In light of this proposal, the issue of initial data model specification can be defined as the problem of selecting a set of internally consistent hypotheses from the three categories of known independence, homogeneity, and distributional assumptions to form a model that can account for the data (Spanos, 1999:756). Having said this, to provide a

[18] See Spanos (1999) for definitions of the notions in the graph.

satisfactory theory of initial model formulation, it remains to explain how the initial selection of these hypotheses can take place.

5.1.1 The Theoretical Approach to Data Model Specification

Theoretical statistics provides the ingredients for two complementary methods for initial selection of the basic hypotheses, one drawing on subject-matter information and the other on data. The first procedure, also cited in Lehmann (1990), emerges from a class of theorems known as characterization theorems. Roughly speaking, a characterization theorem defines a set of sufficient conditions that if they were true of a variable (or a set of variables), the distribution of the variable (or variables) would belong to a certain probability distribution family (Galambos, 1982). A well-known characterization theorem is the Poisson process theorem that describes the conditions under which a univariate distribution has a Poisson distribution. In one form, the theorem goes as follows:

Consider variable Y_t and let t stand for time. For each $t > 0$, if

A$_1$: Y_t is an integer-valued random variable,

A$_2$: $Y_0 = 0$,

A$_3$: Y_t and $Y_{t+s} - Y_t$ are independently distributed, s >0,

A$_4$: Y_t and $Y_{t+s} - Y_t$ are identically distributed,

A$_5$: limit→0 $\dfrac{p(Y_t = 1)}{t} = \lambda$,

A$_6$: limit→0 $\dfrac{p(Y_t > 1)}{t} = 0$,

then, Y_t has a Poisson distribution (Feller, 1977). That is, for any positive integer n,

$$p(Y_t = n) = f(n) = e^{-\lambda_t} (\lambda_t)^n (n!)^{-1}$$

The theorem offers a way of deciding whether a variable Y_t has a Poisson distribution by checking if the information about the distribution of Y_t warrants assumptions A$_1$ to A$_6$. If so, Y_t has a Poisson distribution. In this way, the theorems provide a general procedure for using subject matter information to decide on the appropriateness of a distribution assumption, which leads to a narrowing of the set of appropriate data models. This approach underlies serious specification studies in econometrics, even though no reference is usually made to the theorems. To highlight the important role of the theorems in model formulation, we reconstruct a specification study from the econometric literature, and then state some of the limitations of the method in practice.

The study is adopted from Hausman et al. (1984) who examine the effect of research and development (R&D) on the technological innovation activity of a firm. The authors use patent applications as an indicator of inventive activity and seek to model its relationship with R&D. Let Y_t represent the number of patents applied for or received during period $[0,t)$ and X_t the expenditure on research and development during the period. To model the relation of Y_t with X_t, the authors proceed by listing a number of conceptual and simplifying assumptions that seem plausible about Y_t. Specifically, they propose that

A$_1$: Y_t is a discrete random variable taking a finite number of positive values;
A$_2$: The value of Y_t at time zero $t=0$ is zero (innovation takes time);
A$_3$: The numbers of patents received during *nonoverlapping* time intervals are *independent* of each other (independence assumption);
A$_4$: If Y_t is the number of patents received during $[0,t]$ and Z_t the number of patents received during $[t_1, t_{1+t}]$, Y_t and Z_t have the same distribution (homogeneity assumption);
A$_5$: The probability of receiving two or more patents in a sufficiently small interval is negligible; and
A$_6$: The probability of receiving *n* patents during $[t, t+s]$ is proportional to the length of $[t, t+s]$, barring extremely large intervals.

These hypotheses evidently match with the conditions of the Poisson process theorem. Thus, the authors conclude, as a first conjecture, that Y_t has a Poisson distribution, and model the dependence of Y_t on X_t using a Poisson regression model (Hausman et al., 1984:911),

$$
\begin{aligned}
&p(Y_t = n / x_t) = e^{-\lambda_t} (\lambda_t)^n (n!)^{-1} \\
&\ln(\lambda_t) = \alpha + \beta x_t
\end{aligned}
\tag{5.1}
$$

for any integer *n*. The authors then consider the effect of weakening the independence assumption, and investigate the possibility of adopting a more robust model such as the negative binomial regression model. A vast number of phenomena are similar to patent data, including the number of spells of sickness in a year, the number of records purchased per month, the number of cars owned, the number of jobs held during a year and so forth. Thus, in one stroke, the Poisson theorem provides a unified approach to creating an initial data model for a large number of economic phenomena. Many other specification studies in econometrics can easily be interpreted as an application of a characterisation theorem.[19]

This study illustrates how the theoretical approach to initial selection of a distribution assumption, which emerges from the characterisation theorems, enables one to draw on subject matter information to narrow down the class of data models that could possibly be true of a set of variables. The method is, nonetheless, subject to some limitations in practice. A trouble relates to the probabilistic conditions that enter the theorems. As explicit in the above example, the theorems assume that the data are identically and independently distributed. In the natural sciences, there may be reliable subject matter information to justify these assumptions *a priori*. But, in the social sciences, theories are imprecise, lack adequate empirical support, and the mechanisms generating the data undergo changes. And, therefore, the fate of these assumptions can rarely be decided on subject matter information alone. If there is any way of deciding on the appropriateness of the independence or homogeneity assumptions, which go into the theorems, it must be by looking at the data.

[19] Another interesting use of the theorems is found in Kiefer (1988), which concerns modelling the duration of unemployment.

Also, the information available about the distribution of a variable is usually imprecise and, as a result, consistent with more than one distribution family. In general, if the information is consistent with the assumptions defining a distribution family (say, exponential), it is also consistent with any distribution family that is robust with respect to it (say, Weibull). So, the approach does not usually lead to the choice of a single distribution hypothesis. These reservations aside, the theorems can effectively narrow down the class of appropriate models in the model reservoir. Even the information that the variable is continuous, finite, positive, or falls within the unit interval substantially reduces the space of appropriate data models within which an exploratory search must take place.

5.1.2 The Empirical Approach to Data Model Specification

The second method, which emerges from theoretical statistics, uses data for initial selection of the basic probabilistic assumptions. To explain the method, let us return to the definition of a data model as a set of internally consistent hypotheses drawn from the three categories of independence, homogeneity, and distribution. Each combination of these hypotheses, which forms a data model, implies a series of consequences that are true of the model *under all its possible parameterisations*. We term such consequences *ex ante* or *pre-estimation* implications, as they can be derived before estimating the model. Theoretical statistics has a rich literature on the *ex ante* consequences of alternative combinations of the basic assumptions defining the model reservoir. With this in mind, a plausible methodological principle is that a model worthy of further consideration must not have *ex ante* consequences that are grossly incompatible with the data. Granting this, the class of candidate data models, warranted by subject matter information, can be substantially narrowed down by investigating the *pre-estimation* consequences of the models. If the *ex ante* consequences of a model are compatible with the data, it is kept as a candidate model. Otherwise, it is excluded. Moreover, each *ex ante* implication of a data model can be traced to one of its assumptions. If an *ex ante* consequence of a model fails to appear in the data, then the failure can be traced to a particular assumption, and this information can be used to systematically search for a model capable of accounting for the data. The search for a first model need not then be entirely blind.

Essential to using data for initial model formulation is a judgment whether the *ex ante* consequences of the model are consistent with the data. In the frequentist setting, this judgment of consistency is made by computing p-values. An exploratory theory of Bayesian model formulation can also follow a similar route. But, since most *ex ante* consequences of a model are of a graphical nature or can be rephrased graphically, and since at this stage the objective is simply to make educated guesses about the nature of the statistical model that might be appropriate for the data, it suffices to work with an informal concept of incompatibility. It will be explained later how the frequentist idea of p-value can justifiably be assimilated within a broader view of Bayesian inference.

The following three subsections describe in some detail the process of data-driven initial model specification using a simple data set on the quarterly US unemployment rate over 25 years from 1948 to 1972, given in Fuller (1976), which is used later to illustrate Bayesian diagnostic learning. An objective behind the illustration is to emphasise the relevance of classical methods to an exploratory theory of Bayesian model formulation. Another objective is to bring to the fore the kind of heuristic principles that are necessary for using data in initial model specification. The exposition will also illustrate modes of inference that cannot be understood in terms of prior to posterior analysis but occupy a central place in a wider view of statistical inference.

5.1.2.1 The Independence Assumption

The choice of an independence and homogeneity assumption restricts the choice of a distribution family. This means the empirical search for an initial data model should begin with looking for an appropriate independence and homogeneity assumption. The starting point in this search is whether the data form a random sample or, in other words, whether the assumptions of C-independence (complete independence) and C-homogeneity (complete homogeneity) are appropriate. To focus on one assumption at a time, we first take *C*-homogeneity for granted. Let *Y* denote the unemployment rate, and *N* the sample size. Given *C*-homogeneity, the task of *ex ante* assessment of *C*-independence involves assessing the implications of the following model:

Unrestricted Data Model

A$_1$ Distribution:	Unrestricted
A$_1$ Independence:	$(Y_1, Y_2, ..., Y_N)$ is *C*-Independent
A$_1$ Homogeneity:	$(Y_1, Y_2, ..., Y_N)$ is *C*-Homogeneous

This model has several consequences that underlie a number of classical tests of independence, usually named distribution-free tests of randomness. Some of these tests are discussed in Bradley (1968) and Lehmann (1975). Here, for illustration, we follow Bradley (1968:271-8). Let us arrange the *N* observations in the order they were obtained. Suppose, for simplicity, none of the observations are identical so that they constitute *N* distinct numbers.[20] The *N* numbers can be arranged in *N*! distinguishable ways, creating a sample space *S* with *N*! elements. If the hypothesised model were true of *Y*, each element in *S* would *a priori* be as likely as the actual sequence. In other words, if one believed that the observations on *Y* were random, before seeing the data, one would have to consider each element in *S* as equally likely. An assumption to the contrary entails the failure of either *C*-independence or *C*-homogeneity (Bradley, 1968:277). The same conclusion is also true of any sample space formed from a sub-sequence of the *N* observations. This consequence leads to several procedures for *ex ante* assessment of *C*-independence. To explain one possible method, consider the *t*-plot of the unemployment data given below:

[20] For how to deal with identical observations see Bradley (1968:48-56).

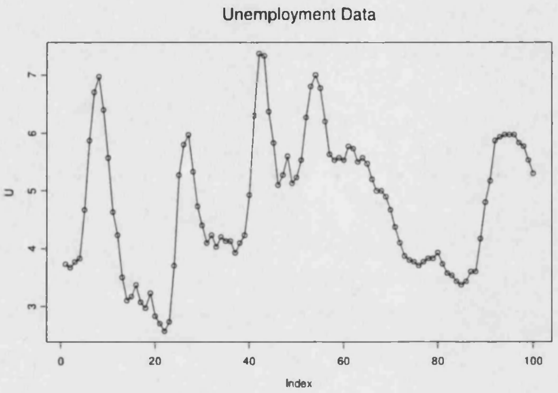


Figure 5.1

If an increase in the ordered sequence of observations is designated by "1" and a decrease by "0", the first quarter of the sequence of the unemployment data plotted in Figure 5.1 can be shown as:

0 1 1 1 1 1 1 0 0 0 0 0 0 1 1 0 1 1 1 0 0 1 0 1

An unbroken sequence of increasing observations (ones) or decreasing observations (zeros) is called a *run*. There are a total of 10 runs in the above subsequence. Let R be the total number of runs of any size in the entire sequence, $A_{R,N}$ the total number of arrangements of N that contains R runs, and $R_{(\geq m)}$ the number of runs of size m or greater. Given the equal probability of each element in S, Levene (1952) establishes that

$$P(R/N) = \frac{A_{R,N}}{N!},$$

$$E(R) = (2N-1)/3, \qquad Var(R) = (16N-29)/90$$

$$E(R_{(\geq m)}) = \frac{2+2(N-m)(m+1)}{(m+2)!}, \qquad m \leq (N-2)$$

and that R is asymptotically normally distributed. These consequences are in the form of expected and probability values, and as such say nothing about a particular sample. Nonetheless, it is plausible to assume that if the assumptions are appropriate, in an 'adequately' large sample the sample values come close to the theoretical values. In general, to bridge between the theoretical quantities and their sample analogues, the following heuristic principle, present in many areas of statistics, commands plausibility:

> **Heuristic Principle I**: If the hypothesized model is appropriate, in an adequately large sample, the theoretical (expected) values implied by the model for variables defined from the sample and the sample values of the variables are expected to be 'close' to each other.

In light of this, the appropriateness of C-independence can be assessed by comparing, say, the actual values of R and $R_{(\geq m)}$ with their expected values $E(R)$ and $E(R_{\geq m})$. A sharp difference casts doubt on the assumption. The sample in Figure 5.1 contains 100 observations, with $E(R)$ and $E(R_{\geq 3})$ being 66.33 and 6.48 respectively. The actual values of R and $R_{(\geq 3)}$ are 32 and 14 respectively, which are considerably different from the theoretical values. The significant difference strongly points to dependence in the data, suggesting the inappropriateness of C-independence.

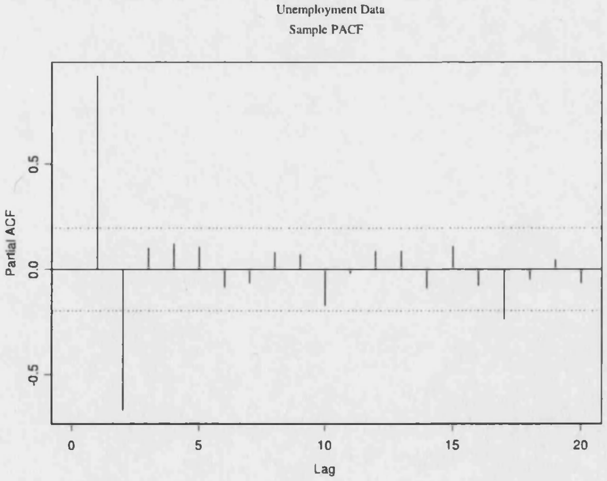


Figure 5.2

There is a wealth of techniques that can be used for pre-estimation assessment of alternative independence hypotheses. Notably, one may look at the sample partial autocorrelation function (SPACF) of various orders to tentatively select an independence assumption. Figure 5.2 (above) shows the plot of the SPACF of the unemployment data.

In general, if a p-order Markov independence assumption were true of Y, the sample partial autocorrelation function would be expected to "cut off" (i.e., be equal to zero) after p lag (Box and Jenkins, 1976). The plot suggests selecting a second order Markov independence condition. However, for illustration purposes, in what follows, we will work with a first order Markov condition, which leads to a simpler data model.

5.1.2.2 The Homogeneity Assumption

The next phase in initial model formulation is to search for a homogeneity assumption. The starting point in this search is an assessment of C-homogeneity, which is the simplest of the homogeneity assumptions. Classical statistics provides a host of distribution-free tests useful for investigating the pre-estimation implications of C-homogeneity. A class of such procedures is developed in Cox et al. (1955). For illustration, we look at these authors' test of trend in location, described in Bradley (1968:175). Suppose the sample consists of N different observations, with N being an even number. If N is an odd number, the middle observation dividing the sequence into two parts is removed. Arrange the observations as an ordered sequence $Y_1, Y_2, ...Y_i, ..., Y_n, Y_{n+1}, ..., Y_{n+i}, ..., Y_{2n}$ with the subscripts indicating the order in which they were obtained. Now, for every $i \leq n$ form the difference-score $Z_i = (Y_i - Y_{n+i})$, and let S be the number of positive difference-scores. Considering the signs of Z_i, the difference-scores can be viewed as the outcomes of n Bernoulli trials. If the unrestricted model is true, Z_i is as likely to be positive as to be negative, i.e., $p(Y_i < Y_{n+i}) = p(Y_i > Y_{n+i}) = .5$. In that case, S can be regarded as the number of successes in n Bernoulli trials, with probability $p = 0.5$ of success on each trial. This results in the binomial data model:

Binomial Data Model

A$_1$ Distribution: Binomial, $S \sim Bin(n,\pi)$; $P(S=s)=\binom{n}{s}\pi^s(1-\pi)^{n-s}$

A$_2$ Independence: $(Z_1, Z_2, ..., Z_n)$ is C-Independent

A$_3$ Homogeneity: $(Z_1, Z_2, ..., Z_n)$ is C-Homogeneous

with first and second moments

$$E(S)=n/2, \qquad Var(S)=n/4.$$

Cox *et al.*'s test of trend in location is based on computing p-value of the observed value of S. As a less formal check, one may assess the appropriateness of C-homogeneity by comparing the expected values $E(S)$ and $Var(S)$ with their sample analogues. In an adequately large sample, a significant departure points to the failure of C-homogeneity. In particular, when S is considerably greater than $E(S)$, the data points to a negative trend in location, and when it is considerably less than $E(S)$, it points to a positive trend in location. Cox *et al.* (1955) also establish analogous procedures for testing trend in dispersion or cyclical trend.[21]

As for the unemployment data, the expected values $E(S)$ and $Var(S)$ are 25 and 12.5 respectively, which are close to the sample values of 24 and 11.06. Similar results are obtained when the data are examined for trend in dispersion or cyclical trend. Thus, the data cast no doubt on C-homogeneity. The choice of the first order Markov condition, however, necessitates replacing C-homogeneity with strict stationarity, which is an extension of C-homogeneity to an independently distributed vector of random variables (chapter 3). With this choice, we obtain the following data model:

Unrestricted Data Model

A$_1$ Distribution: Unrestricted

A$_2$ Independence: $(Y_1, Y_2, ..., Y_T)$ is first order Markov independent

A$_3$ Homogeneity: $(Y_1, Y_2, ..., Y_T)$ is strictly stationary

[21] Kendall (1955) and Mann (1945) provide similar distribution-free tests of randomness, which can be used for *ex ante* assessment of C-homogeneity. See Bradley (1968:287-8) for an exposition.

5.1.2.3 The Distribution Assumption

The outcome of a pre-estimation search among the independence and homogeneity hypotheses is a data model of the form stated above. Given such a model, the pre-estimation search for a distribution family involves inserting alternative distributions, suggested by subject matter information, into the model, and assessing the *ex ante* implications of the model relating to the distribution assumption. In the current case, since Y_t is continuous, the first order Markov assumption restricts the class of plausible distribution families for Y_t to bivariate continuous families. To illustrate, we consider the bivariate normal family. This gives rise to the following data model:

Bivariate Normal Data Model

A_1	Distribution:	$\mathbf{X} \sim N(\mu, \Sigma)$, bivariate normal, $\mathbf{X} = (Y_t, Y_{t-1})$
A_2	Independence:	$(Y_1, Y_2, ..., Y_T)$ is first order Markov independent
A_3	Homogeneity:	$(Y_1, Y_2, ..., Y_T)$ is strictly stationary

The *ex ante* consequences of a distribution family are mainly defined by the invariant features of the density curve, or the cumulative distribution function (cdf) of the family. These include symmetry and skewness. So, with a reasonably sized sample, the appropriateness of a distribution family can be assessed by comparing the density curve or the *cdf* of the family with their sample analogues. The justification for this practice arises from another typical exploratory principle, which can be stated as follows:

Heuristic Principle II: If the data come from a distribution family, when the sample is adequately large, an appropriate plot of the data should show, within sampling error, the invariant features of the density curve or *cdf* of the family such as symmetry, positive or negative skewness, kurtosis, and so forth.

This methodology works well for appraisal of univariate and bivariate distribution families. However, since graphical features are difficult to investigate in high dimensional data, it cannot directly be extended to multivariate families. Nevertheless, the multivariate families have other types of *ex ante* implications that pave the way for their assessment. We briefly refer to three categories of such implications:

A general feature of the exiting multivariate (bivariate) families is that if they are true of a set of variables, the marginal distributions of the variables also belong to the same distribution family. This means an initial assessment of a multivariate family can be achieved by checking the marginal distribution families of the variables. The converse of this result is not though true; if the univariate distributions of a set of variables belong to a distribution family, it does not follow that the joint distribution of the variables also belongs to the same family (Seber, 1984:141).

In the current case, if the bivariate normal family is true of $\mathbf{X} = (Y_t, Y_{t-1})$, the marginal distribution of Y_t is also normal. The density curve of a normal distribution is symmetric. This means that, with a large sample, a judgement about normality can be achieved by checking the symmetry of a histogram or stem and leaf plot of the data. A more informative graph for assessing symmetry is obtained by plotting the upper half of the ordered observations against the lower half. Let $Y_{(1)}, Y_{(2)}, \dots, Y_{(N)}$ represent the ordered observations. If the data arise from a symmetric distribution, a plot of $Y_{(N)}$ versus $Y_{(1)}$, $Y_{(N-1)}$ versus $Y_{(2)}$, and in general $Y_{(N+1-i)}$ versus $Y_{(i)}$ for $i \le N/2$ should create a straight line with a negative unit slope (D'Agostino, 1986:13). Figure 5.2 plots $Y_{(N+1-i)}$ versus $Y_{(i)}$ for the US unemployment data.

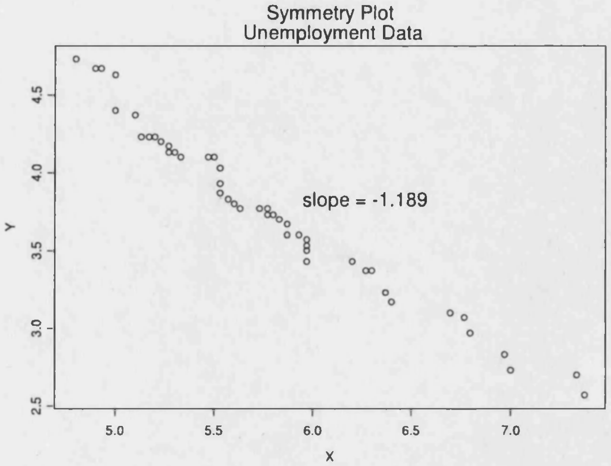


Figure 5.3 Y stands for $Y_{(N+1-i)}$ and X stands for $Y_{(i)}$.

The data points are mostly scattered around a straight line with a negative slope close to one, suggesting that they could have come from a symmetric distribution family such as the normal family. To narrow down the class of symmetric families to the normal family, further assessment of the *ex ante* consequences of the normal family is needed, which can be done, say, by checking the normal probability plot of the data.

A second type of *ex ante* consequences of a multivariate distribution is the imposition of restrictions on the form of the regression function of each of the variables on the rest of the variables. The distribution family determines whether the functions are linear, nonlinear, or how they look like. In the present case, if the bivariate normal distribution is true of *X*, the regression function of Y_t on Y_{t-1} is given by the linear function

$$E(Y_t / Y_{t-1} = y_{t-1}) = \alpha + \beta y_{t-1}. \qquad (5.3)$$

Alternatively, if the variables (Y_t, Y_{t-1}) have, for instance, a bivariate exponential distribution, the regression of Y_t on Y_{t-1} is given by the nonlinear function (Mardia, 1970)

$$E(Y_t / Y_{t-1} = y_{t-1}) = \frac{(1+\theta+\theta y_{t-1})}{(1+\theta y_{t-1})^2}. \qquad (5.4)$$

One can assess the linearity of the regression of Y_t on Y_{t-1} by using a nonparametric regressor to obtain a curve of the dependence of Y_t on Y_{t-1} and checking if it can be approximated by a linear function. Figure 5.4 represents the kernel regression curve of Y_t on Y_{t-1}.

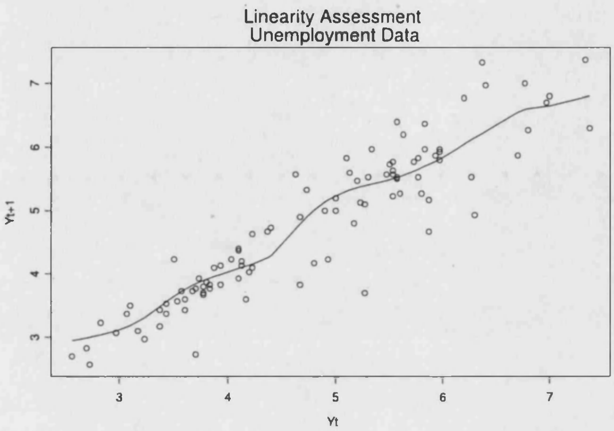


Figure 5.4. Kernel regression of Y_{t+1} on Y_t. The optimal level of smoothing was selected using the leave-one-out cross validation technique.

The curve comes close to a linear function, further confirming the consistency of the data with the normal family. In addition to nonparametric tools, a Bayesian statistician may also use the numerous classical means developed for checking linearity and curvature (Cox and Small, 1978; Abrevaya and Jiang, 2003).

Finally, a third type of *ex ante* consequences of a multivariate distribution family consists of the implications for new variables defined from the variables under study. To give an example, let $\mathbf{X}_i$ be the *i*th point in a sample of data on *X*, $\overline{\mathbf{X}}$ the vector of sample means, and *S* the sample covariance matrix. Further, define a new random variable:

$$d_i^2 = (\mathbf{X}_i - \overline{\mathbf{X}})^T \mathbf{S}^{-1} (\mathbf{X}_i - \overline{\mathbf{X}}) .$$

It has been shown that when *X* has a multivariate normal distribution, for large samples, d_i^2-values are approximately distributed as $\chi^2(p)$, where *p* stands for the dimension of *X*. In that case, a probability plot of the $\chi^2(p)$ percentiles against the ordered d_i^2-values will generate a straight line from the origin (Gnanadesikan, 1977:172-4), allowing direct assessment of joint normality.

Figure 5.5 plots the $\chi^2(2)$ percentiles against the ordered d_i^2-values for the unemployment data.

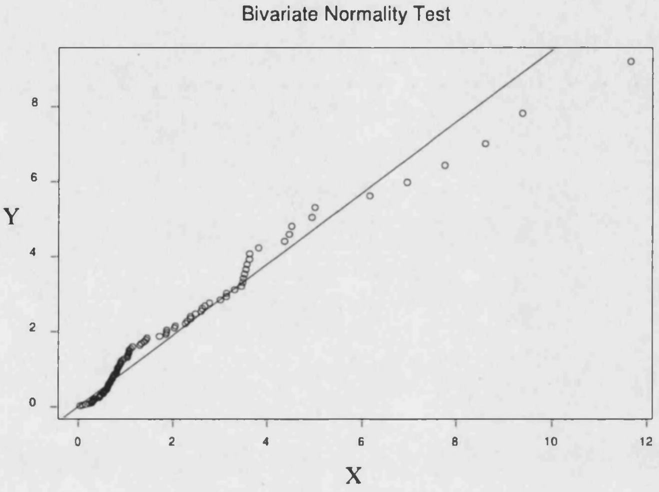


Figure 5.5 X stands for d_i^2-values and Y for $\chi^2(2)$ percentiles.

Though the data points are fairly closely scattered around a straight line, the fit is not perfect. The departure could be because the data have come from another symmetric distribution family, the first Markov condition is inappropriate, or there is noise in the data. All these possibilities can be investigated. Since the concern here is illustrative, we will not further the analysis and take the bivariate normal model as our initial model. In general, if the *ex ante* implications of a distribution family depart from the data, a similar approach can be pursued to assess the appropriateness of alternative distributions.

The unemployment data is very simple, and so is the above analysis. Nevertheless, the analysis gives a reasonable description of how a serious statistician formulates an initial model. Similar methods also guide dealing with complex datasets. An issue in dealing with complex datasets, for instance, is to decide whether to fit a mixture model and, if so, how complex the mixture model should be. Significant insight on this matter can be achieved by using probability plots. When the data come from different members of a distribution family, an appropriate probability plot generates several straight lines, each line representing a specific distribution (D'Agostino 1986:42-6). The complexity of the model can then be based on the number of inferred distributions. Initial data model specification is no longer without principles and procedures.

5.2 Prior Specification

A Bayesian model also requires a (joint) probability distribution for the model parameters. Although the Bayesian literature offers very little on data model specification, there is a substantive body of literature on prior modelling. O'Hagan (1994), to give an example, devotes a full long chapter on prior modelling but says nothing about specification of other model assumptions. This exclusive emphasis on prior modelling is unbalanced. The prior assumption is like any other assumption entering a model, if not the least critical one. If the data model is mis-specified, it is hard to make sense of a good prior. And, if it is correctly specified, when the sample is adequately large, the choice of a particular prior is not often critical. In any case, the central issue in prior modelling is whether there is a method to find a prior density for the parameters of the data model that enables it to best account for the data. Our response to this question will come in a later section. Here, to pave the way, we briefly look at various conventional approaches to prior modelling, explain the merits and shortcomings of each approach, and show why the focus of attention in these approaches are mistaken.[22]

5.2.1 The Summary-Based Method

The aim of prior modelling is traditionally defined as specifying a joint density function that best represents the modeller's opinion about the parameters before seeing the data. A prior density that represents substantive information is called an *informative* prior. The literature provides two methodologies for quantifying a person's qualitative information in terms of a density function. One is the *summary-based* method. The idea behind this method is that a distribution can be characterised in terms of a number of summaries. A univariate distribution, for instance, can be summarised using location measures (mean, median, and mode),

[22] Kadane et al. (1998) review the recent literature on prior elicitation.

dispersion measures (variance, standard deviation, and range), skewness, and fractiles. The summary-based method reverses this summarisation process. It requires expressing certain summaries about the distribution of the parameters and searching for a probability distribution that best fits the summaries (O'Hagan, 1994:143). This strategy underlies several apparently differing prior modelling techniques, whose only difference consists of the type of summaries they require and the way in which the summaries are used to select a density function.

As a simple illustration, following Berger (1980:66), consider the case of a univariate parameter θ, say, the mean of a normal distribution with a known variance. Suppose it is thought the median of the distribution $\pi(\theta)$ is close to zero and the first quartile (1/4 fractile) and the second quartile (3/4 fractile) of the distribution are respectively –1 and 1.[23] These summaries suggest that $\pi(\theta)$ is symmetric around its median. Therefore, it may be concluded that $\pi(\theta)$ belongs to the family of normal distributions, which are symmetric about their median. Since the mean and median of a normal distribution is the same, it follows that $\pi(\theta)$ is a normal distribution with mean zero i.e:, $N(0,\delta^2)$. At this point, the table of normal distribution can be used to conclude from the information on the quartiles that the variance $\delta^2 = 2.16$.[24]

The summary-based method is fraught with difficulties. One problem is that it requires thinking directly about parameters, which is difficult. To appreciate this point, recall the parameter in the simple exponential regression model mentioned earlier. The parameter enters the model in various ways, making it difficult to think directly about its role and distribution. The difficulty is compounded as more complex nonlinear models are considered (Kadane, 1980:90). Second, the distribution summaries obtainable in practice are usually consistent with more than one distribution family. The above summaries about θ are also consistent with the Cauchy distribution $C(0,1)$.[25] Telling these two distributions apart

[23] An "*a*-fractile of a continuous distribution is a point z(*a*) such that a random variable with this distribution has probability *a* of being less than or equal to z(*a*)" (Berger, 1985:79).

[24] When *Z* is a standard normal variable $p(Z < -1/\sqrt{2.16}) = 1/4$.

[25] The median is zero, and it can be checked that $\int_{-\infty}^{-1} 1/\pi([1+\theta^2])d\theta = 1/4$.

requires accurate summaries that cannot be easily obtained. Third, according to a dominant reading of the Bayesian position rooted in de Finetti's representation theorem, the parameters have no independent role but to simplify the relations among the observables (Lindley, 1982:77; Poirier, 1988:131).[26] If so, there is no fact of the matter about parameters to guide formulating a prior density other than their instrumental role in generating an empirically adequate model. Finally, there is no guarantee that the priors resulting from a person's distribution summaries lead to an empirically adequate model. It may be that the data model is correctly specified but, because of the choice of an inappropriate prior, the overall model is inadequate.

5.2.2 The Hypothetical Prediction-Based Method

The difficulties in thinking directly about parameters have given rise to an alternative approach to prior modelling that only demands distribution summaries about observables. Suppose the interest is to model the distribution of variable *X*, with data density $f(x/\theta)$. Let $\pi(\theta)$ stand for the (joint) prior density function of the parameters θ. Further, let *Y* denote some statistic defined from (hypothetical) observations $\{x_1,...,x_N\}$. The distribution of *Y* before seeing the data is given by the prior predictive distribution

$$m(y) = \int_{\Theta} f(y/\theta)\pi(\theta)d\theta \qquad\qquad \theta \in \Theta, \qquad\qquad (5.16)$$

which does not depend on the parameters θ, since they are integrated out. Equation (5.16) contains one known term, which is the data density of the statistic $f(y/\theta)$, and two unknown terms, which are the predictive distribution of the statistic $m(y)$ and the prior distribution $\pi(\theta)$. Now, suppose it was possible to estimate the predictive distribution $m(y)$ for some values of *Y*, or to state some

[26] De Finetti's representation theorem implies that coherent like-minded individuals who share symmetries (like exchangeability) in their beliefs are led to common likelihoods (data models). These data models are simplified in terms of mental constructs (unobservables) called *parameters* (Poirier, 1988:131). A readable introduction to the theorem is given in Heath et al. (1976). Also see Bernardo et al. (1994:172-80).

summaries of $m(y)$, such as mean, median, and fractiles, which were enough to infer the distribution. This would reduce the number of unknowns in equation (5.16) to one unknown, the prior density $\pi(\theta)$. The prior specification problem could then be solved by searching for a density function that renders the two sides of equation (5.16) equal. And, there would remain no need to think directly about the parameters to formulate a prior.

A problem with this strategy is that if the search for a prior is carried out in the class of all possible density functions, it will be difficult to solve the problem analytically. Specifically, it is not clear where to start the search and there can be many different densities equalising the two sides of (5.16). Any attempt at solving the inference problem requires restricting *a priori* the class of densities to which $\pi(\theta)$ belongs. A common restriction is to assume that $\pi(\theta)$ is a member of the distribution family that is conjugate with respect to the data density $f(y/\theta)$.[27] This assumption reduces the search for a prior into the search for a set of hyperparameters of the conjugate family that renders the two sides of (5.16) equal (Winkler, 1980:99). So, an alternative approach to prior modelling is to state certain summaries of the prior predictive distribution of the observables to infer the distribution. The approach next involves *a priori* restricting the distribution family to which the priors belong to a class smaller than the class of all possible densities. One then uses the predictive assessments to infer a set of values for the hyperparameters that equalises the two sides of (5.16).

A simple example, adapted from Winkler (1980:99), illustrates the method. Suppose the data have arisen form a Bernoulli process so that each observation can be considered as either a success $(x=1)$ or a failure $(x=0)$. Let Y stand for the number of successes in N trials. And, suppose the observations are random. We can describe the process generating Y using a binomial data model, with a parameter θ representing the probability of success on any given trial. The

[27] Let F denote the class of data density functions $f(x/\theta)$, defined by θ. A class of P of prior distributions is said to be a conjugate family for F if $\pi(\theta/x)$ is also in the class P for all $f\in F$ and $\pi\in P$ (Berger, 1980:96).

conjugate family for a binomial parameter is the beta family, which gives rise to a Beta-Binomial Bayesian model:

The Beta-Binomial Model

A$_1$ Data Distribution: Binomial, $p(y/\theta)=\binom{N}{y}\theta^{y}(1-\theta)^{n-y}$, $Y=\sum_{i=1}^{N}X_i$

A$_2$ Independence: $(X_1,X_2,...,X_N)$ is C-independent

A$_3$ Homogeneity: $(X_1,X_2,...,X_N)$ is C-homogenous

A$_4$ Prior Distribution: Beta, $\pi(\theta)=B(\alpha,\beta)^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}$,

where B is the beta function, $0\leq\theta\leq1$, $\alpha>0$, and $\beta>0$. The prior predictive distribution of Y is given by the 'beta-binomial' distribution (N,α,β),

$$p(y)=\binom{N}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y)\Gamma(N+\beta-y)}{\Gamma(\alpha+\beta+N)} \tag{5.17}$$

for $y=0,1,...N$, where Γ is the gamma function. Let y_i be the number of successes in i trials. Thus, $y_2=1$ represents one success in two trials. Given this notation and the properties of the gamma function, equation (5.17) entails the following simple equalities:

$$\frac{p(y_1=1)}{1-p(y_1=1)}=\frac{\alpha}{\beta},\quad \frac{p(y_2=0)}{p(y_2=2)}=\frac{\beta(\beta+1)}{\alpha(\alpha+1)},\text{and}\quad \frac{p(y_2=1)}{p(y_2=2)}=\frac{2\alpha\beta}{\alpha(\alpha+1)}.$$

These equalities can be used to infer α and β from some estimates of $p(y_i)$. For instance, the estimates $p(y_1=1)=.5$, $p(y_2=0)=.25$ and $p(y_2=2)=.25$ imply that α and β are equal to one. Over the last two decades or so, the predictive method has been extended to some common models such as the normal linear regression model (Kadane, et al., 1980; Kadane et al., 1998). For these models, it is now known what prior predictive assessments to obtain and how to use them to infer a prior fitting the assessments.

The predictive approach to prior modelling conquers one serious problem with the summary-based method by relinquishing the need for directly thinking about parameters. But it has its own limitations. For one thing, the form of the prior predictive distribution is not known for most interesting models encountered. This limits the usefulness of the method in practice (Winkler, 1980:99). Moreover, as the complexity of the data model grows, a larger number of predictive estimates are needed for finding necessary priors, making the method impractical (Kadane, 1980:91). Above all, the method requires pre-specifying a distribution family to which the priors belong. This raises the possibility that none of the members of the family can yield an empirically adequate model. On this score, the predictive approach offers no improvement on the summary-based method. Like the older approach, this method is also concerned with eliciting beliefs about parameters rather than with building an empirically adequate model.

5.2.3 Default Priors

The analysis of these two major approaches to prior modelling demonstrates the difficulties in specification of an informative prior. In addition to these methods, a line of research in Bayesian statistics has been to establish formal rules for specifying priors that contain no information and let the data rapidly dominate the posterior distribution. Historically, the origin of these rules is traced to the theory of objective Bayesianism, according to which, given any information set, there is only one probability distribution in relation to the information set (Jeffreys, 1931:10). And, when there is no information about a parameter, there exists a unique prior density representing the state of initial knowledge (ignorance). Such priors go by the name of 'noninformative', 'default', 'reference' or 'invariant' priors.

The earliest formal rule for prior specification is the principle of insufficient reason that assigns equal probabilities to all possible outcomes when there is no information to the contrary. The rule is subject to a re-parameterisation (partitioning) paradox; applying it simultaneously to all the equivalent representations (coarsenings and refinings) of the parameter space yields inconsistent probability assignments (Kass, et al., 1996:1347). Consider a single parameter θ and a one to one transformation of it such as $\phi = \theta(1-\theta)^{-1}$. If ignorance is claimed about θ, the rule requires choosing a uniform distribution. The change of variable formula then entails the prior density for ϕ to be $\pi^*(\phi) = (1+\phi)^{-2}$, which is not uniform. But, if one is ignorant about θ, one is also ignorant about ϕ, and in either case, according to the principle, one should select a uniform density. Since there is no such a thing as the 'correct' representation of the parameter space, the principle fails to identify a unique representation of the initial state of knowledge (Leamer, 1978:61).

What is ideally required is a rule that chooses a prior that is parameterisation invariant. In the context of the above example, this means it should not matter whether the rule is first applied to θ to obtain $\pi(\theta)$ and $\pi^*(\phi)$ is derived by means of the change of variable formula or it is first applied to ϕ to obtain $\pi^*(\phi)$ and $\pi(\theta)$ is derived by means of the change of variable formula. In either case, the priors should assign equal probabilities to the corresponding regions under both parameterisations.[28] Recognising this minimal requirement, Harold Jeffreys (1946) pioneered an approach to non-informative prior modelling, which is nowadays referred to as the invariance approach.

The approach links the choice of a prior to the model chosen for the data. To be precise, it considers one-to-one differentiable transformations of the random variables or the model parameters that do not change the model, and accordingly defines certain invariance requirements. It next searches for a prior that satisfies the requirements (Seidenfeld, 1979:419). To explain the core of the approach, following (Dawid, 1983), denote a data model by the triple $M = (X, \Theta, P)$, where X is a variable, Θ the parameter space, and $P = \{f(x,\theta), \theta \in \Theta\}$ the distribution family to which $p(x)$ belongs. Let $Y = g(X)$ be a one-to-one differentiable transformation of X (e.g.,$Y = X + c$) and $\Phi = h(\Theta)$ the parameter space induced by the transformation of X (i.e., $\Phi = \Theta + c$). Although the change transforms

[28] The priors must be related according to $\pi(\theta)d\theta = \pi^*(\phi)d\phi$.

$M = (X, \Theta, P)$ into a new model $M^* = (Y, \Phi, P)$, the distribution families in both cases are still the same; if $p(x)$ belongs to distribution family P (say, the normal family), so does $p(y)$. This means if M is true of a situation, M^* is also true of that situation, and in this sense the models are equivalent. Moreover, in the state of ignorance it is as likely that $\theta \in A \subset R$ as $\phi \in A \subset R$. It is therefore required that the prior satisfies the invariance condition $\pi(\theta \in A) = \pi^*(\phi \in A)$, which is known as the *context invariance* condition.

Jeffreys (1961:181) proposes a general rule that fulfils the context invariance condition and a few others. The rule is to take the prior density to be proportional to the square root of the expected Fisher information measure. In the univariate case, it is given by

$$\pi(\theta) = [I(\theta)]^{1/2}, \tag{5.18}$$

where $I(\theta) = E[-\partial^2 \log f(x/\theta)/\partial\theta^2]$ is the expected Fisher information for the parameter θ, and the expectation is taken with respect to the probability distribution function for x, $f(x/\theta)$. In the multi-parameter case, $I(\theta)$ is replaced with the determinant of the expected Fisher information matrix. This prior is invariant with respect to one-to-one transformations of the random variables or parameters appearing in the data model. That is,

$$\pi^*(\phi) = [I(\theta)]^{1/2} \left| \frac{d\theta}{d\phi} \right|. \tag{5.19}$$

Computing the Jeffreys prior for ϕ directly produces the same prior as computing the prior for θ and subsequently using the change of variable formula to obtain $\pi^*(\phi)$.

There has been a great deal of controversy surrounding the use and status of invariant priors. Most of these arise from the fact that invariant priors are inevitably improper; that is, they do not integrate to one. As a consequence, the context invariance condition in the form stated above is not strictly valid (Dawid, 1983). The most that can be assumed is that if in the state of initial knowledge it is as likely that $\theta \in A \subset R$ as $\phi \in A \subset R$, then the priors $\pi(\theta)$ and $\pi^*(\phi)$ must be proportionally related to one another, i.e., $\pi(\theta) \doteq h(c)\pi^*(\phi)$. This weaker condition is, however, satisfied by many other priors than Jeffreys' prior, which makes invariant priors non-unique. Having noted this multiplicity, Jeffreys proposed to select a prior on the basis of an international agreement (1955: 277).[29]

This proposal overlooks the possibility that a prior, chosen on the basis of international agreement, may not give rise to an empirically adequate model. A more reasonable proposal for selecting from among invariant priors is to tie the acceptability of the priors to the overall adequacy of the model. There is no difference between the prior assumption and other assumptions entering a model, and just as the plausibility of other assumptions are to be judged by looking at the overall adequacy of the model, the appropriateness of a prior must also be judged in the light of the adequacy of the model. From this perspective, the insufficient reason principle, Jeffreys' rule and other possible formal rules for formulating priors constitute valuable modelling tools. Since formulating informative priors is difficult, it is sensible to pick out first a prior using these rules and assess if it gives rise to an adequate model. If the model is adequate, the posterior distribution can be used as a prior in future inferences. If the model turns out to be inadequate, it is necessary to search for alternative priors. Thus, we regard invariant priors as default priors. This is not, however, to deny that invariant priors must be used with care, especially because they can lead to improper posteriors (O'Hagan 1994:79).

5.3 Some Limitations

This section began by asking if there is a theory of initial model formulation. The dominant belief is that there is not. Against this background, the section noted that theoretical statistics provides a rich reservoir of models, characterises the conditions under which a model can be true, and offers valuable information on

[29] Seidenfeld (1979) offers an appraisal of the invariance approach.

the *ex ante* consequences of the models. The section next showed how these contributions could be used as building blocks for an exploratory theory of initial model formulation. According to the theory, initial model formulation begins by using qualitative assumptions about the distribution of the variables to narrow down the class of data models that can be appropriate. It then involves examining the *ex ante* consequences of the models to find a model that can account for the data. Essential for the theory is certain heuristic principles for linking theoretical concepts with their sample counterparts.

The possibility of a theory of model formulation hinges on the existence of a model reservoir, and the scope of the theory grows with advances in theoretical statistics. As the list of the independence and homogeneity assumptions grows, new distributions are characterised, and new *ex ante* consequences are derived, the scope of the theory expands. Even though there is a relatively large list of univariate and bivariate distribution families, to date only a few multivariate distribution families have emerged in statistics. Of the four volumes of the reference work by Johnson et al. (1994), only the last deals with multivariate distributions, and this is dominated by the multivariate normal distribution. In addition, all the known multivariate families are based on the restrictive assumption that the marginal and conditional distributions of the variables also belong to the same distribution family. This scarcity of multivariate families defines the boundary of parametric inference. It also constrains the scope of the specification approach outlined above, which starts with modelling the joint distribution of the observables, and using it to derive the marginal and conditional distributions, as well as the regression functions of interest. The scarcity also renders prior formulation further difficult, as none of the few multivariate families may actually fit one's prior information.

Due to the scarcity of multivariate distribution families, it has inevitably become common in practice to consider the values of the independent variables as constant, and concentrate on the univariate distribution of the dependant variable conditional on the fixed values of the independent variables. The above exploratory methods assist in deciding on the univariate distribution but are not of much help in specifying the regression function beyond indicating whether it is linear, convex, or concave. Precise specification of the algebraic form of the function becomes a matter for trial and error.

Apart from the limitation arising from the scarcity of multivariate families, initial model formulation requires subjective judgements as to whether the sample size is large enough to permit comparison of theoretical and sample values, whether the discrepancies between the theoretical and actual values are large enough to call for searching an alternative model, and whether an incompatibility between the model's *ex ante* consequences and the data is due to chance or inappropriateness of the model. Because of the necessity of such judgements, investigation of *ex ante* consequences should be used for finding a model capable of accounting for the data, not for rejecting a model as false.

A final word may be needed on the compatibility of the Bayesian theory with the exploratory methods outlined above. Strictly speaking, Bayesian theory is only applicable after having formulated a model or a set of models, and is silent about the steps preceding specification of a model. Since the theory and the exploratory methods operate at two different levels, there is no incompatibility between them. Savage's last papers also reveal a high regard for 'puttering about with the data' (Savage, 1977:5), which can be construed as learning by means other than Bayes' theorem (Draper et al., 1993:25).

6 Bayesian Empirical Model Assessment

The above analysis exposes the complexity of initial model formulation, the uncertain decisions involved in selecting basic hypotheses, the difficulties in prior formulation, and the inconclusiveness of data and subject matter information in locating a single model. There is every reason to expect that the initial model may fail to account for important features of the data, and yield poor predictions. An important aspect of data modelling is therefore to assess the empirical adequacy of the initial model or models. The concern in empirical model assessment is with assessing the relation between a single model and the data, which falls outside the scope of the orthodox Bayesian theory. This section reconstructs and defends a trend in the literature that seeks to broaden the Bayesian framework by enriching it with a Fisherian notion of empirical adequacy and a method for assessing adequacy. The trend began with proposals by Barnard (1962), Anscombe (1964), and Dempster (1971), and culminated in the works of Box (1980, 1983), Rubin (1984), and Gelman et al. (1996). Drawing on the works and ideas of these statisticians, this section defines the notion of empirical adequacy of a Bayesian model, and describes various ways to investigate a model's adequacy. The section then shows how the ideas lead to a general procedure for Bayesian specification searches.

6.1 A General Framework for Model Assessment

The key to a theory of Bayesian empirical adequacy is the notion of *ex post* consequences and a method for judging their conformity with the data. Let denote a Bayesian model by $M(Z,\Phi,\pi)$, with Z being the variable (or variables) under study, Φ the parameter space, and π the (joint) prior density. Further, let $D^o = \{z_1^o,...,z_N^o\}$ be the actual sample, which, in statistics, is perceived as a realization of a vector of random variables $\mathbf{Z} = \{Z_1,...,Z_N\}$. The set of all possible realizations of variables $Z_1,...,Z_N$ is called a sample space, denoted by S. The actual data $\{z_1^o,...,z_N^o\}$ is thus a point in the N-dimensional sample space S. Next, let $T_i(.)$ be a function that maps each point of S into the real line, and let $T = \{T_1(.),...,T_k(.)\}$ be the set of all such functions of interest. We refer to $T_i(.)$ as a diagnostic or checking function. Each $T_i(.)$ takes the points in S into a new sample space S_i, leading to a collection of sample spaces $S^* = \{S_i,...,S_k\}$, defined by the checking functions in T. Any fully specified model for Z implies a probability distribution for the points in S, and through $T_i(.)$ a distribution $p(S_i)$ over S_i. By the *ex post* consequences of a model, we therefore mean the set of probability distributions $C = \{p(S_i),...,p(S_k)\}$ that the model implies for the sample spaces in S^*.

In this setting, the issue of consistency of a model's *ex post* consequences with the data boils down to the consistency of the induced probability distribution $p(S_i)$ with the actual value $T_i(D^o)$, for every checking function $T_i(.)$ of interest. Now the core of the Fisherian theory of goodness of fit test is that the consistency in question has to do with the location of $T_i(D^o)$ in the distribution $p(S_i)$, which is termed as the *reference* distribution, following Box (1980). If $T_i(D^o)$ falls in the central part of $p(S_i)$, the distribution is consistent with the data. If it falls in the (extreme) tail area of the distribution, it is inconsistent with the data, since in that case the actual value $T_i(D^o)$ receives a lower probability as compared to the most points in the sample space S_i (Anscombe, 1963). Having said this, a model $M(Z,\Phi,\pi)$ may be defined as empirically adequate if, for each relevant diagnostic function $T_i(.)$ in T, the reference distribution $p(S_i)$ confers a 'high' probability on the realized value $T_i(D^o)$ as compared to other possible points $T_i(D)$ in the sample S_i.

In the Bayesian setting, the distribution of the observables under a model is given by the predictive distribution or, in other words, the marginal distribution of the data. In view of this, Guttman (1967), Dempster (1971), Box (1980) and Rubin (1984) have suggested taking the predictive distribution as the basis from which to derive the distributions of the statistics included in T. From this perspective, a Bayesian model is adequate if the predictive distribution $p(S_i)$ for each diagnostic function $T_i(.)$ of interest confers a high probability on the realized value $T_i(D^o)$ as compared with the other points $T_i(D)$ in S_i. The empirical adequacy of a Bayesian model thus goes hand in hand with the predictive accuracy of the model; they are in fact the same thing.

In light of this account, the adequacy of a Bayesian model can be assessed by (i) selecting appropriate diagnostic functions $T_i(.)$ to capture relevant features of the data; (ii) deriving the predictive (reference) distributions of the functions $p(T_i(.))$ under the model; (iii) computing the realised values of the functions, i.e., $T_i(D^o)$, and (iv) determining the location of $T_i(D^o)$ in the distribution $p(T_i(.))$. This may be done in more than one way. It may be done by computing the probability $\Pr\{p(T_i(D)) \geq p(T_i(D^o))\}$ or $p(T_i(D) \geq T_i(D^o))$. If these probabilities are not extreme, the model is consistent with the data in respect of the statistic in question (Anscombe, 1963:84).

The justification of this approach to Bayesian model assessment lies, on the one hand, in the fact that for assessing the adequacy of a single model one necessarily needs to look at the consistency of the model's consequences with data. And on the other, it lies in the fact that statistical models have no deductive consequences; a statistical model is logically consistent with any observed data (Dawid, 2002). Therefore, either the idea of assessing the empirical adequacy of a single model is abandoned, in which case the process of model formulation remains a mystery, or it is admitted and one is naturally led to the Fisherian idea (Spanos, 2001). After all, the only way to decide on the consistency of a statistical model with data is by looking at the location of the data in the distribution of the observables under the model.

6.1.1 The Variety of Predictive Distributions

Two types of predictive distributions were defined earlier, prior and posterior predictive distributions. The former describes the distribution of the observables given the information in the data model and prior density but takes no account of the data. The latter distribution, in contrast, describes the distribution of future data given the information in the data model, data and the prior density. There are thus two types of predictive distributions from which to derive the *ex post* implications of a Bayesian model, each of which gives rise to a somewhat different approach to model assessment.

6.1.2 Prior Predictive Checks

Suppose our assumptions A regarding the process generating data D lead us to the density function $p(D/\theta,A)$ and prior $p(\theta/A)$. The joint distribution of D and θ is given by

$$p(D,\theta/A)=p(D/\theta,A)p(\theta/A),$$

and the prior predictive distribution of D by

$$p(D/A)=\int p(D/\theta,A)p(\theta/A)d\theta, \qquad (6.1)$$

which gives the distribution of the totality of all possible samples D that could occur if the assumptions A were true. The belief in the appropriateness of the assumptions A implies that the outcome of contemplated data acquisition would be calibrated with adequate approximation by a simulation involving appropriate random sampling from the distributions $p(D/\theta,A)$ and $p(\theta/A)$ (Box, 1983:59). This means if A were true, the actual data D would fall well within the support of the predictive distribution $p(D/A)$. In light of this, Box (1980) suggests assessing the adequacy of the model by investigating the location of actual data D^o in the prior predictive distribution $p(D/A)$ or by checking the location of some relevant diagnostic function $T(D^o)$ in $p(T(D)/A)$. The following two examples, adapted from Box (1983) and (1980), illustrate the prior predictive approach to adequacy assessment.

The first example concerns the modelling of the number of successes in a sequence of random Bernoulli trials $X_1,X_2,\ldots,X_N$, with X_i being either 0 (failure) or 1 (success). The distribution of the number of successes Y in a sequence of random Bernoulli trials is given by the Binomial distribution, with parameter θ representing the probability of success on each trial. Suppose a member of the beta distribution family with $E(\theta)=.0.2$, and $Var(\theta)=0.01$ represents our belief about θ. As seen earlier, with a beta prior, the prior predictive distribution of Y in N Bernoulli trials is given by the 'beta-binomial' distribution (n,α,β)

$$p(y/A)=\binom{N}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y)\Gamma(N+\beta-y)}{\Gamma(\alpha+\beta+N)}, \tag{6.2}$$

where A stands for the model assumptions.[30] Using the formula for the mean and variance of a beta distribution, it can be shown that the prior belief about the distribution of θ implies that $\alpha=3$ and $\beta=12$. The prior predictive approach to model assessment involves assessing the model adequacy by locating the probability of the observed data $p(y^o/A)$ in the distribution (6.2) by computing the probability

$$\Pr(p(y/A)\leq p(y^o/A)). \tag{6.3}$$

Consider two scenarios. In the first scenario, the experiment is carried out 10 times and 3 successes are observed. The prior predictive probability $p(3/A)$ is 0.16, which is not unusually small. In fact

$$\Pr(p(y/A)\leq p(3/A))=0.33.$$

The data provides no reason to doubt the model. In the second scenario, suppose there are 8 successes. The prior predictive probability $p(8/A)$ is 0.0018, and

$$\Pr(p(y/A)\leq p(8/A))=0.0021,$$

which is quite small. The data casts doubt on the model, calling for a revision of the assumptions A.

As a different illustration, consider modelling the distribution of a continuous random variable X for which we have the data set D^0 = (34, 32, 38, 35, 39). Suppose a normal distribution with variance $\sigma^2=1$ is thought to fit the data.

[30] Note that the Beta function can be stated in terms of the Gamma as $B[\alpha,\beta]=\Gamma[\alpha]\Gamma[\mathrm{B}]/\Gamma[\alpha+\beta]$.

Suppose a normal prior with mean $\theta_0 = 30$ and variance $\tau^2 = 3$ captures our prior belief about the location parameter θ of the data density. These assumptions lead to the following model:

Simple Bayesian Normal Model
A1 Distribution: Normal, $X \sim N(\theta,\sigma^2)$, $\sigma^2 = 1$
A2 Independence: $(X_1, X_2,..., X_n)$ is C-independent
A3 Homogeneity: $(X_1, X_2,..., X_n)$ is C-homogenous
A4 Prior Distribution: Normal, $\theta \sim N(\theta_0,\tau^2)$, $\theta_0 = 30$, $\tau^2 = 3$.

The posterior distribution of θ is given by

$$\theta \sim N(\varphi,\phi),\ \phi = (\tau^{-2} + n\delta^{-2})^{-1},\ \varphi = \phi(\theta_0\tau^{-2} + \delta^{-2}\sum x_i)$$
$$\theta \sim N(36,0.19).$$

Let $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$. The prior predictive distribution of *X* is given by (Berger, 1980:93-94):

$$p(x/A) = (2\pi)^{-n/2}\sigma^{-n}(\sigma^2\tau^2/\sigma^2 + n\tau^2)^{-1/2}\exp\{\frac{-1}{2}[\frac{s^2}{\sigma^2} + \frac{(\bar{x}-\theta_0)^2}{n^{-1}\sigma^2 + \tau^2}]\}. \quad (6.4)$$

A possible checking function for the model is the quantity in the bracket; namely,

$$T(X) = [\frac{s^2}{\sigma^2} + \frac{(\bar{x}-\theta_0)^2}{n^{-1}\sigma^2 + \tau^2}],$$

The empirical adequacy of the model may then be assessed by computing the prior predictive probability

$$\Pr(p(T(X)/A) \le p(T(x^o)/A)). \quad (6.5)$$

Given the model assumptions, the quantity $T(X)$ has a χ_n^2 distribution,[31] and since $T(X)$ is inversely related to $p(x/A)$, for computing (6.5) it is sufficient to calculate

$$p = P(\chi_n^2 \geq T(x^o))$$

However, the test statistic

$$T(x^o) = \frac{5\times 6.64}{1} + \frac{(35.6-30)^2}{3.2} = 43$$

and

$$p(\chi_5^2 > 43) < 0.001.$$

This low predictive probability reveals that the occurrence of the data under the model is quite unlikely, which calls for a revision of the model $N(36,0.19)$.The entire predictive approach to adequacy assessment rests upon the ability to derive desired predictive distributions and to locate the position of the actual values of the statistics of interest in the distributions. Analytic evaluation of the integrals in the predictive distributions is generally hopeless. One needs to turn to sampling based methods to derive the predictive distributions of interest (Gelfand et al., 1992:148).

A common criticism against the prior predictive approach relates to the similarity between prior predictive *p*-values and the classical goodness of fit tests. Since the same criticism has been levelled against other types of Bayesian predictive *p*-values, we postpone an assessment of this criticism until we thoroughly explain the role of predictive checks in the Bayesian setting. This aside, a limitation of the prior predictive approach is that it only applies to models with proper priors. When the prior is improper, the prior predictive distribution is also improper and, as a result, prior predictive *p*-values are not defined (Bayarri *et al.* 1999). More importantly, prior predictive checks are sensitive to the choice of priors. A choice of inappropriate priors can lead the analyst to wrongly question a well-specified data model. Thus, the approach is primarily suitable for exploring the effect of

[31] If $X_1, X_2,...X_n$ are NIID (standard Normal) $Y = \Sigma X_i^2 \sim \chi^2(n)$.

alternative priors within a model and should not be used to question the data model unless the appropriateness of the priors has somehow been ascertained (Hodges, 1988:264).[32]

6.1.3 Posterior Predictive Checks

The proposal to use posterior predictive distribution for adequacy assessment first appeared in Guttman (1967) and was later developed in Rubin (1984) and Gelman et al., (1996). To explain the posterior predictive approach, let D^o be the observed data on random variable *Y*, and *A* the assumptions forming a candidate model, with parameter vector θ. The posterior predictive distribution of *Y* under the model is given by

$$P(y^f / A, D^o) = \int P(y^f / A, \theta)\pi(\theta / D^o)d\theta, \qquad (6.6)$$

with y^f standing for a future observation. If the assumptions *A* were correctly specified, we could think of the actual data D^o as a random sample drawn from the distribution (6.6). This means if we could simulate random samples of size *N* (the size of D^o) from the distribution, we would expect the samples to be on average 'similar' in 'relevant ways' to the actual sample D^o (Rubin, 1984:116). So, information on the adequacy of the model may be achieved by simulating random samples of size *N* from the distribution (6.6) and assessing the similarity of the hypothetical samples with the realized sample D^o. To elaborate on the process, consider checking if a normal model fits data $D^o = \{x_1, ..., x_n)$. Suppose a pair of conjugate priors is thought to capture the beliefs about the location and scale parameters of the data distribution. Assuming that the data form a random sample, the task involves assessing the adequacy of the normal / chi-squared model (Lee, 1997):

[32] See Bayarri and Berger "Measures of Surprise in Bayesian Analysis" for a discussion of other objections to the prior predictive approach.

Bayesian Normal/Chi-squared Model

A₁ Distribution: Normal, $X \sim N(\theta,\phi)$

A₂ Independence: $(X_1, X_2,..., X_N)$ is C-independent

A₃ Homogeneity: $(X_1, X_2,..., X_N)$ is C-homogenous

A₄ Prior Distribution: Normal / Chi-squared distribution

The posterior distribution for ϕ is given by the inverse chi-squared distribution $\phi \sim S_1\chi_{v_1}^{-2}$ and for θ given ϕ by the normal distribution $\theta \mid \phi \sim N(\theta_1, \phi / N)$, where S_1, θ_1, and v_1 are defined in Appendix *B* The posterior predictive distribution of *X* is given by

$$P(x / A, D) = \iint N(x / A, \theta, \phi) N(\theta_1, \phi / n) S_1 \chi_{v_1}^{-2}(\phi) d\theta d\phi . \tag{6.7}$$

The key part of Rubin's posterior predictive approach is to compare the actual sample with the samples simulated from (6.7). To this end, a value ϕ^* is first drawn from the posterior distribution $S_1\chi_{v_1}^{-2}$, say by means of Markov chain Monto Carlo simulation, and then given ϕ^* a value θ^* is drawn from $N(\theta_1, \phi^* / N)$. Next, using the simulated values, a sample $D^{repi} = \{x_1,...,x_N\}$ is drawn from $X \sim N(\theta^*, \phi^*)$. These steps are repeated to obtain *k* (say, 10,000) random samples and the value of the diagnostic statistic $T_i(D^{repi})$ for each simulated sample is defined.[33]

A judgment of similarity is based on the number of cases in which the simulated value of the diagnostic function $T_i(D^{repi})$ differs from the realized value $T_i(D^o)$. More precisely, similarity is judged by computing the percentage of these *k* simulations for which the value of the function $T_i(D^{repi})$ exceeds (or is less than) the realized value $T_i(D^o)$. This is known as the *posterior predictive p-value* (Gelman et al. 1996):

[33] Gilks, et al., (1996) contins a collection of articles on Markov Chain Monte Carlo techniques.

$$\textit{Posterior Predictive p-value} = \alpha = \frac{1}{k}\sum_{i=1}^{k} I_{T(D_i^{rep},\theta_i) > T(D_i^{o},\theta_i)} \tag{6.8}$$

where I is the indicator function. If α for the diagnostic functions of interest are close to 0 or 1, the model is considered to be suspect. Otherwise, it is considered as empirically adequate. The posterior predictive approach is consistent with the main trust of Bayesian reasoning, which is conditioning on the whole data (Rubin, 1984:1166).

The posterior predictive approach gets around two difficulties of the prior predictive diagnostics. Since regardless of whether the priors are proper or not, posterior parameter distributions are usually proper, posterior predictive distributions are also usually proper. As a result, the use of posterior predictive diagnostics is not limited to models with proper priors. Second, in contrast to prior predictive checks, if the sample is adequately large, posterior predictive distributions are not sensitive to the choice of priors. Thus, posterior predictive diagnostics can be used for assessing data model assumptions. These successes, however, come at a price. As many critics have pointed out, posterior predictive checks use the data twice. They use the data once to derive the posterior predictive distribution of the observables under the model and once to assess the model. They are therefore prone to underestimating the inadequacy of the model (O'Hagan, 2001:7). Even so, it is true that if a model fails to generate data similar to the data used to obtain it, there is something amiss about it. So, posterior predictive checks of the type proposed by Rubin and others provide valuable exploratory tools for specification searches.

6.1.4 Cross-validated Posterior Predictive Checks

The prior predictive approach, which leaves the whole data out as a test set, and the posterior predictive approach, which takes the whole data as the training set, are two extremes of an spectrum in which the *ex post* consequences of a single Bayesian model can be defined. There are many alternatives in between, arising from various other ways in which the data can be divided into a training and test set. Of these middle-way procedures, re-sampling techniques, as stated in the last chapter, have received most attention. In the Bayesian setting, Gelfand et al. (1992) and Bernardo et al. (1994) suggest using cross validation for model selection but the method can equally be used for adequacy assessment. In its simplest form, the technique leaves out the *i*th observation y_i as a test set and fits the model to the remaining data set D^{-i} to derive the posterior predictive distribution of the omitted observation y_i

$$p(y_i / D^{-i}) = \int f(y_i / \theta)\pi(\theta / D^{-i})d\theta \qquad \text{for all } i = 1,\ldots,N \qquad (6.9)$$

The procedure is repeated to derive *N* posterior predictive distributions of the *N* observations in the sample. Adequacy is judged by drawing *k* (say, 10,000) hypothetical observations from the predictive distribution (6.9) for each observation to form *k* hypothetical samples, and using the samples to derive the posterior predictive distributions of the diagnostic statistics $T_i(.)$ of interest. The model is considered as adequate if the actual values of the statistics fall in the main part of their reference distributions.

The cross-validated distribution (6.9) can also be used to define other important types of *ex post* consequences of a Bayesian model. A number of these implications are listed in Gelfand et al. (1992), of which the following two are the simplest:

(i) Let $e_{1i} = y_i - \hat{y}_i$ measure the difference between the realized value y_i and its predicted value $\hat{y}_i$ $(i.e., E(Y_i / D^{-i}))$, and σ_i^2 be $Var(Y_i / D^{-i})$. Standardizing e_{1i} yields $d_{1i} = e_{1i} / \sigma_i$. If the errors e_{1i} are approximately normally distributed, d_{1i} approximately has a standard normal distribution.[34] In that case, 95% of the standardized errors d_{1i} must fall within the interval –2 to +2. If this is not the case, the model fails to fully capture systematic information in the data. Also, the

[34] A well-fitted model will produce residuals that are approximately independent random variables with zero mean, constant variance, and, possibly, a normal distribution (Gilchrist, 1984:138).

squared sum of the standardized errors $D_{2i} = \sum d_{1i}^2$ can be taken as an overall index of adequacy.

(ii) Let $e_{2i} = 1$ if $\hat{y}_i \leq y_i$, otherwise 0. The expectation $d_{2i} = E(e_{2i})$ is $P(Y_i \leq y_i / D^{-i})$. Viewing y_i as a random draw from the predictive distribution $p(Y_i / D^{-i})$ implies that d_{2i} is uniformly distributed over the unit interval, i.e., $d_{2i} \sim U(0,1)$. If the model is correctly specified, the average $A(e_{2i}) = \sum e_{2i} / N$ is expected to be close to 0.5. Extreme values for $A(e_{2i})$, i.e., values close to 0 or 1, point to inadequacy.

In addition, the predictive errors e_{1i} can be used for graphical residual analysis to investigate different aspects of the model. By plotting the errors e_{1i} versus the predicted values $\hat{y}_i$ it is possible to check the appropriateness of the variance homogeneity; by plotting the residuals versus time it is possible to assess the appropriateness of the independence assumption; or by plotting the residuals as a histogram it is possible to check if they are normally distributed (Gilchrist, 1984:138-44).

The cross-validation method overcomes double using of the data but is not free of limitation either. Holding part of the data out as a test (adequacy) set can destroy crucial features of the data such as dependence, which can lead to a wrong estimate of the model accuracy (Chapter 3). Hold-out methods, such as cross-validation techniques, are only suited for unstructured data. The general view transpiring from the remarks about the weaknesses and strengths of each approach to adequacy assessment is that none of the methods outperforms others in all respect. Their applicability depends on the kind of data under study and the aspect of the model being considered.

6.2 Bayesian Specification Searches

The notion of *ex post* consequences of a Bayesian model combined with the Fisherian approach to assessing the compatibility of the model consequences with the data leads to a powerful procedure for searching the space of data models suggested by the initial exploratory analysis. The procedure involves choosing a data model, adopting a (joint) prior distribution for the model parameters, and assessing the compatibility of its *ex post* consequences with the data. If the model is found empirically inadequate, one model assumption is varied at a time, the effect of the variation on the model adequacy is assessed, and the procedure is repeated until a model that successfully accounts for the data is found. In practice, when a data model assumption is varied, it is also often essential to modify other model assumptions to preserve consistency among the basic hypotheses. This learning procedure, which captures the way in which a serious (Bayesian) statistician builds a model, might be named Bayesian diagnostic model searching. This section illustrates the procedure by further analyzing the example discussed in Section V.

6.2.1 Exploring Prior Distributions

The traditional methods to informative prior modeling demand arbitrary choices regarding the distribution families to which the priors belong and distributional summaries and hypothetical predictions that are hard to obtain. This raises the possibility that the resulting priors may not enable the model to best account for the data, even though the basic probabilistic assumptions are correctly specified. As a result, when the candidate model is the outcome of a careful initial exploratory analysis, the first step in *ex post* assessment of the model should be to find a set of priors, which, given its other assumptions, enables it to best account for the data. It is only after this that it can be assessed whether the data model is able to account for the data. In view of the fact that the main objective of modelling is to specify a model capable of accounting for the data, it makes sense to link choice of priors to the adequacy of the model. As an alternative to the traditional methods, it is therefore appropriate to choose a prior by looking at the compatibility of the *ex post* consequences of the model with the data. Specifically, following Box (1980), Hill (1990), and Geweke (1999), it seems plausible to propose a two-stage method for prior specification. First, subject matter considerations are brought in to *tentatively* limit the class of candidate priors; considerations such as the parameter takes real, positive values, falls in the unit interval, and so forth. Second, the effect of candidate priors on the adequacy of the model is investigated, while holding the data model fixed. A set of priors that enables the model to best account for the data is selected.

To illustrate the process, let us return to the US unemployment data. Initial examination of the data suggested that a bivariate normal data model might be appropriate. This implies that Y_t follows a first order normal autoregression model (Spanos, 1986, Ch 22, Appendix):

Normal AR (1) Data Model

$$Y_t \mid y_{t-1} \sim N(\pi, \sigma^2),$$
$$\pi = \alpha + \beta y_{t-1}$$

which contains three parameters α, β and σ^2. To illustrate the search for priors, it is necessary to start with some tentative decision about the distribution families to which the priors might belong. Since the purpose here is illustrative, it seems appropriate to consider conjugate priors. Suppose we start with the prior densities given below:

Bayesian Normal AR (1) Model I

$$Y_t \mid y_{t-1} \sim N(\pi, \sigma^2),$$
$$\pi = \alpha + \beta y_{t-1}$$
$$\alpha \sim N(0,0.001),\ \beta \sim N(0,0.001),\ \tau \sim Gamma(1,30),\ \tau \sim 1/\sigma^2$$

Assessing empirical adequacy requires some statistics to characterise salient features of the data. In general, any summary statistics may be chosen, such as minimum sample value, maximum sample value, standard deviation, skewness, and so forth. However, when the concern is to check a specific assumption, it is vital to adopt statistics that capture those aspects of the data that relate to the assumption in question. The unemployment data shows strong positive dependence. A critical modelling concern is thus to select an appropriate independence hypothesis. This demands using statistics that capture the dependence feature of the data.[35] To this end, we may include among our diagnostic statistics autocorrelation functions of different order. Table 4.3 defines the statistics used here:

Table 4.3 Definition of vector of interest		
Preliminary statistics: $\bar{y}_T = \sum_{t=1}^{T} y_t / T \quad s_T = \sum_{t=1}^{T} (y_t - \bar{y}_T)^2 / T$ $\bar{y}_T^{(2)} = \sum_{t=1}^{T} y_t^2 / T$		
$T_1(.)$	Minimum sample value	$y_{\min}$
$T_2(.)$	Maximum sample value	$y_{\max}$
$T_3(.)$	Standard deviation	$(s_T)^{1/2}$
$T_4(.)$	Skewness	$\sum_{t=1}^{T} (y_i - \bar{y})^3 / T(S_T)^{3/2}$
$T_5(.)$	Excess kurtosis	$\left(\sum_{t=1}^{T} (y_i - \bar{y})^4 / T(S_T)^2\right) - 3$
$T_6(.)$	1st order autocorrelation	$\sum_{t=1}^{T-1} (y_t - \bar{y}_T)(y_{t+1} - \bar{y}_T) / \sum_{t=1}^{T} (y_t - \bar{y}_T)^2$
$T_7(.)$	2nd order autocorrelation	$\sum_{t=1}^{T-2} (y_t - \bar{y}_T)(y_{t+2} - \bar{y}_T) / \sum_{t=1}^{T} (y_t - \bar{y}_T)^2$
$T_8(.)$	3rd order autocorrelation	$\sum_{t=1}^{T-3} (y_t - \bar{y}_T)(y_{t+3} - \bar{y}_T) / \sum_{t=1}^{T} (y_t - \bar{y}_T)^2$

In principle, any of the predictive approaches can be used to search for priors. However, since in the current case the data shows strong dependence, cross validations techniques are not appropriate; they destroy the dependence feature of the data. For simplicity, we adopt Rubin's approach both for prior modelling and data model assessment. To derive the posterior predictive distributions of the

[35] See Rubin (1984:1168) on how a statistic may be defined to tell whether the data come from a normal or a Cauchy distribution. Also, Geweke (2001:5-6) contains a discussion on the choice of diagnostic statistics for assessing models of financial returns.

statistics, 10,000 samples are simulated from the posterior predictive distribution of the observable under the model, (with 5000 burnt in), and the values of the statistics for each sample is calculated. The values are used to calculate the quantiles of the predictive distributions of the statistics. Table 4.4 gives the quantiles as well as the predictive *p*-values for the observed values of the statistics.[36]

Table 4.4

Posterior predictive distribution of vector of interest

Normal AR (1) Model I

		Data	Median	(2.5%, 97.5%)	*p-value*
$T_1(.)$	Minimum value	2.57	1.557	(032,2.37)	0.9937
$T_2(.)$	Maximum value	7.37	8.189	(7.306,9.449)	0.0365
$T_3(.)$	Standard deviation	1.172	1.412	(1.197,1.648)	0.0152
$T_4(.)$	Skewness	0.2068	0.077	(-0.292,0.441)	0.757
$T_5(.)$	Excess kurtosis	-0.9009	-0.386	(-0.876,0.46)	0.0185
$T_6(.)$	1st order autocorrelation	0.919	0.541	(0.358,0.680)	1
$T_7(.)$	2nd order autocorrelation	0.743	0.4402	(0.2608,0.5746)	1
$T_8(.)$	3rd order autocorrelation	0.534	0.3192	(0.151,0.460)	0.9997

The observed value of the skewness and maximum value statistics are within the support of their distributions. However, the observed values of the rest of the statistics either fall outside the support of their distributions or lie in their extreme tail areas. Notably, none of the realised values of the autocorrelation functions falls within the interval (2.5%, 97.5%) of the predictive distributions; the simulated values of the functions are invariably smaller than their observed values. The model strikingly fails to account for most aspects of the data, and is empirically inadequate.

[36] To perform the simulations discussed in this section I have used the Bugs software, freely available on http://www.mrc-bsu.cam.ac.uk/bugs. I am greatly indebted to the Bugs programs accompanied Congdon's book (2001) as well as the programs found in the Bugs Manuals, available from the above website.

Another way of assessing the compatibility of the model with the data, suggested in Gelman et al. (1996), is to plot the simulated values of each statistic in form of a histogram to obtain a nonparametric estimate of the distribution of the statistic. The consistency of the distribution with the data is determined by locating the observed value of the statistic in the histogram. The following histograms show the distributions of the minimum value and standard deviation statistics under the model.

Model AR(1) I

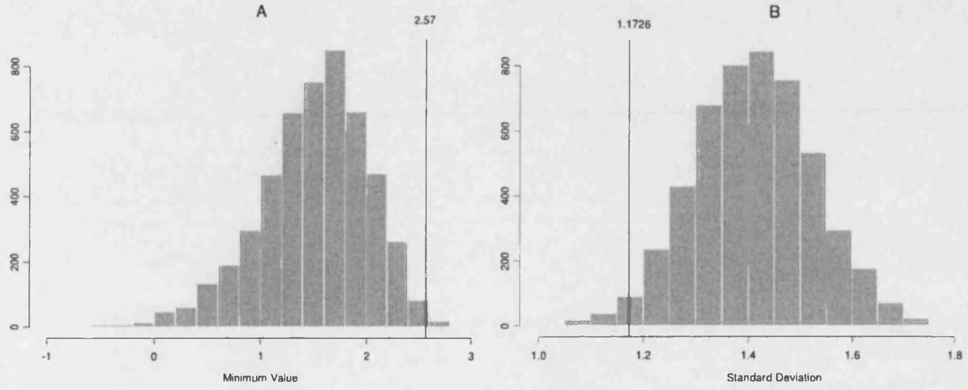


Figure 6.1 Posterior predictive distributions and the observed values for the sample minimum value and sample standard deviation statistics.

The lines indicate the position of the actual values of the statistics in their distributions, showing that the simulated samples D^{rep} almost invariably differ from the actual sample D. The failure of the model demands searching for alternative priors.

Experiments with alternative hyperparameters suggest that the ability of the model to account for the data is not sensitive to the choice of hyperparameters for the prior densities of α and β but is highly sensitive to those in the distribution of

τ. A relatively extensive experiment with alternative values for the hyperparameters suggests the priors given in the following model:

Normal AR (1) Model II

$Y_t \mid y_{t-1} \sim N(\pi, \sigma^2)$,

$\pi = \alpha + \beta y_{t-1}$

$\alpha \sim N(0,0.001)$, $\beta \sim N(0,0.001)$, $\tau \sim Gamma(0.1,0.1)$, $\tau \sim 1/\sigma^2$

The posterior predictive distributions implied by these new priors are given in Table 4.5 below.

Table 4.5

Posterior predictive distribution of vector of interest

Normal AR (1) Model II

		Data	Median	(2.5%, 97.5%)	*p-value*
$T_1(.)$	Minimum value	2.57	2.343	(1.672,2.824)	0.799
$T_2(.)$	Maximum value	7.37	7.524	(6.973,8.245)	0.302
$T_3(.)$	Standard deviation	1.172	1.176	(1.056,1.303)	0.4741
$T_4(.)$	Skewness	0.2068	0.149	(-0.083,0.377)	0.687
$T_5(.)$	Excess kurtosis	-0.9009	-0.685	(-1.02, -0.201)	0.122
$T_6(.)$	1st order autocorrelation	0.919	0.785	(0.707,0.840)	1
$T_7(.)$	2nd order autocorrelation	0.743	0.6391	(0.5572,0.701)	0.9993
$T_8(.)$	3rd order autocorrelation	0.534	0.4638	(0.3709, 0.520)	0.9581

As the quantiles in the Table show, the new priors enable the model to better account for the features of the data captured by the statistics. Unlike the previous model, the actual values of the statistics minimum sample value, standard deviation, excess kurtosis, and the 3rd order autocorrelation fall within the support of the distributions, i.e., the (2.5%, 97.5%) predictive interval. Nevertheless, the model still fails to account for the 1st order and 2nd order autocorrelation functions. Since experiments with alternative priors for τ do not improve on the adequacy of the model, and the performance of the model is not sensitive to the choice of priors for α and β, there is every reason to think that the data model is not correctly specified.

In general, if experiment with a wide range of hyperparameters fails to produce good priors, other distribution families consistent with the subject matter information should be considered. If, after an adequate search among alternative distribution families, a model still fails to account for the data, a revision of the data model assumptions becomes necessary (Geweke, 2001:7). It should be clear from this analysis that the diagnostic approach to adequacy assessment presents a powerful alternative to the traditional prior modelling methods. The approach forgoes the need for qualitative distribution summaries or hypothetical predictions. And more importantly, it overcomes the risk of rejecting a correctly specified data model because of the choice of inappropriate priors. It ties the choice of priors to the model adequacy.

6.2.2 Exploring Data Model Assumptions

When a model fails to account for the data regardless of the choice of priors, the focus of investigation must be turned towards alternative data models. The investigation involves varying the data model assumptions one at a time, searching for a set of priors that best enables the model to account for the data, and checking the model adequacy. Although the failure of the above model may be due to any of the basic assumptions, because of its specific failure in accounting for the dependence feature of the data, it is more plausible to first investigate the effect of varying the first order Markov condition. We proceed by replacing it with the second order Markov condition. Modifying the distribution hypothesis appropriately, this hypothesis leads to a second order normal autoregression model:

Normal AR (2) Model

$$Y_t \mid y_{t-1}, y_{t-2} \sim N(\pi, \sigma^2),$$
$$\pi = \alpha + \beta y_{t-1} + \gamma y_{t-2}$$

Experiments with alternative priors suggest that the following set of priors best enables the model to account for the data:

$$\alpha \sim N(0,0.01),\ \beta \sim N(0,0.01), \gamma \sim N(0,0.01), \tau \sim Gamma(1,3), \tau \sim 1/\sigma^2$$

Table 4.6 gives the (2.5%, 97.5%) predictive intervals of the posterior predictive distributions of the statistics resulting from these priors. The predictive intervals are computed from 10,000 samples simulated from the posterior predictive distribution of the observable under the model, (with 5000 burnt in).

Table 4.6
Predictive Distribution of Vector of Interest
Normal AR (2) Model

		Data	Median	(2.5%, 97.5%)	*p-value*
$T_1(.)$	Minimum value	2.57	2.21	(1.265,2.771)	0.876
$T_2(.)$	Maximum value	7.37	7.668	(7.081,8.442)	0.1836
$T_3(.)$	Standard deviation	1.172	1.172	(1.055,1.296)	0.5
$T_4(.)$	Skewness	0.2068	0.1872	(-0.088,0.448)	0.558
$T_5(.)$	Excess kurtosis	-0.9009	-0.46	(-0.854,0.158)	0.0127
$T_6(.)$	1st order autocorrelation	0.919	0.688	(0.573,0.768)	1
$T_7(.)$	2nd order autocorrelation	0.743	0.576	(0.482,0.6503)	1
$T_8(.)$	3rd order autocorrelation	0.534	0.376	(0.272,0.4666)	0.9999

These predictive quantiles are not improved by considering alternative values for the hyperparameters or alternative prior distribution families. The current parameterisation seems to enable the model to best fit the data. If so, the model does not improve on the second AR(1) model. In fact, contrary to the latter model, it accounts neither for the excess kurtosis nor for the 3rd autocorrelation statistic.

Replacing the first order Markov condition with higher order Markov conditions does not create an empirically more adequate model. Nor does the data show any heterogeneity to consider alternative homogeneity assumptions. Experiments with alternative distributions such as student *t*-distribution also fail to yield a better model. In all these cases, however, the residual ACF function has some large spikes at low lags, indicating that the errors are correlated. This suggests using an Autoregressive Moving Average (ARMA) model.[37] To continue, consider an ARMA (1,1) model,

[37] An ARMA model can theoretically be regarded as an efficient approximation to an Autoregression model of some order *p* (Spanos, 1999:452).

Normal ARMA(1,1) Model

$Y_t \mid y_{t-1}, y_{t-2} \sim N(\pi, \sigma^2)$,

$\pi = \alpha + \beta y_{t-1} + \gamma \varepsilon_{t-1}$

Experiments with alternative priors soon lead to the following densities,

$$\alpha \sim N(0,0.3),\ \beta \sim N(0.1,0.1), \gamma \sim N(0.1,0.1), \tau \sim Gamma(0.01,0.01), \tau \sim 1/\sigma^2.$$

Table 4.7 gives the (2.5%, 97.5%) predictive intervals of the predictive distributions of the statistics that result from these priors. The intervals are computed from 10,000 samples simulated from the posterior predictive distributions, (with 5000 burnt in).

Table 4.7

Predictive Distribution of Vector of Interest

Normal ARMA (1,1) Model

		Data	Median	(2.5%, 97.5%)	*p-value*
$T_1(.)$	Minimum value	2.57	2.538	(2.315,2.538)	0.624
$T_2(.)$	Maximum value	7.37	7.399	(7.189,7.629)	0.398
$T_3(.)$	Standard deviation	1.172	1.177	(1.149,1.204)	0.39
$T_4(.)$	Skewness	0.2068	0.1576	(0.085,0.229)	0.413
$T_5(.)$	Excess kurtosis	-0.9009	-0.886	(-1, -0.755)	0.91
$T_6(.)$	1st order autocorrelation	0.919	0.914	(0.906,0.921)	0.909
$T_7(.)$	2nd order autocorrelation	0.743	0.741	(0.724,0.757)	0.598
$T_8(.)$	3rd order autocorrelation	0.534	0.5395	(0.514,0.564)	0.355

The Normal ARMA (1,1) model accounts for all the aspects of the data captured by the diagnostic functions. In particular, it accounts for the dependence features of the data. Moreover, the performance of the model is not sensitive to particular hyperparameters, as a very wide range of hyperparameters preserves the ability of the model to account for the data. We at last have a candidate model fitting the data.

The unemployment data set is a very simple one. The analysis, nevertheless, illustrates the essence of the Bayesian diagnostic approach to model specification, which offers a powerful tool for searching the space of candidate models to find a model capable of accounting for the data. The diagnostic approach to specification searches, combined with the procedures introduced for initial model formulation, furnishes the key elements of an exploratory theory of Bayesian model formulation.

7 Model Selection

Exploratory searches may generate several models equally fitting the data, raising the issue of how a model should be chosen from among the candidates. We earlier described the Bayesian solution to this problem, and now return to it to discuss some controversies surrounding it and highlight certain complexities in establishing a theory of statistical learning. At the start, following Bernardo et al. (1994), it is . useful to make a distinction between two possible views (perspectives) that can be held with respect to a set of candidate models:

> ***Closed view***: the set of candidate models $\{M_1,...,M_k\}$ is complete in the sense that it includes the true model.
>
> ***Open view***: the set of candidate models$\{M_1,...,M_k\}$ is incomplete in the sense that it excludes the true model, either because the model is not among the candidates or because there is no true model anyway.

The closed view stands on two assumptions: a metaphysical assumption that there exists a *true* model and an epistemological assumption that the true model is *actually* among the candidate models. The model selection problem is thus defined as that of finding the true model from among the candidates. The open view emerges from rejection of at least one of these assumptions and can be interpreted in two ways. One interpretation takes for granted the existence of a true model but acknowledges that it may not be among the candidates. In this case, the model selection issue involves selecting the model that best approximates the true model. The other interpretation rejects the reality of a true model outright, considering the candidate models simply as a set of models contending to account for the data. In this case, the model selection issue is simply to find a simple model that best fits the data and yields accurate predictions.

The Bayesian theory of model selection takes the closed view for granted. It interprets the probability of a model as the probability that it is true (Hill, 1990:61). And, it requires the probabilities over the candidate models to add up to one, meaning that the true model is among the candidates (Wasserman, 2000:103). Consequently, it recommends selecting the model that scores the highest probability in the light of the data. Both assumptions underpinning the closed view are flawed.

The metaphysical assumption of a true model encounters problems of interpretation from the subjectivist viewpoint. From this perspective, probability is the product of thinking consistently about the universe, with no counterpart in the external world (Dawid, 2002:8), and there is no true probability model or data generating process involving parameters that attain an objective existence (Poirier, 1988:122, Leamer, 1990:188). Therefore, for a subjectivist, the truth of a model can only be defined in terms of features of the observables. To elaborate on this, suppose it was possible to observe endlessly a socio-economic or physical process generating data sets of size *N*. Suppose also it was possible to simulate endlessly samples of the same size from a candidate model purporting to describe the system. The model could be said to be true if the stylised features of the simulated samples (such as sample mean, median, minimum value, maximum value, covariance, and so forth) *arbitrarily closely* resembled those of the actual samples. This or something similar seems to be the only way to define a true model in the subjectivist framework. If so, the question arises about the rationale for supposing a unique model generating samples most closely resembling the actual samples. To define 'arbitrarily closely', it is necessary to introduce some distance function. There are, however, many possible distance functions, and depending on the choice of metric, different models may turn out to be true. There is no natural choice of a distance function. All in all, even in the abstract it is not clear how to defend the existence of *a* true Bayesian model.

The epistemological assumption that the true model is among the candidate models is also indefensible for several reasons. The number of models that can be considered in practice is restricted by the finiteness of the reservoir of known models. None of the models may approximate the 'true' model. Moreover, in empirical modelling, due to the possibility of overfitting, the complexity of the models considered must always be tied to the sample size. With small samples, only simple models can be considered, since highly parameterised models are prone to overfitting. This restriction arising from the smallness of actual samples constrains the set of models that can be considered in practice, giving rise to the possibility that the allowed set may neither include the true model nor even a good approximation thereof (Spiegelhalter, 1995:72). Constructing models is also costly, time consuming, and constrained by computational capabilities of the day. The cost of developing a complex model with a better chance of approximating the reality may outweigh the practical benefit that may ensue. Such real pragmatic considerations compel the analyst to consider only a handful of models that may be very different from the true model. Thus, even if the metaphysical quandaries surrounding the existence of a true model are ignored, there are still serious reasons to doubt that the model is among the candidates considered.

The advocates of the Bayesian approach have argued that these objections do not undermine the heuristic role that the closed view plays in the advancement of science. It has been claimed that scientists proceed by adopting the working hypothesis that one of the models under consideration is true in order to analyse the merits of the models and to conduct further research. This tentative assumption transforms the set of models into a closed set, allowing the scientists to assign to them probabilities that add up to one (Wasserman, 2000:103). But, the claim that the only way to consider the merits of alternative models is to think of them as an exhaustive set containing the true model is unfounded. Models can be compared in respect of their predictive accuracy, simplicity, broadness, computability, and so forth. It is by no means necessary to think that one model is true and the other false to contrast the performance of two models.

Another attempt to retain the closed view involves adding to the candidate models $\{M_1,\dots,M_k\}$ a 'catchall' model M^C to represent 'all other models'. This formally transforms the candidate models into an exhaustive set but raises two questions that are difficult to answer. First, it is not clear how to assess the probability of the model $p(M^C)$. What is the probability that the 'true' model is not among the candidate models (Winkler, 1994:109)? Second, Bayesian model selection requires specifying the probability of the data given the model, i.e., $P(D / M^c)$. How can the probability of the data conditional on a model or set of models that is totally unknown be estimated (Anscombe, 1963)? The proposal makes no headway in addressing the problems facing the closed view.

A satisfactory account of model selection should take into account the fact that the candidate models might exclude the 'true' model. A departure from the closed view requires reinterpreting the probability of a model, redefining the goal of inference, specifying the features a model must have to be conducive to the goal, and describing methods for selecting a model with the requisite features. Interestingly, the Bayesian literature provides the elements of an alternative account of model selection that takes some steps in these directions. The account, defended by Geisser (1980), Lane (1986), and Bernardo et al. (1994), has its roots in de Finetti's representation theorem. On this theorem, as said earlier, statistical inference is primarily concerned with observables, and parameters enter the model just to simplify the relations among the observables, and have no independent meaning (Lindley, 1982:77). Since the distribution of the observables under a Bayesian model is given by the predictive distribution, the probability assigned to a model is best understood as the confidence that one has in the model's ability to yield accurate predictions. This permits comparing the relative probability of any set of models, regardless of whether the set is exhaustive or not (Lane, 1986:256). From this perspective, the primary objective of inference is to generate an accurate predictive distribution (Lane, 1986:254; Poirier, 1988:132), and a highly desirable feature of a Bayesian model is its ability to generate accurate predictions.

Epistemic considerations aside, constructing accurate models (and hence accurate predictive distributions) is costly, time consuming, and subject to computational and tractability constraints. A satisfactory account of model selection should also take these features of the real-life inference situations into account. Taken together, these considerations suggest redefining the issue of Bayesian model selection as the problem of selecting a model that is likely to produce the most accurate predictions subject to computational, time, cost, and other pragmatic constraints facing the analyst.

Transforming these remarks into a formal theory of model selection may require introducing a preference function weighting the competing goals of predictive accuracy, tractability, and affordability, and treating the whole model selection problem within the matrix of the expected utility theory. The call for establishing such a theory of model selection is by now old (Anscombe, 1963:89; Lindley, 1968) and still being insisted on (Draper, 1996:763; Hodges, 1987:262). Nevertheless, no serious contender has yet emerged. This partly has to do with the fact that pragmatic considerations are terribly difficult to quantify (Poirier, 1988:137). In the end, a fully formal theory of model selection may be as elusive as a 'true model' (Pesaran, et al., 1985).

8 Objections Revisited

The theory of diagnostic searches characterised in this chapter is founded on the core idea of the Fisherian concept of goodness of fit test, which has been criticised by Bayesian and non-Bayesian statisticians. To complete the discussion, we review some of the basic criticisms levelled against the use of p-values and data-driven model building in general.

The central objection to the use of *p-values* is that they imply an abrogation of the likelihood principle (LP) implied by two basic principles: the conditionality principle (CP) and the sufficiency principle (SP). Consider a parameter θ standing for the proportion of successes in a sequence of independent Bernoulli trials, say, the proportion of non-defective items produced by a machine. Further, consider two scenarios for collecting data to estimate θ. In the first scenario, E_1, N items are collected and the number of non-defectives k is counted, with N being predetermined. In the second scenario, E_2, sampling from the machine continues until k non-defective items are obtained, where $k > 0$ is a pre-determined integer, and the sample size happens to be N. The first experiment leads to a binomial distribution and the second to a negative binomial.

Within this context, the CP states that if we decide which of the experiments E_1 and E_2 to do by the flip of a coin, the finial inference must be the same as if the experiment had been chosen without flipping the coin (Cox, 1958). The SP, on the other hand, says when there exists a sufficient statistic for θ, two samples that yield the same value for the statistic provide the same evidence for θ.[38] These principles, as shown by Brinbaum (1962), necessitate the LP, which for the current purpose, can be stated as:

The Likelihood Principle: Consider two experiments $E_1 = \{Y_1, \theta, f_1(y_1 \mid \theta)\}$ and $E_2 = \{Y_2, \theta, f_2(y_2 \mid \theta)\}$ involving the same parameter θ. Suppose that for particular realizations y_1 and y_2 of the data, $L_1(\theta, y_1) = cL_2(\theta, y_2)$ for some constant c not depending on θ. Then, $Ev[E_1, y_1] = Ev[E_2, y_2]$, where $Ev[E_j, y_j]$ denotes the *evidence* about θ arising from experiment E_j and realised data y_j.

Informally, the principle "states that two experiments providing evidence about the same parameter θ which give rise to data realisations yielding likelihoods which are proportional, must provide the same evidence regarding θ" (Poirier, 1988:125). Since both the CP and SP seem plausible, the LP has become for many statisticians the yardstick against which to gauge the acceptability of a statistical procedure. Agreement with the LP is argued to be a minimal requirement that no statistical procedure can fail to fulfil.

Some simple examples reveal that conventional frequentist-based hypothesis testing procedures, based on assessments of p-values, abrogate the LP. A simple example, due to Lindley et al. (1976), is concerned with estimating θ in experiments similar to those described above. Suppose in the first experiment 12 items are collected and 9 non-defective items are found while in the second it has

[38] A sufficient statistic for θ is a function of the data which summarises all available sample information concerning θ. For example, if an independent sample $X_1,\ldots,X_N$ for $N(\mu, \sigma^2)$ distribution is to be taken, it is known that $T(\overline{X}, S^2)$ is a sufficient statistic for $\theta = (\mu, \sigma^2)$, where $\overline{X}$ stands for the sample mean and $S^2 = \sum_i (X_i - \overline{X})^2 / N - 1$ (Berger,1985:35). This common definition, which underlies the sufficiency principle, assumes that the model is known to be 'true'. Otherwise, a different definition of sufficiency is needed, and the sufficiency principle will no longer be valid (Hill, 1986:217).

taken sampling 12 items to collect 9 non-defectives. The likelihoods for these experiments are given respectively by

$$L_1(\theta,k)=\frac{12!}{9!3!}\left[\theta^9(1-\theta)^3\right]=220\left[\theta^9(1-\theta)^3\right],$$

and

$$L_2(\theta,k)=\frac{11!}{2!9!}\left[\theta^9(1-\theta)^3\right]=55\left[\theta^9(1-\theta)^3\right].$$

These likelihoods are proportional to each other, i.e., $L_1(\theta,k)=4L_2(\theta,k)$. Therefore, according to the LP, both experiments provide the same information about θ, and must lead to the same inferences about the parameter. However, consider testing the null hypothesis:

$$H_0\equiv\theta=1/2.$$

The p-value $p_{\theta=0.5}(Y\geq 9)$ is 0.075 under E_1 and 0.0325 under E_2. If the significance level is set at 0.05, the first experiment suggests accepting the null hypothesis but the second suggests rejecting it. Thus, the frequentist-based hypothesis testing procedures, which require calculating p-values, violate the LP. To see the cause of the conflict, note that even though the observed data are the same in E_1 and E_2, the sample spaces are different. In E_1 the sample space is given by $\{0,1,...,N\}$ and in E_2 by $\{m,m+1,...\}$ (m denotes the number of defectives). Since in computing p-values the whole sample space is considered, not the realised data alone, the difference leads to conflicting inferences (Poirier, 1988:126).

Based on this example, Lindley et al. (1976) and others have criticized the frequentist testing methods, arguing that inferences about statistical hypotheses must be conditioned only on the observed data, which requires abandoning p-values.[39] And when Box (1980) proposed prior predictive p-values for adequacy

[39] See Barnett (1999, pp.181-183).

assessment, Lindley repeated the criticism that it would lead to an abrogation of the LP (Lindley, 1980:423). This critique is misplaced. The LP conditions upon the choice of a model and is therefore only relevant to the estimation phase of inference, where the truth of the model is taken for granted. When the concern is to construct a model that fits the data and no decision has yet been made about whether the model is true or not, the principle has no regulative force at all (Box, 1983:74). The conflict between the LP and *p*-values can be resolved by recognising that the former belongs to the estimation phase of inference while the latter to the model formulation phase (McCullagh, 1995:178). There is then no conflict between the LP and the use of *p*-values. As a final point, the LP also looses its regulative force if one takes the role of parameters to be solely instrumental (Lane, 1986:257).

A second objection against the use of *p*-values is that in large samples they lead to rejection of any model by locating minor deficiencies that are otherwise unimportant (Pratt, 1965). This is not really a deficiency at all. In practice, all models are imperfect, and it is therefore highly desirable to have exploratory methods that can reveal deficiencies in currently held models as the sample grows (Hodges, 1990:87-88). The deficiencies with a model might be ignored for various practical reasons but it is still useful to discover them with an eye to ultimately improving it (Gelman et al., 1996: 800).

Thirdly, it has been objected that there is no guidance to decide when a *p*-value is extreme enough to warrant rejecting a model. This criticism fails to appreciate that Bayesian *p*-values are not for testing or rejecting models. They are just to show whether a model fits the data and, if not, help searching for a model with a better goodness of fit (Dempster, 1983:124). The right question to ask is when a *p*-value is extreme enough to justify the search for an alternative model. There is, however, no purely epistemic response to such questions. The decision whether to take a discrepancy between a model and the data seriously and search for an alternative model is to a large extent driven by pragmatic considerations, which are by no means unique to the use of *p*-values (Anscombe, 1963:89); they are needed at every stage of modelling.

Finally, a serious matter about any exploratory procedure concerns the borderline between model searching and data mining. It is always possible to find a model with a better goodness of fit by exploring increasingly more complex models but such a model may not necessarily perform better over future data. Why should then one search for a model that best fits the data? Several things can be done to meet this concern. First of all, predictive searches must be carried out within the class of models warranted by the existing subject matter information. Secondly, having found a model fitting the data, an essential aspect of modelling is to assess the sensitivity of the model to the underlying assumptions that are in doubt. In the end, the only way to gain serious confidence in a model is to try it over new and diverse data. There is never a substitute for new data. Model building is a complex problem and there is rarely a simple solution to a complex problem.

9 Conclusion

The last chapter began the study of the bounded rationality paradigm that seeks to explain economic phenomena by modelling the economy as a society of intuitive statisticians. The paradigm raises the question if there can be a 'tight enough' theory of model formulation. The chapter showed that constructing an interpretable model requires starting with a parametric model, and model formulation cannot be left to the data. This led us to consider if there can be a 'tight enough' theory of parametric model formulation. To this end, the present chapter studied the theory of Bayesian inference, which, as traditionally understood, is a form of parametric inference, and is commonly believed to offer a model of learning from experience.

The chapter started by examining some foundational issues to provide a correct interpretation of the Bayesian theory. It argued that the theory was concerned solely with coherent analysis of uncertainty attitudes towards a closed set of specified models. And, it has no implication for dynamics of beliefs except that partial beliefs at any moment in time ought to accord with the laws of probability. Considering the limitations of the Bayesian theory, the chapter next outlined various phases of parametric inference including initial model formulation, model fitting (estimation), empirical model assessment, re-specification analysis, and model selection. Coherent analysis is relevant only to model fitting and model selection. Any attempt at explaining other central aspects of learning necessitates broadening the scope of the Bayesian theory.

As a step towards establishing a broader theory of inference within the Bayesian framework, the chapter decomposed a Bayesian model into a data model – which consisted of a set of internally consistent hypotheses of independence, homogeneity, and distribution – and a (joint) prior distribution. Using this decomposition, it outlined a theory of model formulation which dealt with major aspects of learning from data, notably initial model specification, empirical model assessment, and model re-specification.

Contrary to a dominant belief, the chapter pointed to three contributions form theoretical statistics that furnished the essential elements of a theory of initial model formulation. The first contribution was a large collection of independence and homogeneity hypotheses, as well as a large list of probability distribution families, which could be used as building blocks for constructing countless models. The existence of the model reservoir made it possible to define initial data model specification as the problem of selecting a set of internally consistent basic hypotheses from the three categories of known independence, homogeneity, and distributional assumptions to form a model, which can account for the data. The second contribution was the class of characterisation theorems while the third was a rich literature on *ex ante* consequences of alternative models. These contributions led to two general exploratory methods for initial selection of the assumptions entering a model.

Initial model specification involves a variety of subjective decisions at several levels, which necessitate assessing the empirical adequacy of the model before using it. To provide a framework for adequacy assessment, the chapter introduced the notion of *ex post* consequences of a Bayesian model, consisting of the predictive distributions the model implies for various functions of the data. The empirical adequacy of a Bayesian model was defined in terms of consistency of the predictive distributions with the actual values of the statistics. These ideas, joined with the Fisherian idea of goodness of fit testing, led to a powerful exploratory method for searching the space of candidate models warranted by pre-estimation considerations. The method involved beginning with a candidate model, assessing its adequacy, and monitoring the effect of varying one assumption at a time until an empirically adequate model is found.

These concepts and methods provided the necessary elements for an exploratory theory of model construction, according to which the process of model specification starts with examining the *ex ante* consequences of known basic hypotheses to construct an initial model or a set of initial models capable of accounting for the data. The process next involved assessing the *ex post* consequences of the candidate models, which survived initial analysis, to locate a model or models that accurately accounted for the data.

It was also pointed out that there was a pragmatic side to statistical inference. Model construction is costly, time consuming, constrained by computational capabilities, and influenced by the purported use of the model. If the goal is to explain how a statistician approaches a data set and constructs a model, the entire modelling process ought to be thought of as a constrained optimisation problem. This account of the model formulation process has significant implications for establishing a theory of statistical learning and thus the bounded rationality project, some of which are stated below:

First, in the above theory of model formulation, background information enters inference in many forms: most notably, in the form of a reservoir of models (Arthur, 2000), knowledge of the conditions under which the models are appropriate, and knowledge of the *ex ante* implications of the models. An extremely significant point is that a theory of parametric learning takes such information as *given*, which means there can be no general theory of parametric inference that can also explain where the models or, more precisely, basic probabilistic hypotheses come from in the first place. Only after a reservoir of models is given, it is possible to speak of a theory of parametric learning. The necessity of a model reservoir, whose generation cannot be explained by a theory of parametric learning, might have been the principal reason for Fisher and other statisticians' negative feeling concerning the possibility of a theory of model specification (Lehmann, 1990:161). The search for a theory of statistical learning, therefore, encounters a dilemma. If a nonparametric approach to learning is pursued, it would be impossible to build an interpretable model of several variables with ordinarily available samples. If, on the other hand, a parametric approach is taken, the question arises as to where the models come from in the first place.

Second, the scope of a theory of parametric learning is defined by the scope of the model reservoir. So far, only a few multivariate distribution families have emerged, and, because of this scarcity, any set of models considered in practice is likely to exclude the 'true' model or a good approximation thereof. It thus seems fair to question the relevance of the convergence results established in the learning literature; all these results are based on the presumption that the true model is among the candidate models.[40] The relevance of the convergence results becomes even more suspect when we realise the necessity of subjective and pragmatic considerations in modelling data.

Third, since pragmatic considerations influence decisions about the adequacy of a model, the hypothesis that the agent behaves like a Bayesian statistician, even if true, would not be adequate for predicting his model. To predict the agent's model, it is also necessary to know his goals, preferences, and constraints. This makes it even more difficult to establish a precise and informative theory of how he actually models the economy.

All in all, the claim that by modelling people as intuitive Bayesian statisticians we can predict the probabilistic models that they construct of the economy should be treated with scepticism. The most that can be predicted on the basis of this hypothesis, the history of the observables, and the expected utility maximisation principle, is that he takes an action that is optimal with respect to his utility function and view of the environment. The serious issue with the IS hypothesis is

[40] When what is at issue is the structural specification of how known and unknown quantities are related one cannot count on "the data to swamp the priors" (Draper, 1995).

not that people are not perfect statisticians but that, even if they were, the hypothesis would still fall short of producing informative predictions. More will be said on this in the next chapter.

Appendices

Appendix A: Lindley's Paradox

Lindley's paradox shows a disagreement between sampling theory and Bayesian methods, first noted by Jeffreys (1939). The paradox illustrates that a "sharp null hypothesis may be strongly rejected by a standard sampling [...] theory test of significance and yet be awarded a high odds by a Bayesian analysis based on a small prior probability for the null hypothesis and a diffuse distribution of one's remaining probability over the alternatives" (Shafer, 1998:2257). As an illustration, following Bernardo and Smith (1994:394), suppose for data $D = \{x_1,\ldots,x_n\}$ the set of candidate models are M_1 and M_2, corresponding to the simple and composite hypotheses about θ in $N(x_i \mid \theta,\phi)$ defined by

$$M_1:\quad p_1(D) = \prod_{i=1}^{n} N(x_i \mid \theta_0,\phi), \qquad \theta_0,\ \phi \text{ known;}^{41}$$

$$M_2:\quad p_2(D) = \int\prod_{i=1}^{n} N(x_i \mid \theta,\phi)N(\theta \mid \varphi,\eta)d\theta, \quad \phi,\ \varphi,\ \eta \text{ known.}$$

Since $\bar{x} = N^{-1}\sum_{i=1}^{N} x_i$ under both models is a sufficient statistic, the Bayes factor in favor of M_1 against M_2 is given by

$$B_{12} = \frac{N(\bar{x} \mid \theta_0,n\phi)}{\int N(\bar{x} \mid \theta,n\phi)N(\theta \mid \varphi,\eta)d\theta}$$

$$= \left(\frac{\eta+n\phi}{\eta}\right)^{1/2} \frac{\exp\left\{2^{-1}(\eta^{-1}+(n\phi)^{-1})^{-1}(\bar{x}-\varphi)^2\right\}}{\exp\left\{2^{-1}n\phi(\bar{x}-\theta_0)^2\right\}}$$

For any fixed sample D, $B_{12} \to \infty$ as the prior precision η in M_2 approaches zero. This in turn pushes the posterior probability $p(M_1 \mid D)$ towards unity, regardless of the data. In many cases, however, the null hypothesis is rejected by the sampling significance tests (See Lee 1997:128 for an example).

The general, and in fact more important, lesson learnt from the paradox is that, in any Bayesian model comparison, the Bayes factor can depend on the prior distributions specified for the parameters of each model (Bernardo et al. 1994:394), and the effect of the priors on the Bayes factor remains even when the sample size grows (Kass et al. 1993:555).

[41] Here, ϕ is taken to be precision, defined as $1/\sigma^2$.

Appendix B: Bayesian Normal/Chi-squared Model[42]

Consider the case where we have a set of observations $D=\{x_1,...,x_n\}$ thought to come from distribution $N(\theta,\phi)$, with θ and ϕ both unknown. So,

$$p(x/\theta,\phi)=(2\pi\phi)^{-1/2}\exp\{-\frac{(x-\theta)^2}{2\phi}\}.$$

The likelihood function is given by

$$\ell(\theta,\phi/x)\propto p(x/\theta,\phi)\propto\phi^{-n/2}\exp\{-\frac{\sum(x_i-\theta)^2}{2\phi}\}$$

$$=\phi^{-n/2}\exp[-\frac{\{\sum(x_i-\bar{x})+n(\bar{x}-\theta)^2\}}{2\phi}]=\phi^{-n/2}\exp[-\frac{\{S+n(\bar{x}-\theta)^2\}}{2\phi}],$$

where $S=\sum(x_i-\bar{x})^2$.

The conjugate prior distribution of ϕ is (a multiple of) an inverse chi-squared on ν_0 degrees of freedom. (The term 'degree of freedom' is just a name for a parameter). That is,

$$p(\phi)\propto\phi^{-\nu_0/2-1}\exp(-S_0/2\phi).$$

The conjugate prior distribution of θ conditional on ϕ is normal with mean θ_0 and variance ϕ/n_0. Then

$$p(\theta/\phi)=(2\pi\phi/n_0)^{-1/2}\exp\{-\frac{(\theta-\theta_0)^2}{2(\phi/n_0)}\}.$$

The joint prior distribution is thus a normal /chi-squared distribution with density function

$$p(\theta,\phi)=p(\phi)p(\theta/\phi)\propto\phi^{-(\nu_0+1)/2-1}\exp[-\frac{1}{2}\{S_0+n_0(\theta-\theta_0)^2\}/\phi]$$

$$=\phi^{-(\nu_0+1)/2-1}\exp\{-\frac{1}{2}\{Q_0(\theta)/\phi\},$$

where $Q_0(\theta)=n_0\theta^2-2(n_0\theta_0)\theta+(n_0\theta_0^2+S_0)$.

[42] Based on Lee, 1997:65-71

The posterior is

$p(\theta,\phi/D) \propto p(\theta,\phi)\ell(\theta,\phi/D)$

$\propto \phi^{-(\nu_0+n+1)/2-1} \times \exp[-\frac{1}{2}\{(S_0+S)+n_0(\theta-\theta_0)^2+n(\theta-\bar{x})^2\}/\phi]$

$=\phi^{-(\nu_1+1)/2-1} \times \exp\{-\frac{1}{2}\{Q_1(\theta)/\phi\},$

where $\nu_1=\nu_0+n$

and $Q_1(\theta)=(S_0+S)+n_0(\theta-\theta_0)^2+n(\theta-\bar{x})^2$

$=(n_0+n)\theta^2-2(n_0\theta_0+n\bar{x})\theta+(n_0\theta_0^2+n\bar{x}^2+S_0+S)$

$=S_1+n_1(\theta-\theta_1)^2$

$=n_1\theta^2-2(n_1\theta_1)\theta+(n_1\theta_1^2+S_1)$,

where

$n_1=n_0+n$;

$\theta_1=(n_0\theta_0+n\bar{x})/n_1$; and

$S_1=S_0+S+n_0\theta_0^2+n\bar{x}^2-n_1\theta_1^2$
$=S_0+S+(n_0^{-1}+n^{-1})^{-1}(\theta_0-\bar{x})^2.$

The posterior for ϕ is

$\phi \sim S_1\chi_{\nu_1}^{-2}$

and that for θ given ϕ is

$\theta/\phi \sim N(\theta_1,\phi/n).$

Chapter 5

Homo Economicus as an Intuitive Statistician (3)

Data Driven Causal Inference

1 Introduction

> I would rather discover a single causal relationship than be king of Persia
>
> Democritus[1]

The last two chapters studied some aspects of the bounded rationality program that views the economy as a society of intuitive statisticians – the intuitive statistician hypothesis. This hypothesis raises the question whether there is a 'tight enough' theory of statistical inference. To establish a general framework for addressing this issue, this thesis tentatively conjectured that the agent first seeks to learn the probability distribution of the variables representing his choice situation, and next uses the probabilistic information to learn about the causal structure of the situation. The last two chapters studied some of the issues surrounding learning the probability distribution of a set of variables. This chapter studies the second stage of learning that is concerned with inferring the causal structure of a set of variables from their joint distribution.

The first chapter studied the regression method of causal inference. According to this method, to infer whether *X* causes *Y*, one has to include in the regression equation of *Y* on *X* various combinations of potential confounders of *X* and *Y*. If the coefficient of *X* differs from zero regardless of what potential confounders are in the equation, *X* is said to cause *Y* (Cox, 1992). The method fails on several grounds. First, it cannot establish whether an association between *X* and *Y* is because of a direct causal link or latent common causes. Second, conditioning on potential confounders can turn an otherwise consistent estimate of the effect of *X* on *Y* into an inconsistent estimate. Also, controlling for the effects of the response variable *Y* can lead to wrong causal conclusions (Cox, 1958:48, Pearl, 2000:76). It has been argued that these problems can be overcome only by relying on subject matter information about the underlying system.

[1] Quoted from Whittaker, 1990.

An approach pioneered by Spirtes, Glymour and Scheines (SGS, hereafter) and Judea Pearl and his co-researchers is claimed to evade the difficulties facing the traditional methods to causal inference. These authors have argued that the reason for the failure of the traditional methods lies in two things: the lack of an efficient language for representing causal structures *and* the absence of a precise characterisation of the connection between probability and causation. Once an adequate language for representing causal structures is developed and the principles connecting causation and probability are defined, reliable causal conclusions can be derived from data alone. The claim for the necessity of theoretical knowledge in casual inference is exaggerated:

> In the social sciences there is a great deal of talk about the importance of 'theory' in constructing causal explanations ... In many of these cases the necessity of theory is badly exaggerated (Spirtes et al. 1993:133)
>
> In the absence of very strong prior causal knowledge, multiple regression should not be used to select the variables that influence an outcome or criterion variable in data from uncontrolled studies. So far as we can tell, the popular automatic regression search procedures [like stepwise regression] should not be used at all in contexts where causal inferences are at stake. Such contexts require improved versions of algorithms like those described here to select those variables whose influence on an outcome can be reliably estimated by regression (Spirtes et al. 1993: 257).[2]

The approach proposed by these authors, which is termed the Graph Theoretical (GT) or Bayes net approach, has been the source of many advances. In particular, it has led to the development of an efficient language for representing causal structures, a precise formulation of the principles underpinning the traditional methods of causal inference, and to a variety of new algorithms for causal inference. It has also advanced our understanding of the important issue of statistical indistinguishability of causal models. This chapter uses this approach to study the boundaries of data-driven causal inference. By data-driven causal inference, we mean any effort to draw causal

[2] Similar remarks are found in Pearl and Verma (1991).

conclusions from probabilistic data using only general subject-matter-independent principles supposedly linking causation and probability. A claim for a data-driven approach to causal inference, therefore, raises two fundamental issues. The first is whether there are any *universal* principles correctly linking probabilistic and causal dependencies. And, the other is whether the principles are sufficient for inferring the causal structure of a set of variables from their joint probability distribution.

This chapter investigates both topics by focusing on the principles underlying the GT approach, which are the most general principles that can possibly be true of the connection between probability and causation. After a brief description of the GT approach, the chapter takes up the issue of model equivalence. It argues that for every causal model consistent with the data, there are simple rules that allow generating a class of statistically equivalent causal models that have very little or nothing in common. As a consequence, even if the validity of the GT principles is not challenged, one can learn very little from data alone. The chapter next investigates the general validity of the principles. It argues that none of the arguments put forward for the principles justifies their universal validity. In addition, it shows how the possibility of selection bias undermines the claim that the GT approach outperforms other causal inference methods by being able to establish whether a correlation is *definitely* due to latent common causes. Moreover, it shows why, because of the possibility of mistaking the concomitant of a cause for the cause, the GT approach cannot establish the existence or absence of a causal link either. In the end, by reflecting on the limitations of the GT approach, the chapter sketches out an alternative account of causal inference from observational data, explains the role that the GT techniques play in the account, and spells out some implications of the analysis for the bounded rationality project.

The plan of the chapter is as follows: Section 2 defines some formal notions and introduces the axioms put forward by the GT approach for connecting causal and probabilistic dependencies. Section 3 explains the process of causal inference in the GT approach. Section 4 concentrates on the notion of statistical indistinguishability of causal models. Section 5 examines possible justifications for the axioms, and studies issues arising from aggregation, selection bias and concomitants. Section 6 concludes the chapter.

2 Preliminaries and Principles

The GT approach is built on formal notions that are new to the traditional literature on causal inference. This section begins by describing the notion of causation used here, and defines the concepts of 'causal structure' and 'causal model'. It next describes the method of path analysis which will be used to introduce some key graph theoretic notions. The section then characterises the class of candidate causal models in the GT approach, the data used for causal inference, and the principles used to link the data with a causal model.

2.1 Causal Structure

Causation is perhaps the most philosophically controversial topic, and this makes it impossible to offer an account of rival views in a handful of pages. For brevity, we only explicate the view of causation adopted here. Our view of causation has affinity with the manipulative account of causation, defended in the writings of philosophers such as Collingwood ([1940], 1948), Gasking (1955), and von Wright (1971). According to this account, a causal relationship primarily obtains between single events. An event x causes an event y if it was *in principle* possible to alter y by wiggling x. Or in Collingwood's terms, "that which is 'caused' is an event in nature, and its 'cause' is an event or state of things by producing or preventing which we can produce or prevent that whose cause it is said to be ([1940], 1948:285)." If it were not even hypothetically possible to alter y by wiggling x, x would not be a cause of y. This intuition, central to the manipulative account, is basic to analysis of actions in decision theory and evaluation of policies in economics.

The causal relation 'event x causes event y' is transitive, irreflexive, and antisymmetric. Particular events can be classified into types of events and types of events can be coupled with their complementary type events to form variables. Consider the rise in the Dow-Jones Industrial Average last Monday. We may classify this event into event type of 'rises in the Dow-Jones Industrial Average'; call it D. And, we may further put together this type event with its complementary type event 'declines in the Dow-Jones Industrial Average' to define the random variable 'the Dow-Jones Industrial Average'; call it $X \equiv (D, D^c)$.[3] Similarly, we may join together the type events 'rises in the FTSE 100' and 'declines in the FTSE 100' to define the random variable 'the FTSE 100; call it $Y \equiv (C, C^c)$. We say that variable X causes variable Y if and only if at least one member of types (D, D^c) causes at least one member of types (C, C^c) (Sobel, 1995:8). Having said this, we hereafter suppress the talk of particular events to the background.

Let $\mathbf{V} = \{X_1, ..., X_n\}$ be the set of variables necessary for describing a choice situation or a certain aspect of the economy. A proper subset of $\boldsymbol{V}$, $\boldsymbol{X}$, is said to be a *full* cause of X_m $(X_m \notin \mathbf{X})$ with respect to $\boldsymbol{V}$ if (i) there is a set of values $\boldsymbol{x}$ for $\boldsymbol{X}$ and a value x_m for X_m such that were it possible to set $\boldsymbol{X}$ at value $\boldsymbol{x}$, X_m would take on value x_m regardless of the value of other variables in $\boldsymbol{V}$ and (ii), no proper subset of $\boldsymbol{X}$ satisfies condition (i). In line with Spirtes et al. (1993:44), variable X_i is said to be a *direct cause* of X_m relative to $\boldsymbol{V}$ if X_i is a member of a full cause $\boldsymbol{X}$ of X_m in $\boldsymbol{V}$. By the same token, X_i is said to be an *indirect cause* of X_m relative to $\boldsymbol{V}$ if there is an ordered sequence of variables in $\boldsymbol{V}$ starting with X_i and ending at X_m such that each variable in the sequence is a direct cause of the next variable in the sequence, provided that m is greater than two. Also, X_i is said to be a *common cause* of X_m and X_n in $\boldsymbol{V}$ if X_i is a direct or indirect cause of both X_m and X_n.

[3] We may intuitively think of event x as a value of (random) variable X.

Following Spirtes et al. (1993:45), we define a *causal structure* over variables ***V*** as an ordered pair $\langle \mathbf{V},\mathbf{E} \rangle$, where ***E*** is a set of ordered pairs of ***V*** such that $\langle X,Y \rangle$ is in ***E*** if and only if *X* is a direct cause of *Y* with respect to ***V***. The variables in ***V***, which have no direct cause in ***V***, are called *exogenous*, and the rest are called *endogenous*. A structure $\langle \mathbf{V},\mathbf{E} \rangle$ is *deterministic* if the value of each endogenous variable in ***V*** is uniquely determined by its direct causes in ***V***. A structure that is not deterministic but forms part of a deterministic structure is called *pseudo-deterministic*. Each variable X_i in a pseudo-deterministic structure $\langle \mathbf{V},\mathbf{E} \rangle$ is a deterministic function of its direct causes in ***V*** and a disturbance term ε_i, which represents the net effects of variables outside ***V*** on X_i. A particular causal structure, called a *causally sufficient* structure, plays a special role in the GT literature:

> **Causal Sufficiency**: A set of variables ***V*** is called causally sufficient for a population if and only if in the population every common cause of any of two or more variables in ***V*** is in ***V***, or has the same value for all units in the population (Spirtes et al. 1993:45).

An (extra) assumption in the GT literature is that the disturbance terms associated with the variables in a causally sufficient structure are independently distributed. For the time being, when we refer to a causally sufficient structure, we also assume the independence of the errors. Finally, in a pseudo-deterministic structure, as soon as the functions linking the endogenous variables to the exogenous variables are defined, specification of a joint probability distribution for the disturbance terms generates a unique probability distribution for ***V***. With this remark, a causal model can be defined as:

> **Causal Model**: Let *S* be a causal structure defined over variables ***V***, *F* a distribution family over ***V***, and Θ a parameter space compatible with *S*. The triple $M = \langle S, F, \Theta \rangle$ is said to be a *causal model*. Each particular parameterisation of *M* forms a causal hypothesis.

2.2 Path Models

The causal structure is unknown but the presumption is that if enough data become available, the joint probability distribution of the variables under study can be estimated. The issue of data-driven causal inference involves using the estimate to learn about the structure. To explain how this problem is solved by the GT approach, it is useful to begin with a description of the more familiar field of path analysis, which also addresses a similar inference issue. In a nutshell, path analysis starts with a conjecture about the causal structure of the variables under study, translates the structure into a system of equations, introduces certain causal principles to derive the implications of the model, and tests them against the data.[4] As an illustration, consider variables $\mathbf{V} = \{X_1, ..., X_5\}$. Model I describes a possible structure that can be true of these variables:

$$\begin{aligned} &X_1 \\ &X_2 = \alpha X_1 + \varepsilon_2 \\ &X_3 = \beta X_1 + \varepsilon_3 \\ &X_4 = \gamma X_2 + \phi X_3 + \varepsilon_4 \\ &X_5 = \varphi X_4 + \varepsilon_5 \end{aligned} \qquad \text{Model I}$$

where the term ε_i in each equation represents the effect of unrecorded variables on X_i. According to this model, X_1 is a direct cause of X_2 and X_3 but an indirect cause of X_4 and X_5. X_2 and X_3 are direct causes of X_4, and X_4 is a direct cause of X_5. Since there is no reciprocal causal influence among the variables, the model is called a *recursive model.*

[4] Path analysis was developed by Sewell Wright (1934) and advanced by others including Simon (1954) and Blalock (1972). Blalock (1964) gives an introduction to the field. Irzik (1987) contains a philosophically oriented discussion of path analysis.

To estimate the model, path analysis assumes that (i) the disturbance term ε_i in each equation is uncorrelated with the exogenous variables in the equation; (ii) the disturbance (error) terms across the equations are uncorrelated; (iii) the errors are normally distributed with mean zero; and (iv) the endogenous variable in each equation *linearly* depends on the exogenous variables in the equation. A recursive model satisfying these conditions is called a path model. In addition, path analysis assumes that (v) the existence of a direct causal connection between two variables appears as a non-zero coefficient and (vi) that the absence of a direct causal connection always appears as a zero coefficient (Goldberger, 1971:35). These assumptions lead to two principles that allow deriving the implications of a path model for the data (see Appendix *A* for a proof):

(i) **The Screening off Principle**: If in a path model *X* cause *Z* only through the mediate of a set of variables ***Y***, then *X* and *Z* are statistically independent conditional on ***Y***. In short, direct causes screen off their remote causes. Given the linearity assumption, this means that the partial correlation $\rho_{XZ.\mathbf{Y}}$ is zero.

(ii) **The Common Cause Principle**: If in a path model *Z* is a common cause of *X* and *Y* and neither *X* is a cause of *Y* nor *Y* is a cause of *X*, then $\rho_{XY.Z} = 0$.

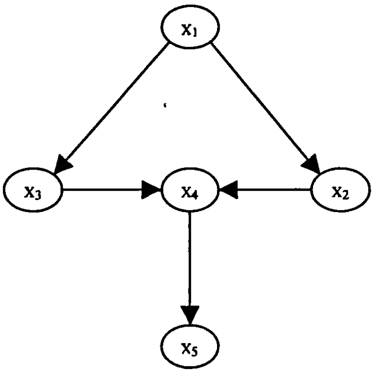
Assuming that Model I satisfies the conditions listed above, the model entails the following zero partial correlations:

$$\rho_{X_2X_3.X_1} = 0;\ \rho_{X_4X_1.X_2X_3} = 0;\ \rho_{X_5X_2.X_4} = 0;\ \rho_{X_5X_3.X_4} = 0;\ \rho_{X_5X_1.X_4} = 0$$

The practice in path analysis is to derive these zero partial correlations and test them against the data. If the vanishing partials are approximately zero in the data, the data is said to confirm the model and if they are significantly different from zero, the model is considered as incompatible with the data. Path analysis solves the causal inference problem by finding a model whose vanishing partials are all consistent with the data. A limitation of this approach is that conflicting causal models can imply the same vanishing partials, which makes it is impossible to infer the true model by testing its zero restrictions. Path analysis can at best eliminate the models whose zero restrictions are not found in the data. But it cannot establish the model that has actually generated the data.

2.3 Graphical Representation

The GT approach builds on the tradition of path analysis and seeks to confront its limitation squarely. To this end, it replaces the language of equations with the language of graphs to represent causal structures. A graph consists of two parts - a set of variables (*vertices* or *nodes*) ***V*** and a set of *edges* (or *links*) ***E***. Each edge in ***E*** is between two distinct variables in ***V***. There are two kinds of edges in ***E***, directed edges $X \rightarrow Y$ and bi-directed edges $X \leftrightarrow Y$. In either case, *X* and *Y* are called the *endpoints* of the edge and when there is an edge between *X* and *Y*, *X* and *Y* are said to be *adjacent*. If there is an edge between *X* and *Y* and towards *Y*, *X* is called a *parent* of *Y* and *Y* a *child* of *X*. A directed edge between *X* and *Y* (i.e., $X \rightarrow Y$) in graph *G* stands for the claim that *X* is a direct cause of *Y* relative to *G*. Absence of an edge conveys the claim that neither *X* causes *Y* nor *Y* causes *X*. The error terms are not usually represented in a graph. Thus, the structure implied by Model I can be expressed as the following graph:



Graph 2.1

This graphs depicts a *directed acyclic graph* (DAG). It is *directed* because the arrows lead from one variable into another and *acyclic* because one cannot return to any of the variables by following the arrows leading away from it. A sequence of consecutive edges in a directed graph *G* is called a *path.* A *directed* path *P* from *X* to *Y* is a sequence of vertices starting with *X* and ending with *Y* such that for every pair of variables *A* and *B* that are adjacent in the sequence in that order, the edge $A \rightarrow B$ occurs in *G*, and no vertex occurs more than once in *P*. Likewise, an *undirected* path *U* from *X* to *Y* is a sequence of variables starting with *X* and ending with *Y* such that for every pair of variables *A* and *B* that are adjacent in the sequence, *A* and *B* are adjacent in *G*, and no vertex occurs more than once in *U*. *Y* is a *collider* on an undirected path *U* if and only if there exist edges $X \rightarrow Y$ and $Z \rightarrow Y$ in *U*. And *Y* is an *unshielded collider* on *U* if and only if there exist edges $X \rightarrow Y$ and $Z \rightarrow Y$ in *U* and, in addition, *Z* and *X* are not adjacent in *G*. When there is a directed acyclic path from *X* to *Y* or $X = Y$, then *X* is said to be an *ancestor* of *Y*, and *Y* a *descendant* of *X*.

A DAG is another way of representing a recursive causally sufficient structure. If the possibility of feedback is ruled out, the class of DAGs that can be built from a set of variables *V* constitutes the class of all causal models that can be true of *V*. For now, we restrict our analysis to recursive causal structures, and denote the class of DAGs that can be built from *V* by Ω.

2.4 Conditional Independence Data

The GT approach takes the independencies true in the joint distribution of variables *V*, i.e., $P(\mathrm{V})$, as the evidence for making inference about the causal structure true of *V*. Let *X* and *Y* be two variables in *V*. *X* and *Y* are said to be independent if the joint probability density $P(x, y)$ is the product of the marginal density $P(x)$ and the marginal density $P(y)$, for all values x and y such that $P(y)$ is greater than zero. Following Dawid (1979), the independence of *X* and *Y* is shown by $X \perp Y$. That is,

$$X \perp Y \quad \text{if and only if} \quad P(x/y) = P(x) \quad \text{whenever } P(y) > 0.$$

Similarly, X and Y are said to be independent conditional on Z if $P(x/y,z)$ equals the product of $P(x/z)$ and $P(y/z)$, for all values x, y, and z such that $P(y,z)$ is greater than zero. That is,

$$X \perp Y / Z \quad \text{if and only if } P(x/y,z) = P(x/z) \quad \text{whenever} \quad P(y,z) > 0.$$

These definitions are generalized to disjoint sets of variables.[5] The conditional independence relation possesses several basic properties that allow deriving new independencies from an existing set of independencies. Some of these properties, studied by Dawid (1979), are listed in Appendix *B*. In what follows, the set of independencies true in the joint distribution $P(\mathbf{V})$ over variables V is denoted by Ind_P.

2.5 Assumptions Relating Probability to Causal Relations

The GT approach introduces two principles to link independence data to a causal structure (DAG). The first principle is the causal Markov Condition, which generalizes the principles underlying path analysis. Informally, the condition says that, in a recursive causal structure, every variable, conditional on its direct causes, is independent of all other variables in the structure except its effects. Formally, it says:

Markov Condition: A DAG *G* over a set of variables ***V*** and a probability distribution $P(\mathbf{V})$ satisfy the Markov condition if and only if for every *X* in ***V*** and every set ***Z*** of variables in ***V*** such that no member of ***Z*** is a descendent nor a parent of *X*, *X* and ***Z*** are independent conditional on the parents of *X* (SGS, 1993:35).[6]

[5] See Pearl (1988:82-83).

[6] Glymour (1997a:203-6) explains how traditional approaches to causal inference rely on variants of the Markov Condition. Hans Reichenbach (1956) was the first philosopher to discuss the Markov properties of causal systems, but variants of the principle have been discussed by Cartwright (1989), Salmon (1984), Skyrms (1980) and Suppes (1970).

The Markov condition characterises how a DAG represents independence relations. It says a variable X in DAG G, conditional on the state of its parents, is independent of all its non-descendants in G. Applying the condition to graph 2.1 yields the following independencies:

$$X_2 \perp X_3 / X_1,$$

$$X_4 \perp X_1 /(X_2, X_3),$$

$$X_5 \perp (X_1, X_2, X_3) / X_4.$$

These independencies entail additional independencies that are not immediately obtained by applying the Markov condition to graph 2.1. An example is $X_5 \perp X_3 / \{X_2, X_4\}$.[7] Pearl et al. (1988) have introduced a graph theoretic criterion, called *d-separation,* which allows reading from a DAG the entire list of independencies entailed by applying the Markov condition to the DAG. The criterion reads as follows:

Definition: Let X and Y be two variables among the vertices in graph G, and $\mathbf{Z}$ a subset of the vertices in G. A path p is said to be *d*-separated (or blocked) by $\mathbf{Z}$ if and only if (i) it contains a chain $X \rightarrow W \rightarrow Y$ or a fork $X \leftarrow W \rightarrow Y$ such that the middle variable W is in $\mathbf{Z}$, or (ii) it contains an unshielded collider $X \rightarrow W \leftarrow Y$ such that neither the middle variable W nor any of its descendants in G are in $\mathbf{Z}$. $\mathbf{Z}$ is then said to *d*-separate X from Y if and only if $\mathbf{Z}$ blocks every path from X to Y. (Pearl, 1998:238).

Geiger et al. (1990) have shown that there is a one-to-one correspondence between the independence relations entailed by applying the Markov condition to a DAG G and the triples $(X, \mathbf{Z}, Y)$ that satisfy the *d*-separation criterion in G. In graph 2.1, X_2 and X_3 are *d*-separated by X_1. However, X_2 and X_3 are not *d*-separated by X_4, since X_4 is an unshielded collider on the path $X_2 \rightarrow X_4 \leftarrow X_3$. Nor are X_2 and X_3

[7] This follows by first applying the weak union and then decomposition properties of independence relations to $X_5 \perp (X_1, X_2, X_3) / X_4$. See Appendix *B*.

d-separated by $\{X_1, X_5\}$, since X_5 is a descendant of X_4, which is an unshielded collider. Applying the *d*-separation criterion to every DAG *G* in Ω yields all the independencies implied by *G*. We use Ind_G to denote the set of independencies implied by DAG *G* over *V* to distinguish it from Ind_P that denotes the independencies true in $P(\mathbf{V})$.

Using the *d*-separation criterion, appendix *C* shows that the Markov condition applied to a DAG over variables $\mathbf{V} = \{X_1, ..., X_n\}$ implies the following variant of the common cause principle: If X_i and X_j are correlated, and neither X_i is a cause of X_j nor X_j is a cause of X_i, there are common causes of X_i and X_j in *V* conditional on which X_i and X_j are independent. The Markov condition therefore implies that every correlation among a recursive causally sufficient set of variables with independent errors has a causal explanation. The GT theorists generalises this implication to every correlation in the world by making a metaphysical assumption that can be called the *completeness* hypothesis:

The Completeness Hypothesis: For every set of recorded variable ***O***, either the set forms a causally sufficient set with uncorrelated errors or it can be embedded in a larger set of variables ***V*** that is causally sufficient with uncorrelated errors (Scheines, 1997:197; SGS, 1993:51)

Joined with this hypothesis, the Markov condition entails that every probabilistic dependency in the world reflects either a direct causal connection or the presence of latent common causes.[8]

[8] To deal with feedback systems, the GT theorists have recently introduced the so-called Global Markov Condition, which reads as follows: for a directed (cyclic or acyclic) graph *G* over vertices *V* and a probability distribution *P* over *V*, the distribution satisfies the global Markov condition if and only if for any three disjoint sets of ***X***, ***Y***, and ***Z*** in ***V*** if ***X*** is *d*-separated from ***Y*** given ***Z*** in *G*, then, ***X*** is independent of ***Y*** given ***Z*** in *P* (Koster, 1999). Joined with the completeness hypothesis, this implies that every correlation has a causal explanation.

The second principle in the GT approach about the connection between probability and causation is the so-called *Faithfulness* condition. The condition says that every conditional and unconditional independency true in the joint distribution of a set of observables represents absence of direct causal connection. Stated in graph-theoretic terms it says:

Faithfulness condition: Let G be a causal graph over variables V and $P(\mathbf{V})$ a probability distribution generated by G. $\langle G, P \rangle$ satisfies the Faithfulness Condition if and only if every conditional independence relation true in $P(\mathbf{V})$ is entailed by the Causal Markov Condition applied to G (SGS, 1993:56).

Faithfulness excludes all the independencies that are not implied by the topology of a DAG. For a possible case of such independencies, consider the graph below, which describes a conjecture about the relations among minimum wage, the economy, and individual income.

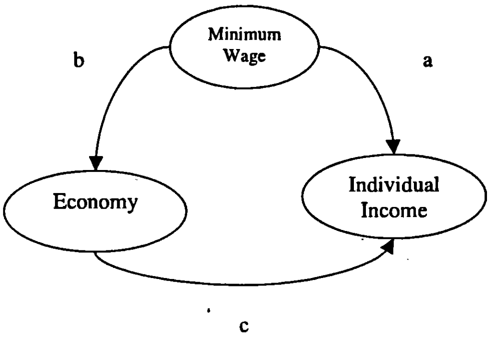


Figure 2.2

Suppose the effect of minimum wage through the economy on individual income were such that it exactly cancelled out its direct effect on individual income, i.e., $a = -(bc)$. In that case, the structure would generate an independency that does not follow from applying the Markov condition to it. If the world contained such structures, it would be wrong to infer absence of causation from independence data. In the current case, one would wrongly conclude that minimum wage does not affect income, even though it does. Faithfulness excludes such structures from the world. It implies that all independencies are structural independencies, following from the topology of the true graph. Appendix *D* shows how Faithfulness underlies other methods of causal inference.

3 Causal Inference

Causal inference in the GT approach proceeds by (i) estimating the joint probability distribution of the variables *V*, (ii) deriving the independencies true in $P(\mathbf{V})$, and (iii) constructing a graph (or graphs) that, given the Markov condition and Faithfulness, is (are) consistent with the independencies. The concern, here, is with the final stage, which has to do with the move from the independencies true in $P(\mathbf{V})$ to a graph that could have generated the distribution. This section describes this stage of inference in some detail so as to prepare the ground for a critical appraisal of the GT approach in Section V.

3.1 Inference with Causal Sufficiency

We begin our exposition by assuming that the variables under study are causally sufficient, and then describe graph-theoretic causal inference in general. To be specific, we work with variables $\mathbf{V} = \{X_1,...,X_5\}$, assuming that *V* is causally sufficient. And, we hypothesise that

$$
\begin{aligned}
&Ind_P = \{X_2 \perp X_3 / X_1, \\
&X_4 \perp X_1 /(X_2, X_3), \\
&X_5 \perp (X_1, X_2, X_3)/ X_4\}.
\end{aligned}
$$

Causal sufficiency implies that the set of DAGs, Ω, that can possibly be true of variables *V* is finite. Thus, the solution to the causal inference problem involves finding a DAG G from Ω that is consistent with the independencies in Ind_P. To explain how such a DAG can be found, note that, given the Markov condition, if a DAG G generated the data, G would not imply any independency that is not in Ind_P. As a result, for any DAG G in Ω, if Ind_G contains an independency that is not in Ind_P, the DAG does not satisfy the Markov condition. The Markov condition therefore excludes all those DAGs in Ω that entail at least one independency that is not in Ind_P. On the other hand, according to the Faithfulness condition, the distribution $P(\mathbf{V})$ is faithful to a DAG G in Ω if every independency in Ind_P follows from the *d*-separation criterion applied to G. This means that if a DAG G in Ω fails to imply *all* the independencies in Ind_P, the DAG is not faithful to $P(\mathbf{V})$. Faithfulness, therefore, excludes all those DAGs in Ω that fail to imply all the independencies in Ind_P. These conditions altogether imply that a DAG G in Ω with independencies Ind_G is consistent with the independencies in Ind_P if and only if there is a one-to-one correspondence between the independencies in Ind_P and those in Ind_G. The inference problem can then be solved by deriving the set of independencies Ind_G implied by each DAG G in Ω and investigating whether they have a one-to-one correspondence with the independencies in Ind_P.

When causal sufficiency is assumed, the above description gives all that there is in the *GT* approach to causal inference. Yet, the above implications of the Markov condition and Faithfulness lead to a basic theorem that simplifies the procedure for constructing a DAG consistent with the independencies in Ind_P. The theorem, proved by Verma and Pearl (1990), says:

Theorem: Distribution $P(\mathbf{V})$ satisfies the Markov and Faithfulness conditions for DAG G if and only if (i) any two vertices X and Y are adjacent in G if and only if they are statistically dependent conditional on every subset of vertices in G not containing them. (ii) $X \rightarrow Y \leftarrow Z$ is an unshielded collider in G, then X, Z are not independent conditional on Y.

The procedure begins with a complete *skeleton*; that is, a graph in which every variable is connected by an undirected edge to every other variable. It tests every pair of variables X and Y, and removes the edge between the variables if $X \perp Y$ is in Ind_P. Next, for every pair of variables X and Y, it tests whether there is a subset Z of variables that does not contain X and Y but renders X and Y independent. If so, the edge between X and Y is removed. The process creates an undirected graph from which some of the edges are removed. In our example, the process results in the graph in Figure 3.1.

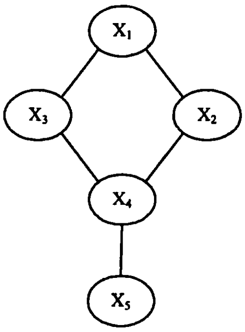


Figure 3.1

In the second phase, the procedure considers every triple of vertices X, Y, and Z in V. If there is an edge between X and Y, and an edge between Z and Y, but no edge between X and Z, and X and Z are not independent given Y, the edges are directed towards Y. In Figure 3.1, there is an edge between X_2 and X_4, an edge between X_3 and X_4, but no edge between X_2 and X_3. The edges are thus directed towards X_4. Otherwise, the resulting DAG will not entail $X_4 \perp X_1/(X_2, X_3)$, which violates Faithfulness. Similarly, the edge between X_4 and X_5 is directed towards X_5 to avoid violating Faithfulness. The edges between X_1 and X_2, and X_1 and X_3 cannot *both* be directed towards X_1, since such an orientation makes X_2 and X_3 dependent conditional on X_1, which contradicts Faithfulness. The independencies Ind_P impose no further restrictions on the edges. The graph in Figure 2.1 is consistent with the independencies in Ind_P.

3.2 Inference without Causal Sufficiency

Causal sufficiency is hardly true, and even if true, it would not be known in advance. To claim any success, a data-driven method of causal inference should deal with the causal inference problem regardless of whether the recorded variables are causally sufficient or not. Without causal sufficiency, a correlation between measured variables *X* and *Y* no longer implies that either *X* causes *Y* or *Y* causes *X*. The correlation might be due to latent common causes. Thus, the general question will be to determine when and how by analysis of a set of measured variables containing *X* and *Y* it is possible to conclude that *X* causes *Y*, or *X* does not cause *Y*, or the correlation between *X* and *Y* is due to latent common causes.

In the GT approach, the burden of generalising the solution to the inference problem under causal sufficiency to cases where the truth of the condition is not known is on the completeness hypothesis (Scheines, 1997:197). According to the hypothesis, for every set of *measured* variables $\boldsymbol{O}$, which is not causally sufficient, there is in reality a DAG $G(\mathbf{O},\mathbf{L})$ with independent errors that is responsible for the dependencies among the observed variables $\boldsymbol{O}$, with $\boldsymbol{L} = \boldsymbol{V} \backslash \boldsymbol{O}$ standing for the latent common causes of $\boldsymbol{O}$. Thus, the joint probability distribution of the recorded variables $P(\mathbf{O})$ is regarded as the marginal of an un-estimated distribution $P^*(\mathbf{V})$ that satisfies both the Markov condition and Faithfulness. From this perspective, the general problem of causal inference involves learning about the true DAG $G(\mathbf{O},\mathbf{L})$ from the marginal distribution $P(\mathbf{O})$.

With causal sufficiency, the object of inference is a DAG in which every adjacency between *X* and *Y* is represented by an arrow, meaning that either *X* causes *Y* or *Y* causes *X*. As causal sufficiency is withdrawn, one needs a different graphical object to state that an adjacency is due to latent common causes. The literature provides several objects suitable for representing latent common causes. We use the so-called *hybrid graph*, which is a graph that, in addition to one-directional edges $\rightarrow$, contains bi-directional edges $\leftrightarrow$ to represent latent common causes.[9] To illustrate the simplest hybrid graph, let $X \leftarrow Z \rightarrow Y$ be the DAG true of variables *X*, *Y*, and *Z*. When *Z* is unknown, the hybrid graph for this DAG is given by $X \leftrightarrow Y$; the bi-directed link represents the latent common cause *Z*.

Learning about the true DAG $G(\mathbf{O},\mathbf{L})$ from the independencies true in $P(\mathbf{O})$ requires knowing what independencies, given the Markov condition, would occur among the recorded variables $\boldsymbol{O}$ if $G(\mathbf{O},\mathbf{L})$ were the DAG generating the data. An answer to this question is given in Pearl and Verma (1991). To explicate the answer, we need to introduce a further graph-theoretic notion – an *inducing path.* An undirected path *U* between *X* and *Y* is an inducing path over $\boldsymbol{O}$ in $G(\mathbf{O},\mathbf{L})$ if and only if (i) every member of $\boldsymbol{O}$ on *U* (except the endpoints) is a collider on *U*, and (ii) from every collider on *U* there is a directed path to *X* or *Y*. Figure 3.2 shows an inducing path between *X* and *Y* over $\mathbf{O} = \{X,Z,Y\}$:

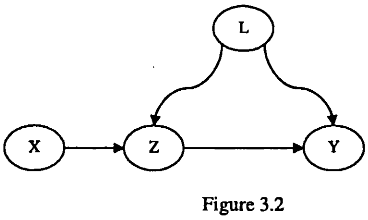


Figure 3.2

Pearl and Verma (1991) have shown that there is an inducing path between recorded variables *X* and *Y* in $G(\mathbf{O},\mathbf{L})$ over $\boldsymbol{O}$ if and only if *X* and *Y* are not independent conditional on any subset of $\mathbf{O}\setminus\{X,Y\}$. This means that if there is a directed path in the hybrid graph between *X* and *Y* that is into *Y*, then *X* is a (possibly indirect) cause of *Y*. If the path is into *X*, then *Y* is a (possibly indirect) cause of *X*. And, if the path is

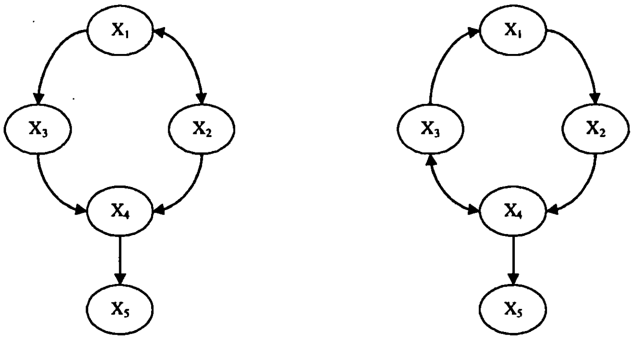
[9] The term 'hybrid graph' has been borrowed from Pearl and Verma (1991). They, however, define a hybrid graph slightly differently; they identify it with a graph in which links may be undirected, uni-directed or bi-directed.

both into X and into Y, then there is a common cause (or causes) in $G(\mathbf{O},\mathbf{L})$ affecting both X and Y. Thus, given the Markov and Faithfulness conditions, one can learn about the true structure $G(\mathbf{O},\mathbf{L})$ by investigating the hybrid graph consistent with the independence data.

The intuitions behind these results can be explained by analysing some simple examples, which will be used later to assess the GT approach. As a first example, we withdraw the causal sufficiency assumption about the variables $\{X_1,...,X_5\}$ studied earlier while retaining the same set of independence relations:

$$\begin{aligned}&Ind_P =\\ &\{X_2 \perp X_3 / X_1,\\ &X_4 \perp X_1 /(X_2,X_3),\\ &X_5 \perp (X_1,X_2,X_3)/X_4\}.\end{aligned}$$

Starting from a skeleton over $\boldsymbol{O}$, these independencies lead to the same graph as the one in Figure 3.1. Faithfulness requires directing the edges between X_2 and X_4, and X_3 and X_4 towards X_4, and the edge between X_4 and X_5 towards X_5. No DAG $G(\mathbf{O},\mathbf{L})$, containing variables d-separating X_4 and X_5, can be true of the data. Any such DAG fails to entail $X_5 \perp X_2 / X_4$, and is unfaithful to $P(\mathbf{O})$. The true DAG $G(\mathbf{O},\mathbf{L})$ thus contains an inducing path between X_4 and X_5 that is into X_5, meaning that X_4 causes X_5. Moreover, since cycles have been ruled out, X_5 is not a cause of X_4. Also, no DAG that renders both dependencies between X_1 and X_2 and X_1 and X_3 spurious can be faithful to $P(\mathbf{O})$. In any such DAG, X_1 is a collider incapable of d-separating X_2 from X_3. Finally, only one of the edges in $X_2 \rightarrow X_4 \leftarrow X_3$ can be due to latent common causes. A DAG that renders both edges spurious fails to entail $X_5 \perp X_1 /(X_2,X_3)$. Figure 3.3 shows two hybrid graphs consistent with the independencies:

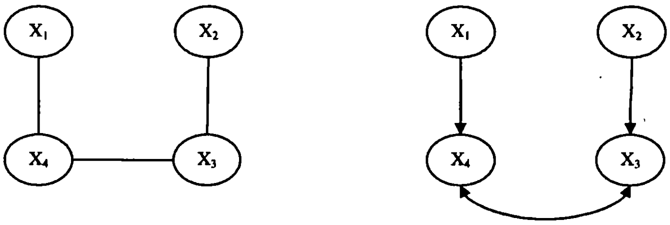


Graph (a) Figure 3.3 Graph (b)

This example shows how, given the Markov condition and Faithfulness, the GT theorists conclude whether a variable causes another variable. To set the stage for our discussion, we also describe an example from Glymour (1997:218) intended to illustrate a case where the conditions entail that an association is *definitely* due to latent common causes. Let $\mathbf{O} = \{X_1,...,X_4\}$ and

$$Ind_P = \\ \{X_1 \perp X_2, X_1 \perp X_3, X_2 \perp X_4\}^{\cdot}$$

Starting from a skeleton over ***O***, these independencies lead to undirected graph (a) in Figure 3.4. Faithfulness requires directing the edges between X_2 and X_3 and between X_4 and X_3 towards X_3. It also requires orienting the edges between X_1 and X_4 and between X_3 and X_4 towards X_4. These create a bi-directed edge between X_3 and X_4, as shown in graph (b) bellow. The bi-directed edge reveals an inducing path in the true *G*(**O**,**L**) that is into both X_3 and X_4, revealing the existence of a common cause for the variables.



Graph (a) Figure 3.4 Graph (b)

This conclusion is based on the consideration that any DAG $G(\mathbf{L},\mathbf{O})$ not containing some variables responsible for the correlation between X_3 and X_4 violates either the Markov condition or Faithfulness. Consider, for instance, a DAG $G(\mathbf{L},\mathbf{O})$ in which X_3 causes X_4. Such a DAG does not entail $X_2 \perp X_4$, which is in Ind_P, and hence violates Faithfulness. Since, by assumption, there is no feedback among the variables, the correlation between X_3 and X_4 must be due to latent common causes.

4 Intrinsic Limitations of Data-driven Causal Inference

Granting that a causal relation is not the same as a probabilistic relation, the Markov and Faithfulness conditions are the most general principles that can be true of the connection between causation and probability. This section continues to assume the universal validity of these principles to examine exactly what kind of conclusions they allow us to infer from data. This requires us studying the issue of statistical indistinguishability (equivalence) of causal models. An analysis of this issue is essential for understanding intrinsic limitations of data-driven causal inference. It will be seen that even if the generality of the principles is not challenged, extremely little can be inferred from data alone. This section proceeds by showing that, given any causal model fitting the data, there is usually a simple rule that can be used to generate a class of equivalent models. These models have very little or nothing in common, because the sign and significance of the coefficient estimates can vary from one model to another.

A notion of model equivalence is the so-called Markovian (or *d*-separation) model equivalence that reads as follows:

Markovian Model Equivalence: Let S_i be a causal structure defined on variables *V*, F_i a multivariate distribution family over *V*, and Θ_i a parameter space compatible with S_i. Two models $M_1 = \langle S_1, F_1, \Theta_1 \rangle$ and $M_2 = \langle S_2, F_2, \Theta_2 \rangle$ are Markovian equivalent if and only if they imply the same Markovian independencies; i.e., if and only if $Ind_{P1} = Ind_{p2}$.

Another stronger concept of model equivalence is the so-called *distributional model equivalence*:

Distributional Model Equivalence: Two models $M_1 = \langle S_1, F_1, \Theta_1 \rangle$ and $M_2 = \langle S_2, F_2, \Theta_2 \rangle$ are distributionally equivalent if and only if for every parameterisation of M_1 generating distribution f_1 there is a parameterisation of M_2 generating distribution f_2 such that f_1 and f_2 are the same.

These notions coincide in the case of causally sufficient recursive models (Pearl, 2000:146). Outside this category, there are Markovian equivalent models that are not distributionally equivalent (Sprites, et al., 1996a; Raykov et al., 1999). Since the aim here is to assess the claims of the GT approach, the discussion will be confined to Markovian model equivalence.[10] We first consider recursive causal models and then turn to non-recursive models.

4.1 Recursive Equivalent Models

Recursive causal models can be divided into causally sufficient and causally insufficient models. An original contribution to the study of statistical

[10] Appendix G shows how in general it is possible to check if two models are distributionally equivalent.

indistinguishability of causally sufficient recursive models (DAGs) is Stelzl (1986), who investigates statistical equivalence of path models. Other early contributions are Frydenberg (1990), Lee and Hershberger (1990), and Verma and Pearl (1990).[11] In path analysis, data are characterised by sample covariance matrices and, as seen, the implications of a model is defined in terms of its zero partial correlations. A path model is compatible with the data if its vanishing partials are compatible with the sample covariance matrix. So, if path models M_1 and M_2 entail the same vanishing partials, and if M_1 is compatible with the data, then M_2 is also compatible with the data and vice versa. On the other hand, if either M_1 or M_2 entails a zero partial correlation that is not implied by the other, the models are not equivalent. This suggests the following definition of path model equivalence:

Path Model Equivalence: Two paths models M_1 and M_2 are equivalent if and only if they constrain the same set of partial correlations to zero.

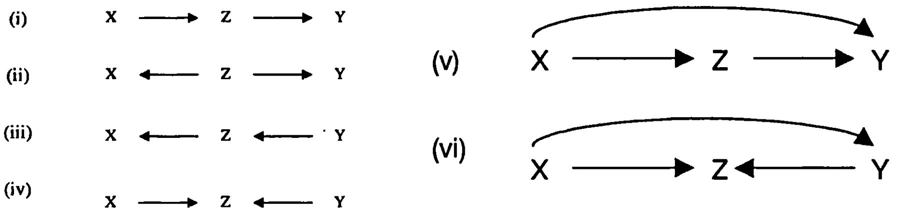


Figure 4.1

Stelzl (1986) noted that the zero partial correlations implied by a path model were invariant with respect to certain changes in the ordering of the variables in the model. He located several invariant properties of vanishing partials and used them to define four rules for transforming a path model into another statistically equivalent model. The invariant properties underpinning Stelzl's rules can be reduced to two very

[11] Further contributions include: Bollen (1989), Breckler (1990), Hershberger, (1994), Jörgeskog and Sörbom, (1993), Luijben (1991), MacCallum, Wegener, Uchino, Fabrigar, (1993), and Raykov, (1997, 1999, 2001).

simple properties. Consider a path model over variables $\{X,Y,Z\}$, with path diagram (i) in Figure 4.1.

Graph (i) implies $\rho_{XY.Z} = 0$ but no other zero restriction. Inverting arrow $X \rightarrow Z$ or both arrows yields graph (ii) or (iii), which have the same zero restrictions as (i). However, inverting arrow $Z \rightarrow Y$ in (i) or $X \rightarrow Z$ in (iii) creates unshielded collider (iv), which does not imply $\rho_{XY.Z} = 0$; the only zero restriction it implies is $\rho_{XY} = 0$. Similarly, consider inverting one or both of the arrows in (iv). This yields one of the models (i) through (iii) that imply a zero restriction not implied by (iv) and does not entail the zero restriction implied by (iv). This suggests that any arrow inversion in a path diagram that creates or destroys an unshielded collider destroys or creates a zero restriction, yielding a statistically different path model.

Now, consider graph (v), which is a complete graph in the sense that there is a link between every two variables in it. A complete graph implies no zero partial correlation (Wermuth, 1980). So, any change in the graph that turns it into another (non-cyclic) complete graph yields an equivalent path model. Redirecting arrow $Z \rightarrow Y$, for instance, gives rise to graph (vi) which is equivalent to graph (v). Removing an arrow from these two models, however, yields a model with a zero restriction not implied by the original model.

Altogether, these analyses point to two types of changes in a path diagram that alter its zero restrictions: (i) deletion or creation of a new link and (ii) creation or destruction of an unshielded collider. In general, Verma and Pearl (1990) and Frydenberg (1990) show that:

Theorem 4.1: Two DAGs G and G^* are Markovian (covariance) equivalent if and only if they (i) they have the same links and (ii) the same unshielded colliders.[12]

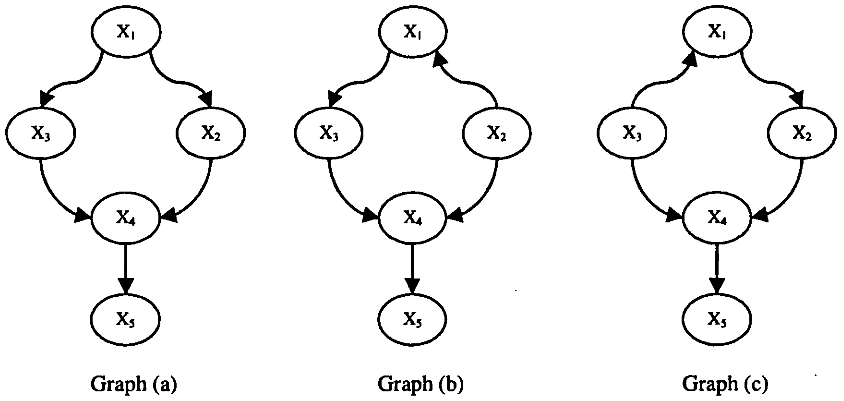
[12] This theorem also follows from Proposition I in Raykov et al. (1999:206).

In light of this, an edge $X \rightarrow Y$ in a DAG G can be inverted to form an equivalent DAG G^* as long as the inversion neither destroys nor creates an unshielded collider. This happens only when every parent of X is a parent of Y and every parent of Y (except X) is a parent of X (Chickering, 1995; and Meek, 1995). The result gives rise to the following rule for converting a DAG G into anpther equivalent DAG G^* (Appendix E outlines a proof):

The DAG Inversion Rule: An arrow $X \rightarrow Y$ in a DAG G can be inverted to form an equivalent DAG G^* only if every parent of X is a parent of Y and every parent of Y (except X) is a parent of X.

Since equivalence relation is reflexive, symmetric, and transitive, by repeatedly applying the rule one can generate all possible models equivalent to a DAG. Applying the rule to the path model described in Section II yields two more equivalent models. The original model corresponds to graph (a) below, with the zero partial correlations:

$$\rho_{X_2X_3.X_1}=0;\ \rho_{X_4X_1.X_2X_3}=0;\ \rho_{X_5X_2.X_4}=0;\ \rho_{X_5X_3.X_4}=0;\ \rho_{X_5X_1.X_4}=0.$$



Graph (a) Graph (b) Graph (c)

Figure 4.2

Assuming causal sufficiency, no other arrow in DAG (a) can be inverted. Take arrow $X_4 \rightarrow X_5$. X_4 has two parents X_2 and X_3 which are not parents of X_5. Inverting this arrow creates new unshielded colliders, which destroy zero restrictions $\rho_{X_5X_2.X_4} = 0$, $\rho_{X_5X_3.X_4} = 0$, and $\rho_{X_5X_1.X_4} = 0$.

The above analysis shows that once a DAG is fitted to the data, there is a simple rule to transform it into another statistically equivalent DAG. Even when the causal sufficiency assumption is taken for granted, the 'true' structure cannot then be discovered from the independence data alone. If causal sufficinency is not assumed, a graph (model) over measured variables $\boldsymbol{O}$ can be changed into another equivalent graph not only by inverting some of the directed edges but also by replacing them with bi-directed edges ↔ representing latent common causes. In discussing Morkovian equivalence of causally insufficient models, we continue to assume the completeness hypothese. The equivalence of two DAGs over observed variables $\boldsymbol{O}$ can then be defined as follows:

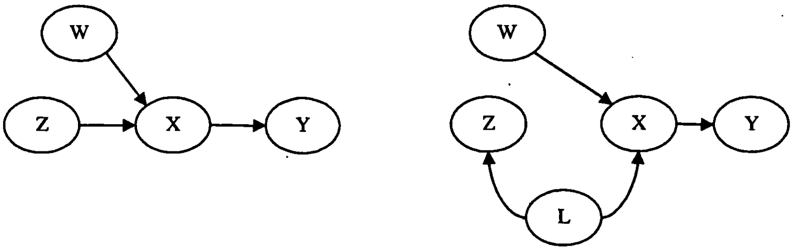
Markovian Equivalence Over O: Two DAGs $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ are Markovian equivalent over $\boldsymbol{O}$ if they imply the same set of *d*-seperation triples over $\boldsymbol{O}$.[13]

Building on Stelzl's (1986), Lee et al. (1990) establish a simple condition for replacing an arrow $X \rightarrow Y$ in the graph of a covariance structural model with a bi-directed edge $X \leftrightarrow Y$ that suggests the correlation between the variables is due to correlated errors. By completeness, correlation among errors indicates latent common causes. Lee et al.'s condition can therefore be viewed as a condition for converting a hybrid graph into an equivalent hybrid graph. Theorem 4.2 restates Lee et al.'s result, and Appendix *F* provides a proof for it.

[13] A necessary and sufficient criterion for testing the *d*-separation equivalence of two semi-Markovian models is given in Spirtes and Verma (1992).

Theorem 4.2: Let $G(\mathbf{O},\mathbf{L})$ be a DAG, *X* and *Y* in *O*, and $X \rightarrow Y$ hold in $G(\mathbf{O},\mathbf{L})$. Let $G^*(\mathbf{O},\mathbf{L}^*)$ be the same as $G(\mathbf{O},\mathbf{L})$ except that $X \rightarrow Y$ is replaced with $X \leftrightarrow Y$. $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ are Markovian equivalent over *O* if for every variable *Z* in *O* that is a parent of *X* in G, *Z* is also a parent of *Y*. Also, if $X \leftrightarrow Y$ is in $G(\mathbf{O},\mathbf{L})$, the bi-directed edge can be replaced with $X \rightarrow Y$ if every parent of *X* is a parent of *Y*.

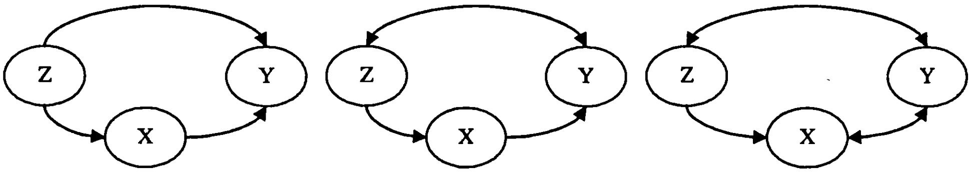
Pearl (2000:146) notes that, when the requirement of the DAG inversion rule holds of an arrow $X \rightarrow Y$ in a hybrid graph, replacing it with a bi-directed edge neither generates nor destroys an unshielded collider, and yields an equivalent hybrid graph. The rule, he argues, can be used to transform a hybrid graph into another equivalent hybrid graph. However, unlike the condition in Theorem 4.2, the DAG inversion rule requires every parent of *X* or *Y* (except *X*) to be a parent of both, which is unnecessarily strong. Consider graph (a) in Figure 4.3. Here, *X* has a direct cause, *W*, which is not a cause of *Z*. But replacing the arrow $Z \rightarrow X$ with a bi-directed edge neither destroys nor creates an independence relation among the recorded variables. Both graph (a) and (b) imply the same independencies over $\mathbf{O} = \{X,Y,Z,W\}$.



Graph (a) Graph (b)

Figure (4.3)

As required by theorem 4.2, in order to replace an arrow $X \rightarrow Y$ with a bi-directed arrow $X \leftrightarrow Y$ without destroying or creating an unshielded collider it is sufficient that every parent of *X* is a parent of *Y*. The theorem, however, does not exhaustively characterise the class of DAGs that are Markovian equivalent with $G(\mathbf{O},\mathbf{L})$ over *O*. This is because creation of a new unshielded collider in certain situations leaves the independencies implied by $G(\mathbf{O},\mathbf{L})$ over *O* unchanged. An example is given by graph (a) in Figure 4.4.[14] Theorem 4.2 permits replacing $Z \rightarrow Y$ with a bi-directed edge to create equivalent graph (b) but does not allow replacing $X \rightarrow Y$ in (b) with a bi-directed edge, since X has a parent that is no longer a parent of Y. Nevertheless, such replacement neither destroys nor creates an independency. Even though graph (c) contains an extra un-shielded collider, all the three graphs are Markovian equivalent over $\boldsymbol{O}$.



Graph (a) Graph (b) Graph (c)

Figure (4.4)

Due to such cases, establishing a rule that defines necessary and sufficient conditions for transforming a hybrid graph into another equivalent hybrid graph demands specifying the conditions under which creating a new unshielded collider does not alter the independencies. Pearl (2000:147) takes some steps towards this aim but acknowledges that his requirements are not sufficient. All the same, theorem 4.2 gives way to the following rule for creation of a partial set of equivalent hybrid acyclic graphs:

The Bi-directed Edge Replacement Rule: An arrow $X \rightarrow Y$ in a hybrid graph $G(\mathbf{O},\mathbf{L})$ can be replaced with a bi-directed edge $X \leftrightarrow Y$ to form an equivalent hybrid graph $G^*(\mathbf{O},\mathbf{L}^*)$ if the parents of Y in $G(\mathbf{O},\mathbf{L})$ include the parents of X. Conversely, under the same condition, a bi-directed edge $X \leftrightarrow Y$ can be replaced with a directed edge $X \rightarrow Y$.

Applying this rule to the example used throughout the chapter adds four more models to the catalogue of the equivalent models listed in Figure 4.2. The new models are

[14] A different example is found in Pearl, (2000: 147)

shown in Figure 4.5. Graph (b) is obtained by first applying the DAG inversion rule to arrow $X_1 \rightarrow X_2$ and then replacing it with a bi-directed edge. Similarly, graph (d) is obtained by first applying the DAG inversion rule to arrow $X_1 \rightarrow X_3$ and then replacing it with a bi-directed edge.

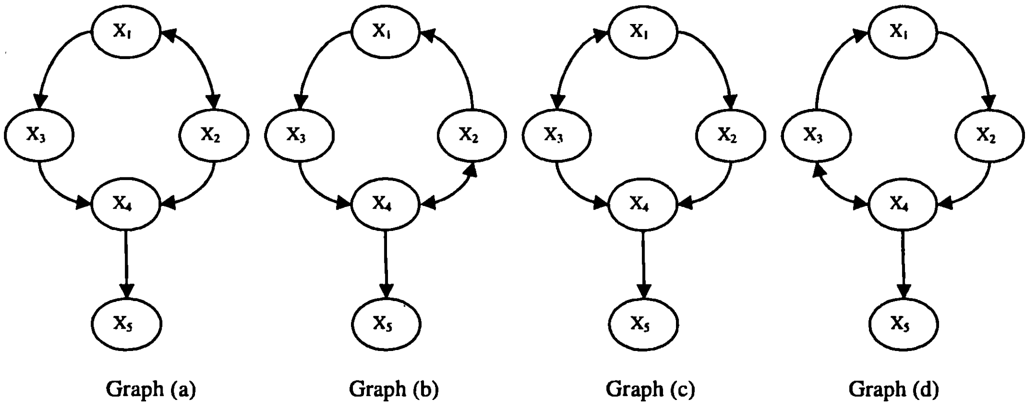


Figure (4.5)

The rule does not permit replacing arrow $X_4 \rightarrow X_5$ with a bi-directed edge, as X_4 has parents which are not parents of X_5. Given the Markov and Faithfulness conditions, the only conclusion that can be inferred from the (hypothetical) independence data is that X_4 is a (possibly indirect) cause of X_5, and X_5 has no causal influence over X_4.

4.2 Non-recursive Equivalent Models

Allowing feedback increases the complexity of causal modelling. Notably, the Markov condition, as defined earlier, does not hold of non-recursive (cyclic) models and must be replaced with a more general one.[15] In addition, feedback adds to the complexity of the conditions under which two cyclic models are *d*-separation

[15] Glymour (1997a: 208) describes a feedback model that does not satisfy the Markov condition.

equivalent. This in turn makes it even more difficult to characterise the necessary and sufficient conditions under which a cyclic model can be transformed into another equivalent model. To keep the discussion short, instead of considering the equivalence of cyclic models in general, building on the works of Frydenberg (1990), Lee et al. (1990), and Raykov et al. (1999), we discuss a specific class of non-recursive models, known as block recursive models, which has been of interest in econometrics (Kmenta, 1986). In graph theoretic terms, a block recursive equation system corresponds to a directed graph that can be partitioned into several subgraphs (blocks) such that there is no feedback across the blocks but the relations among the variables within each block can be either recursive or non-recursive. Graph (a) in Figure 4.6 represents a block recursive equation system. There is no feedback across the blocks separated by the line. If, in addition, the graph (equation system) contains an acyclic subgraph (block), the graph is said to be a *limited block recursive graph* (system) (Lee et al., 1990:317). Following Lee et al. (1990) we name an acyclic subgraph a *focal* subgraph. Theorem 4.3 captures the result available about the *d*-separation equivalence of limited block recursive models. Appendix *G* outlines a proof, based on a theorem due to Raykov et al. (1999):

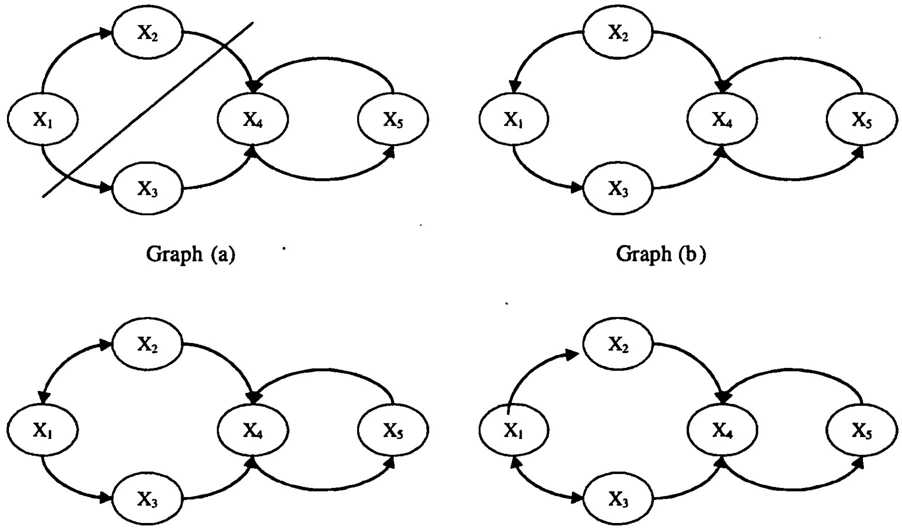
Theorem 4.3: Let $G^*(\mathbf{O},\mathbf{L}^*)$ be the same limited block recursive graph as $G(\mathbf{O},\mathbf{L})$ over $\boldsymbol{O}$ except that $X \leftrightarrow Y$ is in $G^*(\mathbf{O},\mathbf{L}^*)$ instead of $X \rightarrow Y$. Then, $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ are *d*-separation equivalent over $\boldsymbol{O}$ if for every variable Z in $\boldsymbol{O}$ that is a parent of X in $G(\mathbf{O},\mathbf{L})$, Z is also a parent of Y. Furthermore, if $X \leftrightarrow Y$ is in $G(\mathbf{O},\mathbf{L})$, the edge can be replaced with $X \rightarrow Y$ if every parent of X is a parent of Y.

This theorem makes it possible to establish a rule similar to the bi-directed edge replacement rule that allows transforming a limited block recursive graph into another equivalent graph.[16] Figure 4.6 depicts four equivalent models. The set $\{X_1, X_2\}$ forms a focal block. Using the theorem, the arrow $X_1 \rightarrow X_2$ can be redirected to obtain graph (b) or replaced with a bi-directed edge to obtain graph (c).[17] The set

[16] Richardson (1996) provides the necessary and sufficient conditions under which two non-recursive models, limited block-recursive or not, are *d*-separation equivalent.

[17] Graph (b) is obtained by first $X_1 \rightarrow X_2$ with $X_1 \leftrightarrow X_2$ and then replacing it with $X_1 \leftarrow X_2$.

$\{X_1, X_3\}$ also forms a focal block. The arrow $X_1 \rightarrow X_3$ can be replaced with a bi-directed edge to obtain graph (d).



Graph (a)

Graph (b)

Graph (c)

Figure 4.6

Graph (d)

Although the discussion of non-recursive models has been confined to limited block recursive models, the scope of the result is not that limited. It is usually possible to locate a focal block in most non-recursive models. Theorem 4.3 thus applies to most cyclic models.

The above rules permit generating a class of equivalent models for a large class of structural models. As stressed by the founders of the GT approach, the outcome of the GT algorithms is not therefore the true graph but a class of equivalent graphs that could have generated the data. More precisely, the outcome of the algorithms is a *pattern*; that is a graphical object that represents the directed edges common to all the members of the equivalent class but leaves the direction of other edges unspecified. These common edges define what can be learnt from the data using the GT techniques.

4.3 Causal Inference in Practice

A proposal to curb the multiplicity of equivalent models is to consider the temporal order of the variables. A cause is said to temporally precede the effect, which means if *X* precedes *Y*, *Y* cannot be a cause of *X*. Though this suggestion may be of some help, it falls short of narrowing the class of equivalent models to a single model. The suggestion does not apply to feedback models, and it is often difficult to ascertain whether a variable precedes another. In addition, even if the temporal order of the variables were known and only recursive models were permitted, there would still be many models fitting the data. As a simple example, suppose that $\mathbf{O} = \{X, Y, Z\}$ is the set of recorded variables, *X* temporally precedes *Y*, *Y* temporally precedes *Z* and that $X \perp Z / Y$ is true in $P(\mathbf{O})$. The only conclusion that can be derived from this information is that *Y* causes *Z*. Both graph (a) and (b) in Figure (4.7) are consistent with the data. In fact, *L* stands for all the temporally precedent variables that can affect both *X* and *Y*. This means infinitely many models could have generated the independence data.

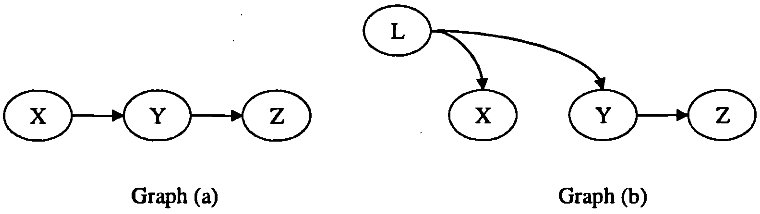
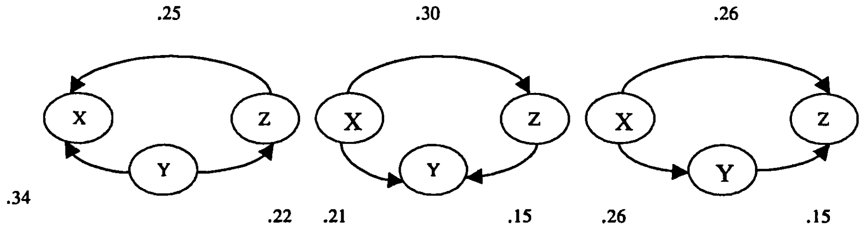


Figure 4.7

Hence, even with the imposition of temporal order, the class of equivalent models may be large. Now, a very important point, which often goes unnoticed, is that *in practice* the class of models (graphs) equivalent with a model fitting the data usually have little or even nothing in common. The reason is that the coefficient estimates do not remain invariant across various members of the equivalence class; they vary as we move from one member of the class to another. Consider the following covariance matrix:

	X	*Y*	*Z*
X	1	.26	.30
Y	.26	1	.22
Z	.30	.22	1

Figure 4.8 depicts three equivalent graphs consistent with this data.



Graph (a) Graph (b) Graph (c)

Figure 4.8

As these graphs illustrate, the parameter estimate for a link between two variables does not remain invariant across the members of the equivalent class. In general, a coefficient estimate may be significant in some members of the class but not in others. Or it may be positive in some members of the class but negative in others. More importantly, the change in the sign and significance of the coefficient estimates is not confined to the coefficients of the edges varying across the equivalent models. The sign and significance of the coefficients of the common edges can also vary from one model to another (Williams et al., 1996:286). In some members of the equivalent class the coefficient estimate associated with a common edge may be significant but not in others. Or in some models it may be positive but in others negative. Appendix *H* further demonstrates these points using two real examples from MacCallum, et al. (1993).

Since probabilities are unknown and one has to rely on their estimates, and since the coefficient estimates vary across equivalent models, in practice the members of an equivalent class usually have little in common. As a result, even by granting the Markov and Faithfulness conditions, little can be learnt from data alone. The claim of the GT approach that one can infer substantive causal conclusions by inspecting the edges common among equivalent models is contingent on the invariance of the coefficient estimates, which is not always the case. Substantive conclusions from data demands subject matter information to narrow down the class of equivalent models fitting the data. One, in particular, needs information on the sign and significance of the coefficients.

5 Assumptions Revisited

The claim that the GT approach can discover the class of equivalent causal models that could have generated the data depends on the universal validity of the Markov condition and Faithfulness. It is now time to investigate if these principles can be applied to any correlation or independency found in the data or, in short, if they are universally valid. This section takes on this issue by examining the positive justifications proposed for the principles, studying some of the objections raised against them, and putting forward some new criticisms. It argues against the universal validity of the conditions, and further demonstrates the necessity of subject matter information in causal inference.

5.1 The Causal Markov Condition

The advocates of the GT approach have set forth several justifications for the Markov condition. It has been argued that variants of the principle underlie other methods of causal inference, and in this respect the GT approach is the same as other causal inference methods (Glymour, 1997:203-5). This claim only means that the conclusions obtained using the GT techniques are as valid as those obtained using other methods. This in itself offers no justification for the condition. It has also sometimes been claimed that if one does not assume the universal validity of the Markov condition, some correlations remain unexplained. Implicit in this defence is that if a correlation has no causal explanation it has no explanation. But this is the very claim that one must defend for establishing the validity of the condition; one cannot simply take it for granted.

The central justification for the Markov condition, however, is said to derive from the fact that it is provably true of recursive, pseudo indeterministic, causally sufficient structures, with independently distributed disturbance terms (Kiiveri et al., 1982).[18] Koster (1999) and Spirtes (1996) have shown that a more general property, called the global Markov condition, is true of both recursive and non-recursive causally sufficient, homogenous and pseudo indeterministic linear structures, with independently distributed errors.[19] In what follows, the focus of analysis will be on the Markov condition defined in Section II, even though the analysis is also relevant to the global Markov condition.

The proof of the Markov condition is a piece of mathematics. To relate it to the world, it is necessary to show that the underlying requirements are true of the world. Of these conditions, recursiveness is not a critical issue (at least in the case of linear models), since the global Markov condition is true of both recursive and non-recursive (linear) structures, which satisfy the other conditions. The pseudo indeterminism requirement has come under attack by critics concerned with the outcomes of quantum mechanical experiments that seem to point to indeterminism. At the quantum level, the world is said to be genuinely indeterministic and the Markov condition does not apply. Since the universal validity of the condition can be challenged without taking sides on indeterminism, we take the pseudo-indeterminism condition for granted, and focus on the causal sufficiency and independence of the

[18] Kiiveri et al. (1982) provided the first proof of the result. A simple proof of the result appears as an appendix in Cartwright (2002:451-452).

[19] For definition of the Global Markov condition see footnote 8. The proofs by Koster (1999) and Spirtes (1996) assume linearity of the structural model.

errors requirements. The founders of the GT approach acknowledge that these conditions may not be true of a set of measured variables. Yet, to universally apply the Markov condition, they introduce the completeness hypothesis:

The Completeness Assumption: For every set of recorded variables ***O***, either the set forms a causally sufficient set with uncorrelated errors or it can be embedded in a larger set of variables ***V*** that is causally sufficient with uncorrelated errors (Scheines, 1997:197).

On this basis, the Markov condition is generalised to every set of variables, at least at the level of description with which social scientists, economists, and biologist are concerned. In light of this, the universal validity of the Markov condition critically depends on the validity of the completeness hypothesis. We therefore concentrate our analysis on this hypothesis. Before proceeding, it should be stressed that exact independencies are not known in practice. One only has access to an estimate of the joint probability distribution of the variables under study, usually obtained from a small sample, and should take approximately zero correlations in place of exact independencies. To make any causal inference from data, the (Population) Markov Condition, which is defined for true probabilities, should be replaced with the so-called *sample* Markov condition:

The Sample Markov Condition: Let $\hat{P}(\mathbf{V})$ be a joint probability distribution estimated from *a finite sample* of observations on variables ***V***. The pair $\langle G, \hat{P} \rangle$ satisfies the sample Markov condition if and only if every variable *X* in ***V***, conditional on its parents, is *almost independent* of every variable *Y* in ***V*** that is not a descendent of *X*.

5.1.1 Aggregation over Heterogeneous Units

The literature points to several circumstances in which completeness can fail. An important case was pointed out by G. Udny Yule in his seminal paper (1903) on the theory of association of attributes in statistics, where he noted that mixing heterogeneous units could lead to creation of spurious correlations at the population level that did not exist at the level of sub-populations. An illustration of such a phenomenon is presented in Table 5.1.

Table 5.1

Aggregation over Heterogeneous Units

	Male Population		**Female Population**		**Mixed Population**	
	Treated	*Untreated*	*Treated*	*Untreated*	*Treated*	*Untreated*
Alive	4/99	16/99	20/99	10/99	24/99	26/99
Dead	8/99	32/99	6/99	3/99	14/99	35/99

In both female and male subpopulations treatment and recovery as well as non-treatment and non-recovery are uncorrelated. When the two subpopulations are mixed together, however, recovery becomes statistically related to treatment and non-recovery to non-treatment. Such examples show that mixing populations, which either have different causal structures or have the same causal structure but different probability distributions, can create associations that do not exist at the sub-population level. Since such associations are by-products of mixing, the mixed population violates the Markov condition.

Spirtes et al. (1993:57) describe in some detail Yule's example, which is similar to the above example, to explain why it presents no problem for the Markov condition. The authors argue that the variables in Yule's example exclude a variable that is the cause of membership in a sub-population. Once the omitted variable is included, and the measured variables are conditioned on it, the spurious correlations disappear (1993:60). In the above example, the analyst has failed to include, for instance, gender. Once he includes gender and conditions treatment (non-treatment) and recovery (non-recovery) on it, the spurious correlations disappear, and the population satisfies the Markov condition.[20]

[20] For further discussion of how correlations arising from mixing heterogeneous units are dealt with see Glymour (1997a:207) and Glymour and Meek (1995:1012).

Although it may be possible in simple situations like the one above to locate classifying variables that can sensibly be considered as common causes, in more complex cases of aggregation over heterogeneous units, which are ubiquitous in the social sciences, there exists no small set of classifying variables capable of explaining away spurious correlations, which can at the same time be considered as common causes of the recorded variables. In social contexts, what is in fact required to explain away a spurious correlation at the aggregate level is a full description of the system at the micro level including the laws governing the behaviour of the individuals, their interactions with each other, and more importantly the socio-economic processes determining the variables affecting individual behaviour. However, a description of the system at the micro-level cannot be considered as a common cause of the variables at the aggregate level. To highlight this point, we borrow an example from the next chapter that studies in detail the complexities arising from aggregating over heterogeneous units. The example revolves around a simple economy studied in Lippi (1988:174), which has two consumers, each having a slightly different demand function. To be specific, the demand function for each individual follows the static routine:

$$Y_{it} = \Pi_i X_{it} \qquad i = 1,2, \tag{5.1}$$

which has no stochastic term. Y_{it} and X_{it} are respectively consumption and income of the *i*th individual in period *t*, and the parameter Π_i for each individual is *different.* Moreover, each consumer operates in a slightly different environment in the sense that the independent micro variable X_{it} for each individual follows a different autoregressive routine,

$$X_{it} = a_i X_{it-1} + v_{it}, \qquad 0 < a_i < 1, \tag{5.2}$$

where the parameter a_i for each individual is *different*, and the v_{it} are orthogonal white-noise processes.[21] As shown in Appendix *E* of the next chapter, in this economy the function relating aggregate consumption $Y_t = Y_{1t} + Y_{2t}$ to aggregate income $X_t = X_{1t} + X_{2t}$ is given by

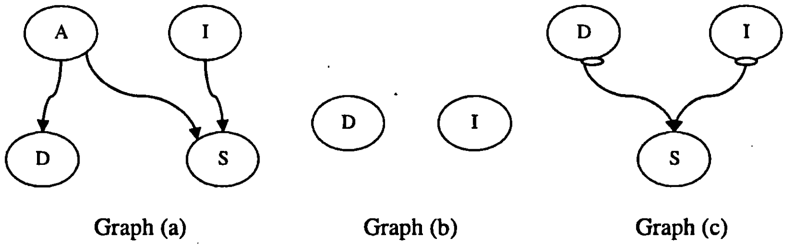
$$Y_t = \alpha Y_{t-1} + \beta X_t + \gamma X_{t-1} + u_t, \qquad (5.3)$$

with u_t being a white-noise process. Contrary to the individual consumption functions, this function contains among its arguments lagged aggregate consumption and income. Moreover, as the number of consumers increases, the complexity of the function grows, including an increasingly larger number of lagged predictors. Now, the point is that the relation between Y_t and Y_{t-1} in (5.3) is not a causal relation, since the last period individual consumption Y_{it-1} does not appear in the individual consumption function and, for that reason, setting Y_{t-1} by intervention at certain value does not affect Y_t. To explain away the spurious correlation, one needs a description of the economy at the micro level, including a description of the choice situation faced by each individual. In real-life situations, providing such a description is impossible. In addition, the description would involve a tremendously large number of classificatory variables (e.g., 'being a farmer', 'being a banker'), which cannot be considered as the common causes of the aggregate variables, say, Y_t and Y_{t-1}. As this example shows, in social contexts, where decision makers are different, and operate in different choice situations, aggregation over heterogeneous units produces variables that neither stand in a causal relation with each other nor are part of a larger causally sufficient set of variables. In such situations, completeness and hence the Markov condition fail.

[21] The stochastic process $\{Z_t, t = 1,2,\ldots\}$ is a white-noise process if $E(Z_t) = 0$ and $Cov(Z_t, Z_s) = \delta^2$ if t=s and $Cov(Z_t, Z_s) = 0$ if $t \neq s$.

5.1.2 Selection Bias

Aggregation over heterogeneous units is only one of the situations in which completeness fails. Another situation in which completeness fails is when there is "selection bias"; that is, when a population is defined by conditioning on some variable Z that is a common effect of two or more of the variables under study (or their causes) that have no mutual influence on each other (Glymour, 1997:208). There has recently been a growing interest in studying the implications of selection bias for causal inference.[22] Here, we concentrate on a problem that selection bias creates for the completeness hypothesis, examine a proposal that some GT theorists have put forward to deal with it, and argue why, because of the possibility of selection bias, an important claim of the GT approach must be abandoned. We first consider an illustration of selection bias discussed in Spirtes, Meek, and Richardson (1996). Suppose a survey of college students is done to determine whether there is a link between *Intelligence* (I) and *Sex drive* (D). Let *Student statues* (S) be a binary variable that takes value 1 when one is studying in a college and zero otherwise. Also, as in graph (a) in Figure 5.1, suppose *Age* (A) causes *sex drive*, and *age* and *intelligence* cause *student status* (here, *age* is taken to be a proxy for a combination of biological and mental states associated with age).



Graph (a) Graph (b) Graph (c)

Figures 5.1

[22] See Cooper (1995, 2000) and Spirtes, Meek and Richardson (1996).

Since the sample is gathered from college students, the variables under study or their causes, i.e., *I* and *A*, influence whether one is in the sample, and this can create a correlation among the *recorded* variables, i.e., *I* and *D*. If graph (a) is an accurate description of the causal relations among $\mathbf{V} = \{A, D, I, S\}$, the correlation between *I* and *D* is spurious, as there is no causal connection among them (graph (b)). Moreover, *V* contains no common cause of *I* and *D* that can screen off the correlation. The Markov condition is not true of the recorded variables *I* and *D*. Nor is there a larger DAG with the common causes of *I* and *D* that satisfies the condition – hence a serious failure of completeness.

A number of proposals have been set forth to counter the danger of selection bias, mainly calling for the use of domain-specific information and sensitivity analysis (Scharfstein et al., 2003). Against this approaches, following a proposal by Wermuth et al. (1994), Cooper (1995) argues that selecting a unit to include in the sample is a causal event. It can be represented by a variable and treated as a genuine part of the causal structure.[23] So, he proposes to incorporate into the structure the process of unit (case) selection, adding an extra assumption to the arsenal of the assumptions underlying the GT approach:

Selection Bias Assumption: Case selection is a causal event that can be modelled within a causal directed graph that has a variable representing whether a case was selected or not (Cooper, 2000).

A similar supposition underlies an attempt by Spirtes, Meek, and Richardson (1996) to extend the GT techniques to data that might be affected by selection bias. On this proposal, the set $\{I, D\}$ does not actually exhaust all the recorded variables in the current case. The recorded set is said to be $\{I, D, S\}$, where *S* is a selection variable taking value 1 for the students in the sample and zero for non-students. The dependence $\neg(D \perp I)$ appeared in the sample should then be interpreted as $\neg(D \perp I/(S=1))$, which means graph (b) ought to be replaced with graph (c) in

[23] In other words, it can be treated as a variable in a higher dimensional probability space.

Figure 5.1 (where the small ovals indicate that each arrow can be replaced with a bi-directed edge ↔). Now, there is many ways to embed graph (c) into a DAG to make it consistent with the Markov condition. Figure 5.2 depicts two possibilities:

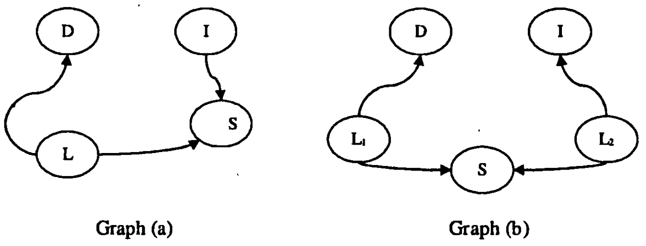


Figure 5.2

Although there may be nothing theoretically wrong with this proposal, it comes with a price for the GT approach. Inclusion of selection variables adds to the complexity of the structure. This enlarges the class of models that, given the Markov and Faithfulness conditions, could have possibly generated the independence data. In that case, the models will have less in common and much less can be learnt about the structure from the data. Specifically, the increase in the class of graphs consistent with the independence data undermines the central claim that the GT techniques are able to establish whether or not a correlation is *definitely* due to latent common causes. Recall, when the orientation of an undirected graph leads to a bi-directed edge, the edge is taken as the evidence that the correlation is *definitely* due to a latent common cause. In the analysis of the second example in Section III, the Faithfulness condition required placing a bi-directed edge between X_2 and X_3 and, following the GT theorists, it was concluded that the correlation was due to latent common causes. However, when the possibility of selection bias is acknowledged, this inference is no longer warranted. This is because the bi-directed edge can simply be due to selection bias. An example of such an explanation is given in graph (b) below, which is also found in Spirtes et al. (1996).

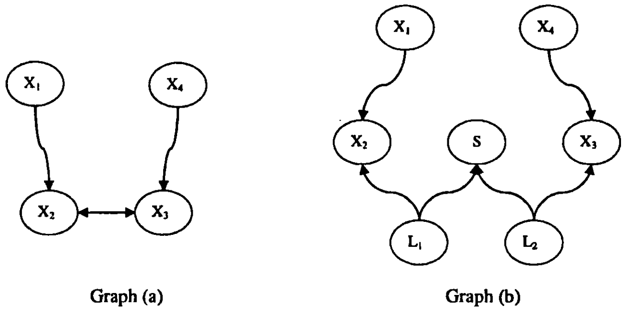


Figure 5.3

Graph (b) implies all the independencies over the recorded variables in graph (a). Yet, it contains no variable affecting both X_2 and X_3. If structures like graph (b) are permitted, it is no longer possible to take a bi-directed edge as the evidence for a latent common cause. Such an interpretation demands ensuring that the bi-directed edge is not the result of selection bias. The GT approach provides no formal guidance how to decide whether a data set is affected by selection bias or not. It too must rely on domain-specific information or sensitivity analysis to counter the threat of selection bias. Finally, allowing selection-variables in a causal structure demands revising the main theorem of the GT approach given in Glymour (1997:219). The theorem assumes that there is no selection bias.

5.1.3 Concomitants

The advocates of the GT approach may admit the inability of their methods to establish the existence of latent common causes but still argue that, regardless of selection bias, the GT techniques are able to establish in certain cases that a variable either directly or indirectly but definitely causes another variable. This claim is also unfounded. To explain this, it is important to recall the distinction between a cause and a concomitant of a cause (Sobel, 1995:29). The possibility of mistaking a cause with a concomitant of the cause creates another situation where completeness fails. In such cases, it is wrong to admit the outcome of the GT techniques that a variable causes another. As an illustration, suppose we are given data on four variables: *Mother's Genotype* (G), *Mother's childhood nutrition* (N), *Mother's occupation* (O), and *Children's intelligence* (I). It is plausible to assume that the following independencies are approximately true in the sample:

$$Ind_p = \{G \perp N, G \perp I / O, N \perp I / O\}$$

Given the Markov condition and Faithfulness, these independencies lead to graph (a) in Figure 5.4, where the ovals at the end of the arrows between *G* and *O* and *N* and *O* indicate that each arrow can be replaced with a bi-directed edge $\leftrightarrow$.

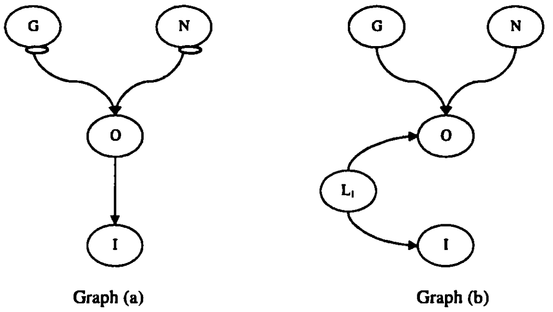


Figure 5.4

Interpreting it causally, graph (a) reads that Mother's occupation causes (possibly indirectly) child's intelligence. Such a claim, though logically possible, is not taken seriously at present. The graph suggests a causal connection from *O* to *I* that does not seem to exist. Assuming completeness, the strategy of the defenders of the Markov condition would be to try to embed the graph into a DAG $G(\mathbf{O}, \mathbf{L})$ with a common cause *L* that screens off the correlation between *O* and *I*. But, if the assumptions of the GT approach are taken for granted, no such DAG can exist. Graph (b) in Figure 5.4 shows a typical DAG capable of explaining away the correlation between *O* and *I*. Any such graph neither entails independence relation $G \perp I / O$ nor $N \perp I / O$, and is not faithful to the distribution of the recorded variables. In the current example, completeness can be restored only at the expense of Faithfulness and Faithfulness can be retained only at the expense of completeness. In either case, the immediate conclusion is that, because of the possibility of mistaking a concomitant of a cause with the cause, it is not warranted to take the existence of an irremovable arrow, such as the one from *O* to *I*, as the evidence of a definite causal connection. Like other approaches to causal inference, the GT method cannot also establish that a variable directly or indirectly but definitely causes another variable.

The way to deal with the problem created by concomitants is not to search for a larger set of variables that includes the original ones. It is to replace them with the right ones. Spirtes, Glymour, and Scheines (1993:63) come close to a similar conclusion when dealing with a counterexample put forward by Wesley Salmon to the common cause principle, calling it *interactive forks.*[24] They argue that the apparent counterexample arises because one has failed to pick up the right variables to describe the situation in hand. This simply means that the Markov condition generates sensible results only when applied to right variables. Moreover, one cannot rely on formal principles to decide on the right set of variables to describe a situation. One needs domain-specific information.[25]

The analysis has so far focused on one component of the completeness conjecture that says for every set of variables ***O*** that is not causally sufficient there is a causally sufficient set ***V*** embedding ***O***. It remains to investigate the other component that says the disturbance terms associated with the variables ***V*** are independently distributed. Pearl seems to suggest that this condition is not an extra assumption. On his view, the independence assumption follows from the causal sufficiency assumption and the common cause principle, which he regards it as basic for linking probability with causation (2000:30). Other GT theorists have also taken a similar line (Richardson et

[24] Consider variables *X*, *Y*, and *Z*. Suppose *Z* causes *X* and *Y* but there is no causal link between *X* and *Y*. Salmon calls such a situation an interactive fork if P(X/Z)<P(X/Z&Y). For some examples see Salmon (1984:168-174).

[25] Another case where the completeness hypothesis may fail is raised in Sober (1987), discussed under the nomenclature of "Co-evolving Processes". An interesting examination of Sober's type of counterexamples is found in Hoover (2003).

al., 1999). But it should not be difficult to gather from our earlier discussion why causal sufficiency and the common cause principle do not entail the independence of the errors. A disturbance term *u* associated with an exogenous variable *X* in *V* represents the aggregate effect of all the variables outside *V* that influence *X*. Aggregation can render dependent variables that are independently distributed at the micro level. Therefore, even if all the variables affecting those in *V* are pairwise independent, when they are aggregated, they might become correlated (Cartwright, 2001). The requirement of independent errors is an additional assumption that lacks a justification. Altogether, these analyses reveal why the completeness hypothesis cannot be taken for granted, and, as a result, why the Markov condition cannot be applied universally.

5.2 The Faithfulness Condition

Faithfulness rules out any structure that, when the Markov condition applied to it, does not entail all the independencies found in the data. The GT literature offers several considerations to support a *priori* exclusion of such structures. Glymour (1997:210) begins his defence of Faithfulness by showing that it underlies other approaches to causal inference. But this provides no justification for causal conclusions drawn using the GT techniques. Scheines (1997:194) defends the condition by arguing that it increases our inferential power, as without it nothing can be learnt from data about the direction of causal influence. Again, the increase in inferential power is in itself no evidence for the soundness of the conclusions, and as such provides no support for Faithfulness.

5.2.1 The Measure Theoretic Argument

The main justification of Faithfulness is of a Bayesian nature. It has been argued that for any linear structural model, the set of parameterisations of the model that lead to violations of Faithfulness is of Lebesgue measure zero. Hence, any Bayesian whose prior over the parameters is absolutely continuous with the Lebesgue measure assigns a zero *prior* probability to the violations of Faithfulness (SGS, 1993:68-9).[26] A quick challenge to this argument, also noted by Scheines et al. (1998:82), is to ask why one has to have a prior that is absolutely continuous with respect to the Lebesgue measure. Obviously, if one adopts a prior that lacks this feature, the measure theoretic argument has no force. Though this criticism is sufficient to challenge the argument, more can be learnt by analysing what is really involved in having a prior that assigns zero probability to violations of Faithfulness. To this end, we follow a line of analysis in Robins et al. (1999) and Robins (2003). Consider a normally distributed causally sufficient set of variables $\mathbf{V}=\{X,Y,Z,U,V,W\}$, and let $\mathbf{O}=\{X,Y,Z\}$ be the recorded variables. Suppose *X* precedes *Y*, and *Y* precedes *Z*. Also, assume there is an extremely large sample of data on *X*, *Y* and *Z* so that estimation problems can be left aside. Finally, suppose the following dependencies and independencies are true in the data:[27]

$$\rho_{XY}=0.5\,; \quad \rho_{YZ}=0.5\,; \quad \rho_{XZ}=0.25\,; \quad \rho_{XZ.Y}=0\,.$$

Explanation (1): A possible explanation of these data is given by graph (a) in Figure 5.5. The graph implies that *X* causes *Y*, *Y* causes *Z* and that these variables have no common causes in *V*.

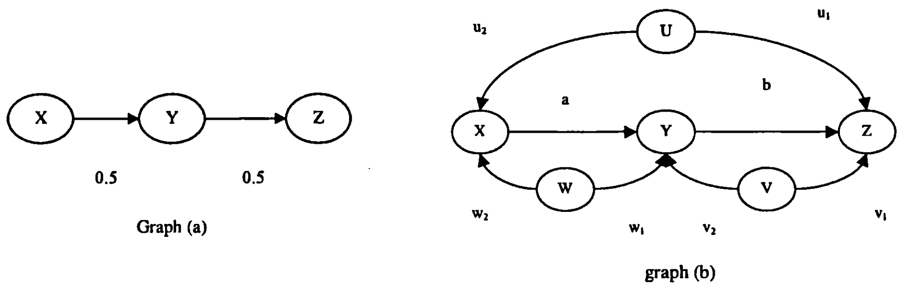


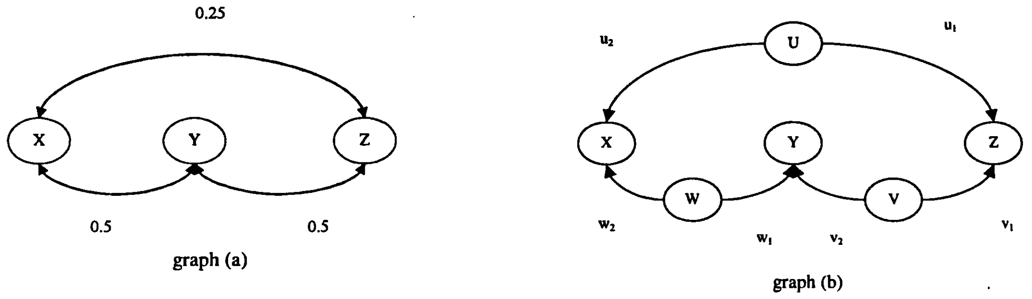
Figure 5.5

[26] Lebesgue measure is the uniform distribution in Euclidean space, e.g., length, area, volume.

[27] This example originates from Sewell Wright (1934) and is described in Irzik et al. (1987:508-9).

Another way of representing the same causal facts is given in graph (b), where the lower case letters denote path coefficients. Represented in this way, the explanation implies that $u_1u_2 = 0$, $v_1v_2 = 0$, $w_1w_2 = 0$, but, $a \neq 0$ and $b \neq 0$.

Explanation (2): A second possible explanation is offered by graph (a) below. According to this graph, neither *X* causes *Y* nor *Y* causes *Z*. The dependencies and vanishing partial $\rho_{XZ.Y} = 0$ are due to particular residual correlations between *X* and *Z*, *X* and *Y*, and *Y* and *Z* – as shown by the numbers on the bi-directed edges linking the variables.



graph (a)

graph (b)

Figures 5.6

If we follow the GT theorists in explaining residual correlation in terms of latent common causes, graph (b) above provides an alternative representation of the causal facts in graph (a). On this graph, *U*, *V*, and *W* are confounders. This means $u_1u_2 \neq 0$, $v_1v_2 \neq 0$, $w_1w_2 \neq 0$, but, $a = 0$ and $b = 0$.

These explanations are both possible. The measure theoretic argument draws on the fact that the subset of coefficient values for $\{u_1, u_2, v_1, v_2, w_1, w_2\}$ that yields the vanishing partial $\rho_{XZ.Y} = 0$, when $u_1u_2 \neq 0$, $v_1v_2 \neq 0$, and $w_1w_2 \neq 0$, has Lebesgue measure zero in R^6. If one has a prior over the parameter space that is absolutely continuous with the Lebesgue measure of the space, then one has to regard Explanation (2) as *a prior* unlikely and accept Explanation (1), which is faithful to the data. The difficulty with this argument is that the move from showing that Explanation (2) is *a priori* unlikely to acceptance of Explanation (1) is not warranted. As noted earlier, Explanation (1) implies that $u_1u_2 = 0$, $v_1v_2 = 0$, and $w_1w_2 = 0$. Now, the Lebesgue measure of each of these events is also zero in R^2. And, if one has a prior over the parameters that is absolutely continuous with the Lebesgue measure, one also has to consider these events as *a priori* unlikely. As far as the measure theoretic considerations are concerned, both explanations are equally unlikely. The only way the balance can be tilted in favour of Explanation (1) is to rule out *a priori* any latent common cause for the recorded variables.[28] If the existence of common causes is not *a priori* ruled out, both explanations are *a priori* equally likely, and no causal conclusion can be inferred from the data. This means, to believe that violations of Faithfulness are *a priori* unlikely, one must believe that *X* and *Y* and *Y* and *Z* have no latent common causes.

In the above analysis, by assuming the existence of an extremely large sample, it was assumed that the true independencies were known. In practice, as said earlier, one only has access to a finite sample, and should take approximately zero dependencies in lieu of exact independencies. This requires substituting the population Faithfulness condition, which is defined for true independencies, with the so-called *sample* Faithfulness condition:

The Sample Faithfulness Assumption: In a large sample if *X* and *Y* are *almost* independent conditional on *Z*, that is evidence that *X* and *Y* are not directly causally connected except through *Z* (Glymour et al. 1999:345)

[28] In other words, one has to assign *a priori* non-zero probabilities to events $u_1u_2 = 0$, $v_1v_2 = 0$ and $w_1w_2 = 0$.

In light of this, what the GT theorists need to exclude *a priori* is the set of parameter values that *nearly* cancel each other out. Such a set always has a non-zero Lebesgue measure, and cannot be excluded on measure theoretic grounds. The Bayesian argument applies, if at all, only when the true independencies are known. It has no force in practice, where almost-zero partial correlations should be taken in place of exact independencies.

5.2.2 Stable Unfaithfulness

Another line of defence of Faithfulness has been pursued in Pearl's writings. Pearl's view of a causal model is influenced by the early views in econometrics. These views define a structural model as a system of equations each representing an *autonomous* causal mechanism that can be manipulated without affecting other equations in the model. Autonomy, the early economists argued, is an essential feature that a model must have to be useful for evaluating actions and policies. Influenced by this tradition, Pearl argues that the reason we search for causal models is the need for evaluating actions and policies, and a key feature that a model ought to have to be useful for analysis of actions and policies is the autonomy of the model equations. Since the equations in unfaithful models break down with a slight change in the conditions sustaining one of the equations, the models lack autonomy and are not useful for evaluating actions and policies. They should not, therefore, be taken seriously (2000: 63).[29]

A number of authors have rightly challenged this claim. Cartwright (1999:118) and Hoover (2001:170) have pointed out that one of the ways that we minimise damages in our social and medical regimes is by arranging the system so that conflicting causal forces counterbalance the effect of each other. Unfaithful structures can therefore be of significant interest in designing efficient social and medical regimes. Moreover,

[29] 'Autonomy' or 'invariance' is defined with respect to a specific set of changes. An elaboration of this point is beyond the scope of this chapter. See Woodward (1993).

what is really at issue here is whether Faithfulness is a reliable guide to discovery of autonomous relations. Surely, a definition of autonomy and a recommendation to avoid using unstable relations in policy analysis cannot serve as a guide in searching for structural relations.

In another charitable reading, Pearl may be taken as arguing that, since unfaithful structures are unstable, they do not last enough to generate data for a reliable estimate of the underlying distribution. So, any independencies embedded in a distribution estimated from an adequately large homogenous sample arise from a faithful structure. It is thus a sound practice to rely on Faithfulness to infer causation from reliably estimated independencies. This line of reasoning assumes that there can be no 'stable' unfaithful independencies. But, this is wrong. We earlier noted the difficulty that 'concomitants' could create for the Markov condition by generating dependencies that lack a causal explanation. Mistaking a concomitant of a genuine cause with the cause can also produce *stable* independencies that do not represent absence of causation. Consider the structure depicted in graph (a) in Figure 5.6 that represents a possible causal structure between Genotype (*G*), Family background (*F*), Heavy smoking (*H*), and Lung cancer (*L*).

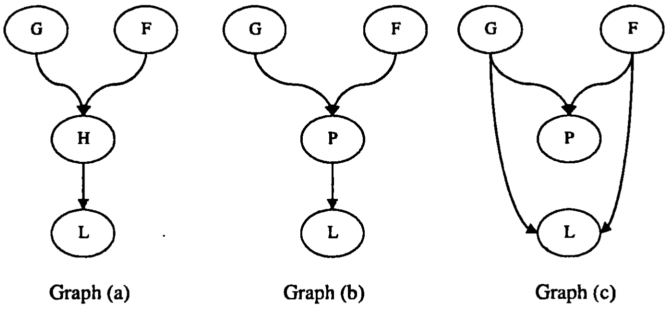


Figure 5.6

According to this structure, conditional on *H*, *L* is independent of *G* and *F*, which means there is no direct causal link from genotype and family background to lung cancer; they cause lung cancer through causing heavy smoking. Now suppose *H* is replaced with one of the concomitants of heavy smoking like 'Having yellowed teeth' (P). Assuming that the conditional independence relation $L \perp \{G, F\} / H$ is true in the data, the conditional independency $L \perp \{G, F\} / P$ will also most likely be true, and if one picks up variables $\{G, F, P, L\}$ instead of $\{G, F, H, L\}$, one ends up with graph (b). The graph entails that, conditional on having yellowed teeth, lung cancer is independent of genotype and family background. However, based on our current state of knowledge, independence relations $L \perp G / P$ and $L \perp F / P$ do not genuinely represent absence of causal connections. Assuming graph (a) is true, when 'heavy smoking' is dropped from the graph, there will be causal links from G and *F* to *L*, as shown in graph (c). Moreover, the spurious independencies $L \perp G / P$ and $L \perp F / P$ are stable in the sense that they are true as long as the structure depicted by graph (a) is true. Such examples, which are by no means rare, illustrate cases of stable unfaithfulness that are neither generated by exact cancellation of parameter values nor by mixing of heterogeneous units (the Simpson Paradox).[30] Pearl's stability argument may be useful for excluding violations of Faithfulness that arise from exact cancellation of parameter values. It is of no force in ruling out stable unfaithful independencies that arise from mistaking concomitants with genuine causes. Like the Markov condition, Faithfulness cannot be applied universally either.

6 Conclusion

The strongest assumptions that can be made about the link between causation and probability are that "every probabilistic dependency has a causal explanation" and "every probabilistic independency reflects lack of a causal connection". This chapter has argued that these hypotheses are not true. A correlation or non-correlation can arise for reasons other than causal reasons, and hence the class of possible explanations for a correlation or non-correlation is larger than the class of possible causal explanations. So, there can never be an entirely data driven causal inference

[30] Simpson's Paradox has been taken up by many authors in detail (Cartwright, 1997, and Hausman, 1998). The philosophical problems arising from the paradox are similar to those discussed under the heading of mixing heterogeneous units.

method. Causal inference must first proceed by eliminating non-causal explanations that a dependency or independency can have. This requires drawing on subject matter information available about the system. In the simple economy we described, it is essential to know the rules governing the behaviour of the individuals as well as the character of the environment in which they operate to decide whether the correlation between the aggregates reflect a causal connection or is an artefact of aggregation. The Markov and Faithfulness conditions become relevant only after non-causal candidate explanations are eliminated.

Even when non-sample explanations are ruled out, because of the existence of causally equivalent models, the Markov and Faithfulness conditions are not sufficient to pin down a single causal model. Given a causal model, there is always a simple rule to generate a class of equivalent models. Since the coefficient estimates of the common edges are not invariant across the models and their sign and significance vary from one model to another, very little or often nothing can be learnt from data without further subject matter information. Specifically, extra subject matter information is needed to reduce the class of causally equivalent models by excluding unlikely but possible causal models.

The reliability of the GT algorithms, and indeed any data-driven method of causal inference, is contingent on the sample size and the joint distribution of the variables under study. The GT algorithms proceed by assuming that the data comes from a multivariate normal distribution. When the sample is large, this assumption may be justified, and one can reliably test independence hypotheses. But, in practice, where the samples are small, the normality assumption can lead to wrong conclusions. As a general rule, since the known multivariate distribution families are very limited, and all make the restrictive assumption that the marginal distributions of the variables belong to the same family, in practice test of independence hypotheses needs great care.

Also, for analysis of policies and actions, one needs to know not only whether an equation in a model represents a causal relation but also the circumstances under which it remains invariant. Recall Haavelmo's remarks about the relation between pressure on the throttle and acceleration of the car (Haavelmo, 1944). For forecasting the effect of taking a car to a territory not yet explored one needs to know not only whether putting pressure on throttle makes the car accelerate but also the circumstances under which the relation remains invariant. The methods studied here are at best suited for discovering a causal structure, understood as a complex of type-level causal connections (Cartwright, 1997:435). They are not suitable for understanding the circumstances under which the structure continues to operate. This needs knowledge of the experimental conditions or the chance set-up, to use Cartwright's phrase (1997:357), which have given rise to and sustain the causal relations.

These analyses have major implications for modelling bounded rationality. They imply that to explain how people are able to make causal inferences from usually small samples one has to search for an approach, which emphasises the interaction between domain-specific causal knowledge and statistical learning (Griffiths, et al., 2004). The causal knowledge is required to restrict plausible causal relationships, their functional form, and strength. This will limit the space of plausible models, making it possible to infer causal conclusions from small samples. But again, as in statistical model formulation, we are faced with the question of where the domain specific knowledge comes from. One thing is clear about this question. The information does not come from a statistical analysis of data. The IS hypothesis does not provide a full account of human causal learning.

Appendices

Appendix A: The Path Analysis Principle

We state the proofs of the two principles for the case where there are only three variables *X*, *Y* and *Z* under investigation.[31] Extension to more general cases is straightforward. Since for the current purpose there is no interest in the first moments, each variable is expressed as a deviation from its mean.

Proof of the fist principle:

Let

$$\begin{aligned} X &= u_x \\ Z &= \alpha_{xz} X + u_z \\ Y &= \alpha_{yz} Z + u_y . \end{aligned} \tag{A1}$$

Assumption (i): u_x, u_z, and u_y are uncorrelated, and
Assumption (ii): u_z and *X*, and u_y and *Z* are uncorrelated.

Given these assumptions, the objective is to establish that $\rho_{xy.z} = 0$.

Multiply both sides of the equation for *Z* with *X*. Taking expectations of both sides of the equation gives

$$E(XZ)/E(X^2) = \rho_{xz} = \alpha_{xz} . \tag{A2}$$

Also multiply both sides of the equation for *Y* with *Z*. Taking expectations of both sides of the equation yields

$$E(YZ)/E(Z^2) = \rho_{yz} = \alpha_{yz} . \tag{A3}$$

Multiplying both sides of the equation for *Y* with *X* and taking expectations of both sides of the equation leads to

$$E(YX)/E(X^2) = \rho_{xy} = \alpha_{xz}\alpha_{yz} . \tag{A4}$$

Therefore,

$$\rho_{xy} = \rho_{xz}\rho_{yz} .$$

[31] This appendix is based on Irzik, 1987.

Finally, recall the expression for partial correlation

$$\rho_{xy.z} = (\rho_{xy} - \rho_{xz}\rho_{yz}) / (1-\rho_{xz}^2)^{1/2}(1-\rho_{yz}^2)^{1/2}. \qquad \text{(A5)}$$

Since the numerator is zero (because $\rho_{xy} = \rho_{xz}\rho_{yz}$), $\rho_{xy.z} = 0$.

The proof for the second principle is quite similar. We replace (A1) with

$$\begin{aligned} Z &= u_z \\ X &= \alpha_{xz}Z + u_x, \\ Y &= \alpha_{yz}Z + u_y \end{aligned} \qquad \text{(A6)}$$

and compute ρ_{xz}, ρ_{yz}, and ρ_{xy}.

Appendix B: The Conditional Independence Properties

Some of the properties of conditional independence, studied by Dawid (1979), include:

(1) *Symmetry*: $(X \perp Y/Z) \Rightarrow (Y \perp X/Z)$;

(2) *Decomposition*: $(X \perp YW/Z) \Rightarrow (X \perp Y/Z)$;

(3) *Weak union*: $(X \perp YW/Z) \Rightarrow (X \perp Y/ZW)$;

(4) *Contraction*: $(X \perp Y/Z) \& (X \perp W/ZY) \Rightarrow (X \perp YW/Z)$;

(5) Intersection: $(X \perp W/ZY) \& (X \perp WY/ZW) \Rightarrow (X \perp YW/Z)$.

For a detailed discussion of these properties see Pearl (1988:82-83).

Appendix C: The Common Cause Principle

Consider a DAG *G* true of variables $\mathbf{V} = \{X_1,...,X_n\}$. Define X_c as a common cause of X_a and X_b in *V* just in case there is a directed path from X_c to X_a and a directed path from X_c to X_b. Let *C* denote the set of common causes of X_a and X_b in the *V*.[32]

> **Claim**: Suppose X_a and X_b are conditionalized on *C*. If X_a is not a cause of X_b and X_b is not a cause of X_a, then every path between X_a and X_b in *G* is *d*-separated (inactive or blocked).

[32] The proof to follow is based on a footnote in Arntzenius, (1999).

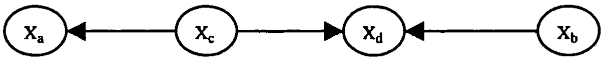
For any path *P* between X_a and X_b, either (i) *P* departs from X_a (i.e., is out of X_a) or (ii) it arrives at X_a (i.e., is into X_a).

Case (i): Suppose *P* is a path out of X_a. Since X_a is not a cause of X_b, the path cannot be a directed path, and therefore along the way to X_b it must reach a collider X_d. Since neither X_d nor any of its descendent is in ***C***, X_d blocks (*d*-separates) the path between X_a and X_b.

Case (ii): Suppose *P* is a path into X_a. Since X_b is not a cause of X_a, the path cannot be a directed path, and therefore somewhere along the way it must change direction. Starting from X_a and moving along the path towards X_b, there are two general possibilities:

(a): *P* changes direction at a variable X_c from which there is a directed path into X_b. In that case, X_c is a common cause of X_a and X_b and in ***C***, *d*-separating the path between X_a and X_b.

(b): Suppose the path from X_c to X_b is not a directed path. In that case, it must contain a collider X_d, as in graph (a).

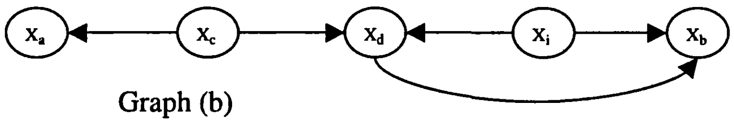


Graph (a)

To take up this possibility, it is enough to concentrate on path P* between X_d and X_b. As before, these paths can be of two types. Either they are into X_d or they are out of X_d.

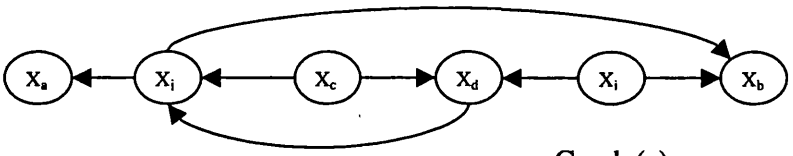
For any path P* that is into X_d, the whole path between X_a and X_b, created by joining the sub-paths between X_a and X_d, and X_d and X_b, is inactive, as neither X_d is in ***C*** nor a descendent of it.

For any path P* between X_d and X_b which is out of X_d there is also two possibilities. Either it changes direction at some points between X_d and X_b or it forms a directed path towards X_b. If it forms a directed path and intersects with no node between X_c and X_a as shown in graph (b) below, node X_c will be a common cause and is included in ***C***. The whole path between X_a and X_b will be *d*-separated.



Graph (b)

On the other hand, if the directed path has a common node X_j with the path between X_a and X_c, there will be a directed path from X_j to X_d. In that case, X_j will be a common cause of X_a and X_b, as shown in graph (c).



Graph (c)

Since X_j is in ***C***, the whole path between X_a and X_b, formed by joining the (sub) path from X_a to X_j with the path from X_j to X_b, is *d*-separated. This exhausts all the possibilities that matter and, therefore, the desired conclusion.

Appendix D: The Use of the Faithfulness Condition in the Traditional Methods

Variants of the Faithfulness condition underlie other data-driven approaches to causal inference. Suppes (1970) defines an event C_{t*} to be a *prima facie* cause of an event E_t if and only if (i) t^* refers to a time point prior to *t*, (ii) the event C_{t*} has positive probability, and (iii) and C_{t*} is positively relevant to E_t, that is, $P(E_t / C_{t*}) > P(E_t)$. He then gives several conditions of spuriousness to distinguish genuine causes of E_t from those events spuriously related to E_t. On this account, the events that could possibly be causes of E_t are those that are positively correlated with it; an event C_{t*} cannot be a cause of E_t if it is statistically unrelated with E_t. This is nothing but the Faithfulness condition.

As another example, consider Granger's theory of causation (Granger, 1980). Let Ω_t denote the complete history of the world up to and including discrete time *t*, excluding deterministic relations among the components of this history. Let X_t stand for a random variable. Granger suggests that X_t causes Y_{t+1} if

$$P(Y_{t+1} \in A / \Omega_t) \neq P(Y_{t+1} \in A / \Omega_t - X_t)$$

for some set A. He next operationalises this definition by replacing the set Ω_t with a limited information set I_t that includes information on the history of the variables considered; i.e., $I_t = (X_t, Y_t, Z_t, \ldots)$, and relativizes his definition of causation with respect to I_t. Thus, he takes a confirmation of the statistical hypothesis of non-causality

$$P(Y_{t+1} \in A / I_t) = P(Y_{t+1} \in A / I_t - X_t)$$

by the data as the evidence that X_t does not causes Y_{t+1}. The inference from the independence of Y_{t+1} and X_t conditional on the information set $I_t - X_t$ to the denial of a causal link from X_t to Y_{t+1} is a special case of the Faithfulness condition (Robins, 2003:89).

Appendix E: The DAG Inversion Rule

Let *G* be any DAG containing the edge $X \rightarrow Y$, and G^* be a graph the same as *G* except that the edge $X \rightarrow Y$ is replaced with $X \leftarrow Y$. Then, G^* is a DAG equivalent to *G* if and only if every parent of *X* is a parent of *Y* and every parent of *Y*, except *X*, is a parent of *Y*.[33]

Part I (if part): Suppose G^* is not a DAG (i.e., contains a cycle). Since the only difference between *G* and G^* is that $X \rightarrow Y$ is replaced with $X \leftarrow Y$, and since *G* is a DAG, then there has to be a directed path from *X* to a variable *Z* which is a parent of *Y*. This means *Y* has a parent in *G* which is not a parent of *X*, contrary to the assumption. So, G^* is a DAG.

Now suppose *G* and G^* are not equivalent. Then, by theorem 4.1 in the text, either *G* or G^* contains an unshielded collider that is not present in the other. Since the only difference between *G* and G^* is that $X \rightarrow Y$ in G^* is replaced with $X \leftarrow Y$, the unshielded collider ought to be formed from $X \leftarrow Y$ and $X \leftarrow Z$, while *Z* is not a parent of *Y*. This implies that *X* in *G* has a parent that is not a parent of *Y*, contradicting the assumption. The same argument applies if *G* contains an unshielded collider that is not in G^*.

[33] The proof here is based on Chickering, 1995.

Part II (only if): Suppose *X* has a parent in *G* that is not a parent of *Y*. Substituting $X \rightarrow Y$ with $X \leftarrow Y$ creates an unshielded collider in G^*. Alternatively, suppose *Y* has a parent that is not a parent of *X*. Substituting $X \rightarrow Y$ with $X \leftarrow Y$ destroys an unshielded collider which is in *G*. In either case, *G* and G^* are not equivalent.

Appendix F: The Semi-Markovian Model Equivalence Theorem

Theorem 4.2: Let $G(\mathbf{O},\mathbf{L})$ be a directed acyclic graph, *X* and *Y* in ***O***, and $X \rightarrow Y$ hold in $G(\mathbf{O},\mathbf{L})$. Let $G^*(\mathbf{O},\mathbf{L}^*)$ be the same as $G(\mathbf{O},\mathbf{L})$ except that the directed edge $X \rightarrow Y$ is replaced in $G^*(\mathbf{O},\mathbf{L}^*)$ with the bi-directed edge $X \leftrightarrow Y$. (i) $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ are Markovian equivalent over ***O*** if for every variable *Z* in ***O*** that is a parent of *X* in *G*, *Z* is also a parent of *Y*. (ii) Furthermore, if $X \leftrightarrow Y$ is in $G(\mathbf{O},\mathbf{L})$, the bi-directed edge can be replaced with $X \rightarrow Y$ just in case every parent of *X* in $G^*(\mathbf{O},\mathbf{L}^*)$ is also a parent of *Y*.

The proof of this theorem follows from a theorem established in Spirtes and Verma (1992). Several graph-theoretic notions are needed to introduce the theorem:

Inducing path relative to *O*: If $G(\mathbf{O},\mathbf{L})$ is a DAG over variables ***V***, ***O*** is a recorded subset of ***V*** containing *X* and *Y*, where $X \neq Y$, then an undirected path *U* between *X* and *Y* is an inducing path relative to ***O*** if and only if every member of ***O*** on *U* except the end points (i.e., *X* and *Y*) is a collider on *U*, and every collider on *U* is an ancestor of either *X* or *Y*.[34]

Inducing path graph over *O*: *G** is an inducing path graph over ***O*** for DAG $G(\mathbf{O},\mathbf{L})$ if and only if there is an edge between variables *X* and *Y* with an arrow directed at *Y* if and only if *X* and *Y* are in ***O*** and there is an iducing path in $G(\mathbf{O},\mathbf{L})$ between *X* and *Y* relative to ***O*** that is into *Y*.

Partially oriented inducing path graph over *O*: Recall, the process of GT inference without causal sufficiency starts by constructing a skeleton over ***O***. For every pair *X* and *Y* in ***O***, it is checked whether *X* and *Y* are independent. If so, the edge between them is removed. It is then searched if there is any subset **Z** of $\mathbf{O}\setminus\{X,Y\}$ such that conditional on **Z**, *X* and *Y* are independent. If so, the edge between *X* and *Y* is removed. The process is repeated for every pair of variables in ***O***. The outcome at this stage is an incomplete undirected graph. Every end point between $X - Y$ admits two possibilities, i.e., '-' and '>'. To make these possibilities explicit, let us represent the undirected edge $X - Y$ between every connected pair *X* and *Y* by $Xo-oY$.

[34] These definitions are adapted from Spirtes, Glymour, and Scheines (1993), Ch 6.

In the next stage, we look at every triple (*X*, *Y*, *Z*). If there is an edge between *X* and *Y*, an edge between *Y* and *Z*, and no edge between *X* and *Z*, we replace $Xo-oYo-oZ$ with $Xo\rightarrow Y\leftarrow oZ$.

Then, for every $-oYo-$, it is checked if there can be a graph consistent with the data such that both 'o' are replaced with arrows, i.e., $\rightarrow Y\leftarrow$. If no such graph is consistent with the data, $-oYo-$ is replaced with $-o\underline{Y}o-$. The graphical object, thus constructed, represents all that can be learnt from the independence data about the underlying causal structure. The graph is referred to as a partially oriented inducing path graph over ***O***.

With these preliminaries in hand, Spirtes and Verma establish the following:

Theorem (Spirtes and Verma, 1992): If *G* is a DAG over ***V***, *G** is a DAG over ***V****, ***O*** is a subset of ***V*** and of ***V****, then *G* and *G** have the same d-separation relations among the variablese in ***O*** if and if they have the same partially oriented inducing path graph over ***O***.[35]

Given this theorem, the proof of theorem 4.2 is strightforward:

(i) Suppose $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ are defined as in the first part of Theorem 4.2 but are not Markovian equivalent over ***O***. By the above theorem, $G^*(\mathbf{O},\mathbf{L}^*)$ has a partially oriented inducing path graph over ***O*** different than that of $G(\mathbf{O},\mathbf{L})$. This can only happen if in $G(\mathbf{O},\mathbf{L})$ *X* has a parent *Z* in ***O*** that is not a parent of *Y and* Z and *Y* are independent (d-separated) conditional on *X*. In that case, $G^*(\mathbf{O},\mathbf{L}^*)$ includes the subgraph $Z\rightarrow X\leftarrow L\rightarrow Y$ that makes *Z* and *Y* dependent conditional on *X*. However, by assumption, every parent of *X* in ***O*** is a parent of *Y* in $G(\mathbf{O},\mathbf{L})$. So, both DAGs have the same partially oriented inducing path graph and are Markovian equivalent over ***O***.

(ii) Suppose $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ are as defined in the second part of the theorem. That is, they just differ in that $X\leftrightarrow Y$ is in $G(\mathbf{O},\mathbf{L})$ but $X\rightarrow Y$ in $G^*(\mathbf{O},\mathbf{L}^*)$. If $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ are not Markovian equivalent, it follows that $G(\mathbf{O},\mathbf{L})$ and $G^*(\mathbf{O},\mathbf{L}^*)$ produce different partially oriented inducing path graphs over ***O***. Again, this can only happen if *X* in $G^*(\mathbf{O},\mathbf{L}^*)$ has a parent *Z* in ***O*** that is not not a parnt of *Y and Z* and *Y* conditional on *X* are indepenent (d-separated). By assumption, every parent of *X* in ***O*** is also a parent of *Y* in $G^*(\mathbf{O},\mathbf{L}^*)$. Both DAGs, therefore, generate the same partially orineted inducing path garph and are Markovian equivalent over ***O***.

[35] This theorem is restated, somewhat differently, as corollary 6.4.1 in Spirtes, Glymour, Scheines (1993:189).

In either case, the condition given in the lemma is sufficient for the equivalence of $G(\mathbf{O},\mathbf{L})$ and $G^{*}(\mathbf{O},\mathbf{L}^{*})$.

Appendix G: The Limited Block Recursive Theorem

A proof for theorem 4.3 is found in Raykov et al. (1999:238-43). The proof is based on a proposition, established by these authors, which provides a general procedure for checking model equivalence. We outline the proof to explain how it can in general be checked whether two models are equivalent. To state the proposition, several technical notions are needed:

The first is the concept of **parameter transformation.** Let M_1 and M_2 stand for two models, with parameter spaces Θ and $\Theta*$ respectively. Call $\mathbf{g}:\Theta \to \Theta*$ a parameter transformation (mapping) if for each $\theta \in \Theta$ there is an $\theta* \in \Theta*$ such that θ is mapped into $\theta*$ by ***g***; that is, $\theta* = \mathbf{g}(\theta)$.

The mapping $\mathbf{g}:\Theta \to \Theta*$ is called **surjective** if for each $\theta* \in \Theta*$ there exists an $\theta \in \Theta$ such that $\theta*$ is mapped into θ by ***g***. A surjective transformation is an "onto" mapping. And, M_1 and M_2 are said to satisfy **$\sum$-condition** if, for all $\theta \in \Theta$, there is a ***g*** such that

$$\sum\nolimits_1(\theta) = \sum\nolimits_2[\mathrm{g}(\theta)],$$

where $\sum_1(\theta)$ is the covariance matrix implied by the parameter vector θ for model M_1 and $\sum_2[\mathrm{g}(\theta)]$ is the derived covariance matrix for model M_2.

Raykov et al.'s theorem (1999:206) can now be stated as follows:

General Model Equivalence Proposition: Two models M_1 and M_2 are equivalent if and only if they fulfill the $\sum$-condition with a surjective transformation $\mathbf{g}:\Theta \to \Theta*$ relating their parameters (Raykov, et al., 1999:206).

Informally, two models are equivalent if a transformation of the parameters of one of them can be found that preserves the implied covariance matrix, and covers the whole parameter space of the other.

The proof for theorem 4.3 then involves establishing that there is a transformation ***g*** such that (i) the model before applying the theorem, denoted by M_1, and the model obtained by applying the theorem, denoted by M_2, satisfy the $\sum$-condition; and (ii) ***g*** is surjective. To state the proof, some further notations and preliminaries are needed:

1. Notations:

M_1: the model before replacement of $X \rightarrow Y$ with bi-directed edge $X \leftrightarrow Y$;
M_2: the model after replacement of $X \leftrightarrow Y$ for $X \rightarrow Y$;
$\mathbf{P} = (P_1,...,P_m)'$: the vector of common explanatory variables (parents) of *X* and *Y*;
$\mathbf{Q} = (Q_1,...,Q_n)'$: the vector of additional explanatory variables (parents) of *Y* $(m,n \geq 0)$.

Every limited block recursive model can in principle be decomposed into three blocks. They are the *preceding* block, *focal* block, and *succeeding* block. So, M_1 can be decomposed into a preceding block (PB) with variables $\mathbf{V}_p$, a focal block (FB) with $\mathbf{V}_f (\equiv (X,Y))$, and a succeeding block (SB) with $\mathbf{V}_s$. Several assumptions are made about M_1:

(I) The relations across $\mathbf{V}_p$, $\mathbf{V}_f$ and $\mathbf{V}_s$ are recursive.
(II) The relations within the focal block $\mathbf{V}_f$ are only recursive.
(III) M_1 is identified

Thus, M_1 can be stated as:

$$\begin{aligned}
\mathbf{V}_p &= \mathbf{A}_{pp}\mathbf{V}_p + \mathbf{E}_p,\\
X &= \mathbf{a}'\mathbf{P} + u,\\
Y &= \mathbf{b}'\mathbf{P} + \mathbf{c}'\mathbf{Q} + \lambda X + v\\
&= (\mathbf{b}' + \lambda.\mathbf{a}')\mathbf{P} + \mathbf{c}'\mathbf{Q} + (\lambda u + v), \quad \lambda \neq 0\\
\mathbf{V}_s &= \mathbf{A}_{ps}\mathbf{V}_p + \mathbf{K}\mathbf{V}_f + \mathbf{L}\mathbf{V}_s + \mathbf{E}_s.
\end{aligned} \tag{F1}$$

where

- $\mathbf{A}_{pp}$ is a $p \times p$ matrix containing all regression coefficients in the PB;
- *a* and *b* are $m \times 1$ vectors containing the partial regression coefficients of *X* and *Y* upon the common explanatory variables of *X* and *Y*;
- *c* is an $n \times 1$ vector containing the partial regression coefficients of *Y* on its additional explanatory variables;
- $\mathbf{A}_{ps}$ is the coefficient matrix relating the SB-variables to the PB variables;
- *K* contains two columns, representing the coefficients of *X* and *Y*, relating the SB-variables to *X* and *Y*;
- *L* is a coefficient matrix relating the SB-variables to each other; and
- *u* and *v* are uncorrelated.

Model M_2 obtained by replacing $X \leftrightarrow Y$ for $X \rightarrow Y$ is defined as

$$\begin{aligned}
&\mathbf{V}_p = \mathbf{A}_{pp}\mathbf{V}_p + \mathbf{E}_p, \\
&X = \mathbf{a}'\mathbf{P} + u, \\
&Y = \mathbf{B}'\mathbf{P} + \mathbf{c}'\mathbf{Q} + w \\
&\mathbf{V}_s = \mathbf{A}_{ps}\mathbf{V}_p + \mathbf{K}\mathbf{V}_f + \mathbf{L}\mathbf{V}_s + \mathbf{E}_s.
\end{aligned} \qquad \text{(F2)}$$

u and w are no longer assumed to be uncorrelated. Note that the replacement leaves all the equations expect the equation for Y unchanged, and in this equation nothing has changed regarding the variables in Q, which do not enter into the equation for X. Before showing that M_1 and M_2 are equivalent, it is useful to state some rules for calculating the required covariance matrices.

2. Simple rules of covariance algebra (Bollen, 1989):

(I) For any random variable X with finite second-order moment,

$$Cov(X,X) = Var(X)\text{, and}$$

(I) For any random variables X, Y, Z and U with finite second-order moments, and any real numbers a, b, c, and d,

$$\begin{aligned}
&Cov(aX + bY, cZ + dU) = \\
&acCov(X,Z) + adCov(X,U) + bcCov(Y,Z) + bdCov(Y,U).
\end{aligned}$$

To establish that M_1 and M_2 are equivalent it must be shown that there is a surjective transformation vector function g, mapping every element of Θ onto Θ^*, and satisfies the $\sum$-condition. Since the replacement of $X \rightarrow Y$ with $X \leftrightarrow Y$ leaves all the elements of the parameter vector θ for M_1 unchanged except $(b_1,...,b_m)$, λ and σ_{vv}, where σ_{vv} is the variance of v, one only needs to find a surjective mapping g for these parameters. For the rest of the elements in θ, the required mappings are identity functions. To define the transformation g for the parameters changed by the replacement, the parameters of M_1 are held as fixed to define the corresponding parameters of M_2 as

$$\begin{aligned}
&B_1 = b_1 + \lambda a_1 \\
&B_2 = b_2 + \lambda a_2 \\
&\ldots \\
&B_m = b_m + \lambda a_m \\
&\sigma_{uw} = \lambda\sigma_{uu} \\
&\sigma_{ww} = \lambda^2\sigma_{uu} + \sigma_{vv}
\end{aligned} \qquad \text{(F3)}$$

With *g* thus defined, it remains to show that M_1 and M_2 satisfy the $\sum$-condition and $\sum^1_{pp}$at *g* is surjective. For model M_1, let

$\sum^1_{pp}$: the covariance matrix of the preceding block;
$\sum^1_{ff}$: the covariance matrix of the focal block;
$\sum^1_{ss}$: the covariance matrix of the succeeding block;
$\sum^1_{pf}$: the covariance matrix of the variables in preceding and focal block;
$\sum^1_{ps}$: the covariance matrix of the variables in preceding and succeeding block;
$\sum^1_{fs}$: the covariance matrix of the variables in the focal and succeeding block.

The covariance matrix implied by the model for parameter vector θ can be partitioned as:

Model 1

	$\mathbf{V}_p$	$\mathbf{V}_f$	$\mathbf{V}_s$
$\mathbf{V}_p$	$\sum^1_{pp}(\theta)$		
$\mathbf{V}_f$	$\sum^1_{fp}(\theta)$	$\sum^1_{ff}(\theta)$	
$\mathbf{V}_s$	$\sum^1_{sp}(\theta)$	$\sum^1_{sf}(\theta)$	$\sum^1_{ss}(\theta)$

Similarly, the covariance matrix implied by model M_2 for $\theta^* = g(\theta)$ can be partitioned as,

Model 2

	$\mathbf{V}_p$	$\mathbf{V}_f$	$\mathbf{V}_s$
$\mathbf{V}_p$	$\sum^2_{pp}[g(\theta)]$		
$\mathbf{V}_f$	$\sum^2_{fp}[g(\theta)]$	$\sum^1_{ff}[g(\theta)]$	
$\mathbf{V}_s$	$\sum^2_{sp}[g(\theta)]$	$\sum^2_{sf}[g(\theta)]$	$\sum^2_{ss}[g(\theta)]$

To establish the $\sum$-condition, it must be shown that:

$$\sum\nolimits_{ij}^{1}(\theta)=\sum\nolimits_{ij}^{2}[\mathbf{g}(\theta)],\quad (i,j=s,f,p) \tag{F4}$$

The transformation $\mathbf{g}:\Theta\rightarrow\Theta^{*}$, defined by equation (F3) leaves $\sum^{1}{}_{pp}$, $\sum^{1}{}_{ss}$, $\sum^{1}{}_{fs}$, and $\sum^{1}{}_{ps}$ unchanged. For these matrices, equation (F4) is trivially true. It remains to show that

(i) $\sum\nolimits_{ff}^{1}(\theta)=\sum\nolimits_{ff}^{2}[\mathbf{g}(\theta)]$;
(ii) $\sum\nolimits_{fp}^{1}(\theta)=\sum\nolimits_{fp}^{2}[\mathbf{g}(\theta)]$.

The process of establishing (i) and (ii) is similar. So, we describe the steps in establishing (ii). Let V_{pi} be any variable from the preceding block $\mathbf{V}_p$. Since the equation for X in both models is the same, the covariance of X with V_{pi} remains unchanged by $\boldsymbol{g}$. To establish (ii), it is therefore enough to show that the covariance of Y with each V_{pi} in $\mathbf{V}_p$ satisfies the $\sum$-condition.

Let $\mathbf{A}_{pp}(i)$ be the row of coefficients in the coefficient matrix $\mathbf{A}_{pp}$ relating V_{pi} to its predictors. Using the covariance rules (I) and (II), for any V_{pi} in $\mathbf{V}_P$ in model M_1 we have

$$Cov(Y,V_{pi})=\mathbf{A}_{pp}(i)Cov(\mathbf{V}_p,\mathbf{P})(\mathbf{b}'+\lambda.\mathbf{a}')+\mathbf{A}_{pp}(i)Cov(\mathbf{V}_p,\mathbf{Q})\mathbf{c}. \tag{F5}$$

Applying transformation $\boldsymbol{g}$, defined by (F3), to the right hand side of this equation yields

$$\mathbf{A}_{pp}(i)Cov(\mathbf{V}_p,\mathbf{P})\mathbf{B}+\mathbf{A}_{pp}(i)Cov(\mathbf{V}_p,\mathbf{Q})\mathbf{c}. \tag{F6}$$

Applying the covariance rules (I) and (II) to model M_2 to compute the covariance of Y with any variable V_{pi} in $\mathbf{V}_P$ yields

$$\mathbf{A}_{pp}(i)Cov(\mathbf{V}_p,\mathbf{P})\mathbf{B}+\mathbf{A}_{pp}(i)Cov(\mathbf{V}_p,\mathbf{Q})\mathbf{c}, \tag{F7}$$

which is identical with (F6). So, condition (ii) holds. A similar reasoning can be used to establish (i). The two models satisfy the $\sum$-condition. It remains to show that $\boldsymbol{g}$ is surjective.

Recall in equation (F3), we hold the parameters of the original model M_1 fixed to define the parameters of the transformed model M_2. To establish the surjectivity of

g, the parameters of model M_2 is held fixed in order to define the parameters of model M_1, which yields

$$\begin{aligned}
&b_1 = B_1 - (\sigma_{uw}/\sigma_{uu})a_1 \\
&b_2 = B_2 - (\sigma_{uw}/\sigma_{uu})a_2 \\
&\ldots \\
&b_m = B_m - (\sigma_{uw}/\sigma_{uu})a_m \\
&\sigma_{vv} = \sigma_{ww} - (\sigma_{uw}/\sigma_{uu})^2\sigma_{uu} \\
&\lambda = \sigma_{uw}/\sigma_{uu}
\end{aligned} \tag{F9}$$

The surjectivity of g is established by deriving the covariance $Cov(Y,V_{pi})$ of each variable V_{pi} in $\mathbf{V}_P$ of model M_2 using rules (I) and (II), restating the result using equation (F9), and checking that the result is the same as the one obtained by applying the rules to M_1 to compute $Cov(Y,V_{pi})$ for each V_{pi} in $\mathbf{V}_P$. This will show that the two models are equivalent.

The proof of the second part of the theorem follows a similar path, with the difference that we start with model M_2. To define g, the parameters of M_2 are held fixed and the parameters of M_1 are accordingly derived (as done in F9). It is then shown that the implied covariance matrices are the same. Raykov et al.'s method is quite general. It can be used for checking the Markovian equivalence of any two structural models.

Appendix H: Examples from MacCallum et al. (1993)

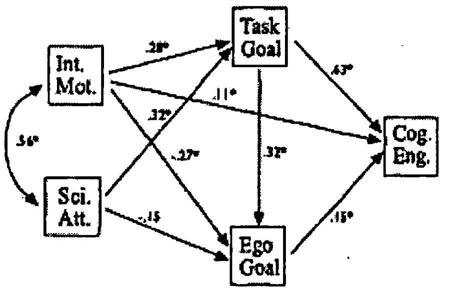
Two sets of equivalent models are reproduced from MacCallum et al. (1993). The first illustrates how the significance of a parameter estimate for a path between two variables varies across equivalent models. The second illustrates how the sign of a parameter estimate varies across the models.

Model I

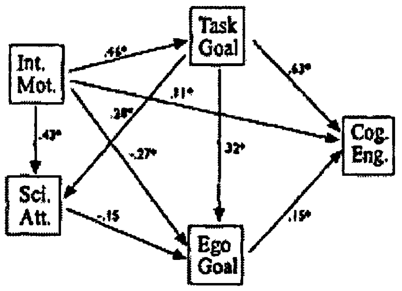
MacCallum et al. narrate this model from a study by Meece, Blumenfeld, and Hoyle (1988).

In the original model, the exogenous variable of *intrinsic motivation* expresses the degree to which a student's learning is characterized by intrinsic interest whereas the exogenous variable of *science attitudes* reflects the extent of a student's interest and enjoyment in science. *Task mastery goals* and *ego-social goals* represent differing goal orientations in achievement situations. The primary aim of task mastery goals is to independently master the task at hand, whereas the primary aim of ego-social goals is to demonstrate high ability socially. Finally, *active cognitive engagement* represents a student's use of cognitive strategies representative of self-regulated learning (e.g., reviewing difficult material).[36]

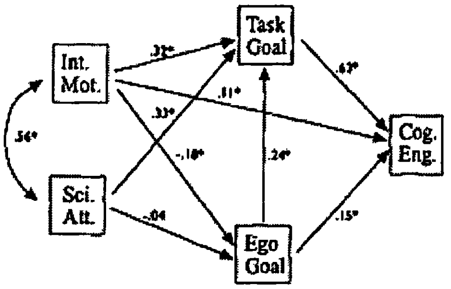
Original Model



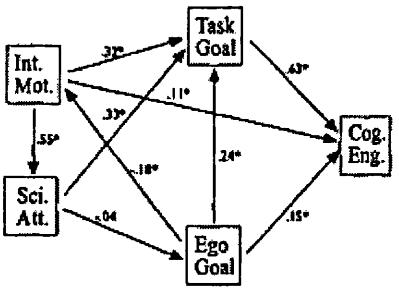
Model 2B



Model 2C



Model 2D



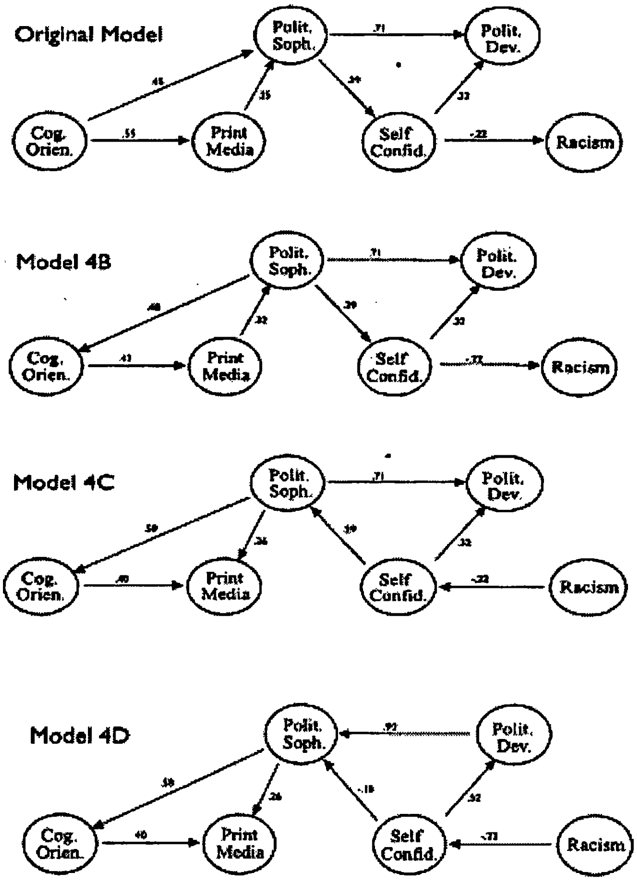
The significance of the coefficient path from *science attitude* to *ego-social goal* varies across the models.

Model II

MacCallum et al. narrate this model from a study by Sidanius (1988) in the field of social and personality psychology.

[36] These definitions are exactly quoted from MacCallum et al. (1993:191)

The firs variable in the original model is *cognitive orientation*, representing the interest a person has in politics and the importance he places in understanding politics. *Print media usage* reflects the degree to which people make use of the print media in obtaining political information. *Political sophistication* represents one's knowledge of politics and the complexity with which one thinks about politics. *Political deviance* represents a person's level of political extremity, and *self-confidence* reflects overall global confidence. Finally, *racism* is the extent to which an individual holds racist attitudes and beliefs towards ethnic groups.[37]



Original Model

Model 4B

Model 4C

Model 4D

Compare the sign of the parameter estimate for the path between *political sophistication* and *self-confidence* in the Original Model and Model 4D.

[37] These definitions are exactly quoted from MacCallum et al. (1993:194).

Chapter 6

The Economy as an Interactive System

An Appraisal of the Microfoundations Project

1 Introduction

> In order to have a useful theory of relations among aggregate, it is necessary that they be defined in a manner derived from the theory of individual behaviour. In other words, even the definition of such magnitudes as national income cannot be undertaken without a previous theoretical understanding of the underlying individual phenomena (Arrow, 1968).

This thesis began with stating some of the difficulties standing in the way of establishing a structural model of the economy: Firstly, for social and practical reasons, the economy cannot be subjected to controlled experiments in order to establish causal relations true of the variables representing the state of the economy. Secondly, because of the theoretical difference between a causal and statistical relation, the causal structure cannot be discovered by atheoretical statistical analysis of aggregate data. And finally, aggregate data are inherently imprecise, a fact that aggravates the difficulties in drawing causal inferences from observational data. The key to achieving the objectives of macroeconomics is to find a way around these difficulties, which make establishment of causal relations at the economy level problematic.

These difficulties, according to mainstream economists, can be evaded by beginning with a model of individual behaviour. It is argued that we often intuitively know how human beings make decisions, and even if intuition fails to lead us to the laws of behaviour, we can study human behaviour in an experimental environment to establish an empirically accurate theory of behaviour. Having established a theory of individual behaviour, we can transform it into a theory of the economy using aggregation procedures. Aggregate data can then be used to estimate the model and obtain a quantitative model of the economy. Since the structure is determined by the laws of behaviour and the model is based on behavioural laws, it correctly describes the economy. Specifically, it describes how aggregate variables relate to each other, classifies them into exogenous and endogenous categories, defines the conditions under which the aggregate equations remain invariant, and fixes the interpretation of the aggregate model parameters. So, the model provides all the information necessary for policy analysis.

The enterprise of deriving the correct macroeconomic theory from microeconomic theory, termed the *microfoundations project*, is the hallmark of modern theoretical macroeconomics. Two central assumptions underlie the project. One is that there exists or it is possible to establish an empirically adequate theory of microeconomic behaviour. The other is that microeconomic theory can be turned into a theory of the economy using aggregation procedures, without having to introduce any substantive assumption about the economic structure. The last four chapters studied some of the commonly accepted tenets in economics about individual behaviour. This chapter takes up the second hypothesis which has to do with the move from the micro to the macro level.

The search for microfoundations is the concern of all those economists who think that macroeconomics is something more than the art of summarising data and can establish structural models suitable for analysis of policies. New classical and Keynesian economists have both searched for microfoundations. Despite this, most systematic attempts to found models of large-scale economic phenomena on assumptions about individual behaviour have taken place in new classical economics. As a result, this chapter confines itself to an analysis of the efforts made in new classical economics. Nevertheless, since the analysis has to do with the general issue of moving from individualistic assumptions to a theory of the economy, it is equally relevant to any attempt at deriving a macroeconomic theory from a microeconomic theory.

There is more than one view of microeconomic theory in new classical economics. According to one vision of microeconomics, the basic unit of economic analysis is a single decision-maker, either a consumer or a firm. The consumer is modelled as an expected utility maximizer and the firm as an expected profit optimiser. On this account, a call for microfoundations is a call for a model of the economy in which the starting point is an expected utility or profit maximisation problem. To model some aspect of the economy, a utility (or profit) maximisation problem is set up for an individual and solved subject to his budget constraint in order to derive a model of the micro variables of interest. The model is then elevated to the economy level. This approach is referred to as the "representative agent" modelling approach to macroeconomics.

An alternative view of microeconomic theory in new classical economics is presented by the Walrasian general equilibrium theory in which the decision problems of various sectors of the economy, each represented by a representative agent, are simultaneously solved. To account for uncertainty about future, the theory is supplemented by the rational expectations hypothesis. From this perspective, the microfoundations project is an attempt to derive the laws of the economy from the assumptions of the Walrasian theory and the rational expectations hypothesis. Since the Walrasian theory makes minimal assumptions about the structure of the economy, this account of the microfoundations thesis is known as the strict microfoundations thesis (Rizvi, 1994:357).

This chapter criticises both interpretations of the microfoundations thesis. In a nutshell, the representative agent modelling approach conceives of the economy as a society of identical isolated individuals. And the strict microfoundations approach conceives of the economy as a collection of few sectors, each being populated by identical decision makers, who only interact through equilibrium prices. However, most macroeconomic phenomena arise from informational differences, behavioural heterogeneities, coordination failures, and interactions among market participants. Any satisfactory explanation of macroeconomic phenomena, therefore, calls for thinking of the economy as a society of heterogeneous interactive individuals. In such a society, the relations true of the aggregates are fundamentally different from those true of the micro variables, and there is no way that the former can be derived from the latter alone. Besides the microeconomic relations, one also needs to know a great deal about the structure of the society in order to derive the correct form of the aggregate relations.

The rest of this chapter is organised as follows: Section 2 characterizes the structure of the representative agent modelling approach, describes a representative agent-based macroeconomic model, and studies the conditions under which the behaviour of a collection of individuals can be modelled in terms of the behaviour of a single agent. It then explains why the modelling approach is inappropriate for studying the economy. Section 3 takes up aggregation issues arising from individual heterogeneity. In particular, it studies the theory of exact aggregation to examine how far one can go in accounting for individual heterogeneity while preserving a simple, and manageable, relation between the laws of micro variables and those of macro variables. Section 4 first discusses the strict microfoundations thesis and then examines some of the issues that interaction creates for the microfoundations project. Section 5 concludes the chapter.

2 The Representative Agent Modelling Approach

Most modern economies consist of millions of decision makers, either as individuals or organised groups, each pursuing their own disparate interest in a limited part of the economy. These individual and group activities are somehow coordinated, leading to certain regularities at the economy level, which form the subject matter of macroeconomics. If we were in a position to simultaneously study the behaviour of every decision-making unit in the economy and model its interaction with other decision-making units, we would be able to predict the emergence of macroeconomic regularities by simulating the evolution of the economy. However, we are not omniscient and this avenue is closed to us. All the same, many individuals or groups often encounter similar choice situations, have similar tastes and demographic characteristics, and behave similarly. Moreover, individual idiosyncratic differences sometimes neutralise each other in the real life. A satisfactory understanding of the economy does not then necessarily require simulating the whole system including the details of each decision-making unit. It is sufficient to work with an idealised, smaller, model economy in which the behaviour of each group of 'similar' decision-making units is represented by an average unit (agent). Some economists, like Jevons, have taken this consideration to an extreme. According to Jevons, "accidental and disturbing causes will operate, in the long run, as often in one direction as the other, so as to neutralize each other." Thus, "the general forms of the laws of economics are the same in the case of individuals and the nations" (Jevons, 1965 [1871]:16-17).[1] Hicks has even gone further to suggest that microeconomic theory has greater relevance for aggregate data, arguing that the variations in circumstances of individual households are averaged out to negligible proportions in the aggregate, leaving only systematic effects of variation in prices and budgets (Hicks, 1956).[2] Such thoughts have led to the emergence of a modelling approach that views the economy as a single average individual, implying that whatever is true of the individual is also true of the economy, hence the nomenclature of representative agent modelling.

2.1 The Structure of the Representative Agent Approach

In new classical economics, two central assumptions regarding individual behaviour are the expected utility optimisation hypothesis (Friedman and Savage, 1952) and the rational expectations hypothesis (Lucas and Sargent, 1979). In this school of thought, the point of departure in building a representative agent model is to specify the optimisation problem of an agent (a household or a firm) and solve it subject to his budget constraint and rational expectations. The solution yields the individual decision rules and the relationships among the individual variables. In the second stage, the well-defined *individual* model is taken to be exactly true at the aggregate level. Aggregate variables are inserted into the model to transform it into a qualitative model of the economy or some aspect of it. In the final stage, aggregate data are used to estimate the model parameters, turning it into a quantitative model of the economy. If the model fits aggregate data, the conformity is taken as evidence for the truth, or at least, the empirical adequacy of the microeconomic model. If it does not fit the data, the blame is placed on the individual assumptions built into the model, like the form of the utility function or the variables entering it. In this manner, the new classical representative agent methodology seeks to meet all the challenges of macroeconomic modelling. It aims to specify the form of the relations linking aggregate variables, the conditions under which the model equations remain invariant, and the proper

[1] See also Marshall, 1890 [1961:174].

[2] See Deaton et al. 1981:149.

interpretation of the macro model parameters. On this interpretation of the microfoundations thesis, only those macro models that are grounded on utility optimisation subject to rational expectations are regarded as acceptable for policy evaluation.

2.2 A Historical Example

Before proceeding to study the requirements of the representative-agent modelling methodology, we study a classic representative agent model that has been the source of many controversies and insights in the recent history of macroeconomics. The study will later help us bring to the fore various assumptions underpinning a representative agent model. [3] An issue in macroeconomics concerns the relation between aggregate consumption and aggregate income. Several empirical studies during the third quarter of the last century implied that aggregate income was a good predictor of current aggregate consumption (Blanchard et al., 1989, chapter 6). This result seemed to contradict the idea that people form expectations rationally, and make their consumption decisions according to the permanent income hypothesis. In a classic paper, Robert Hall (1978) set out to shed light on this issue by testing the implications of the permanent income hypothesis for aggregate income and consumption data. He did this by following the representative-agent modelling method.[4]

According to the permanent income hypothesis, a household chooses how much to spend at time t as part of a plan that takes into account future uncertainty in income by optimising over time with regard to available wealth.[5] To make this idea precise, let r be the real rate, T the length of economic life, and $u_i(.)$ a strictly concave one period utility function. Furthermore, let C_{it} be consumption by consumer i in period t, Y_{it} income in period t, A_{it} assets apart from human

[3] A thorough analysis of the representative agent modelling approach is given in Hartley (1997).

[4] A discussion of Hall's methodology is given in Granger (1999:42-48).

[5] Here, the permanent income hypothesis is taken to be the idea that permanent income is the annuity value of current financial and human wealth, and consumption is equal to permanent income.

capital, and δ the consumer's rate of subjective time preference so that £1 now and £$(1+\delta)$1 next period are equally valued. The permanent income hypothesis says that, in each period t, family i decides on its consumption plan by maximising the expected lifetime utility

$$E_t \sum_{\tau=0}^{T-t} (1+\delta)^{-\tau} u_i(C_{it+\tau}), \tag{2.1}$$

subject to the amount of available wealth

$$\sum_{\tau=0}^{T-\tau} (1+r)^{-\tau} (C_{it+\tau} - Y_{it+\tau}) = A_{it} \,. \tag{2.2}$$

E_t in (2.1) denotes mathematical expectation conditional on all information available at t including $C_{it-\tau}$, $Y_{it-\tau}$, and $A_{it-\tau}$, for $\tau = 0,1,2,\ldots$. Hall also assumes that the real rate of interest r is constant, the subjective rate of time preference δ is equal or less than r, incomes Y_{it} are stochastic and are the only source of uncertainty, and lets T go to infinity. The first-order necessary condition for maximisation of equation (2.1) subject to constraint (2.2) is the famous Euler equation

$$E_t u_i'(C_{it+1}) = [(1+\delta)/(1+r)] u_i'(C_{it}), \tag{2.3}$$

where $u'(C) = du(C)/dC$. This equation says that the expected marginal utility next period is the same as marginal utility this period, except for a trend associated with the rate of time preference δ and the real rate of interest r. Another way to express the same idea is

$$u_i'(C_{it+1}) = \gamma u_i'(C_{it}) + \varepsilon_{it+1}, \tag{2.4}$$

where $\gamma = (1+\delta)/(1+r)$ and ε_{it+1} denotes the difference between the marginal utility next period and its current expected value. Assuming that expectations are rational, ε_{t+1} is a random variable with expected value zero at time t, when consumption C_{it} is decided. Equation (2.4) implies that no information available in period t apart from the level of consumption C_{it} helps predict future consumption C_{it+1}. Once C_{it} is taken into account, individual income and assets at time t or earlier and past consumptions C_{it-j}, for $j > 0$, are irrelevant for predicting the next period marginal utility.

Hall simplifies matters by assuming the utility function $u_i(.)$ is quadratic, $u_i(C_{it}) = -(\overline{C_i} - C_{it})^2/2$, where $\overline{C}$ is the bliss level of consumption.[6] This leads to the individual consumption function:

$$C_{it+1} = \lambda C_{it} + \varepsilon_{it+1}. \tag{2.5}$$

According to this equation, the change in individual consumption is the amount warranted by innovations in expectations about future labour income. Formally, this means that individual consumption obeys a random walk.[7] So, no other variable observed in period t or earlier will have a nonzero coefficient when added to the equation.

Hall next assumes that if individual consumption exhibits random walk behaviour, aggregate consumption also by and large mimics random walk behaviour. Therefore, if the above assumptions are approximately true of a typical household, the equation,

$$C_{t+1} = \lambda C_t + \varepsilon_{t+1}, \tag{2.6}$$

[6] A bliss utility level is a level beyond which the marginal utility of consumption is negative (Deaton, 1992:179). Note that equation (2.5) is based on the assumption that δ equals r; otherwise, the equation includes an intercept.

[7] A random walk sequence is an example of a martingale sequence. A sequence Z_t is a martingale if $E[Z_t / Z_{t-1}, Z_{t-2},...] = Z_{t-1}$. Z_t is then a random walk if $Z_t = Z_{t-1} + u_t$ where $Cov(u_t, u_s) = 0$ for all $t \neq s$.

provides a good approximation of the behaviour of aggregate consumption C_t. The permanent income hypothesis, Hall concludes, rules out any systematic influence of any variable on current aggregate consumption other than last period aggregate consumption.[8] Equation (2.6) can thus be tested by embedding it in a wider model such as

$$C_{t+1} = \lambda C_t + \alpha C_{t-1} + \beta Y_t + e_{t+1},$$

and checking the significance of the coefficients α and β. Hall tested equation (2.6) by regressing aggregate consumption changes on lags of aggregate consumption, income and stock prices. He found little predictive power for income but rejected nonpredictability for stock prices. He concluded that while data on income confirm the permanent income hypothesis, the data on stock prices disconfirm it (Hall, 1989:157). Flavin (1981) also studied the relation between aggregate income and consumption in a similar setting and argued that there was enough predictive power for aggregate income to reject the permanent income hypothesis.

2.3 The Requirements of the Representative Agent Approach

Hall's analysis is a typical example of the representative agent modelling approach to the study of large-scale economic phenomena. An analysis of this approach calls for addressing three related issues: The first is concerned with the conditions under which the behaviour of a collection of individuals can be modelled as the behaviour of a single individual. The second is related to the plausibility of the conditions. And the third, and in fact the most important, issue has to do with the usefulness of the representative agent models for understanding large-scale economic phenomena. We begin with the first query.

Significant decisions always involve uncertainty and, as a result, one has to work, as done by Hall, with a dynamic model of individual behaviour. Nevertheless, it is

[8] Hall's exercise is an example of testing for non-Granger causality (Sargent, 1987:94).

convenient to first study the conditions required for the existence of a representative agent in a static setting, and then investigate the additional conditions that may be needed in a dynamic setting. Consider an economy of n consumers and m goods. Each individual i has utility function $u_i(.)$, income (expenditure) X_{it} at time t, and demands $\mathbf{Y}_{it} = (Y_{it1},...,Y_{itm})$ for m goods at time t. Further, suppose everyone in the economy faces the common price vector $\mathbf{P}_t = (P_{t1},...,P_{tm})$.[9] Each agent i maximises his utility subject to his budget constraint, arriving at the individual consumption function:

$$\mathbf{Y}_{it} = f_i(X_{it}, \mathbf{P}_t). \tag{2.7}$$

The aggregate demand of m goods will be

$$\mathbf{Y}_t = \sum\nolimits_i f_i(X_{it}, \mathbf{P}_t) = G(X_{1t},...,X_{nt}, \mathbf{P}_t) \tag{2.8}$$

where $\mathbf{Y}_t = \sum_i Y_{it}$. Finally, let $\mathbf{X}_t = \sum_i X_{it}$ denote aggregate expenditure. Our question regarding the circumstances under which a representative agent exists has two different components. The first concerns the conditions under which there exists a macro function $F(\mathbf{X}_t, \mathbf{P}_t)$ such that

$$\mathbf{Y}_t = G(X_{1t},...,X_{nt}, \mathbf{P}_t) = F(\mathbf{X}_t, \mathbf{P}_t). \tag{2.9}$$

The second relates to the conditions under which the aggregate consumption function $F(\mathbf{X}_t, \mathbf{P}_t)$ can be derived from maximization of a utility function subject to total income $\mathbf{X}_t$ and price vector $\mathbf{P}_t$. Before addressing these questions, note that the setting is quite general in the sense that the individual function f_i can take any form and the variables X_{it} and Y_{it} can be interpreted in different ways. For instance, as in Hall's model, Y_{it} can be current consumption and X_{it} lagged

[9] A weighted sum of the individual demand functions with each function multiplied by the price of the corresponding commodity is equal to expenditure $\sum p_m f_{mi} = x_i$.

consumption. To preserve consistency, for the time being, we take Y_{it} to be consumption and X_{it} income.

Gorman (1953) establishes the necessary and sufficient conditions for the existence of a macro function of the form $F(\mathbf{X}_t, \mathbf{P}_t)$ in a static setting. Theorem (2.1) states these conditions:

Theorem 2.1: Aggregate consumption function (2.9) exists if and only if the individual demand functions (2.7) take the form:

$$\mathbf{Y}_{it} = a_i(\mathbf{P}_t) + b(\mathbf{P}_t)X_{it} \tag{2.10}$$

that is, if and only if the individual demand functions are (i) linear in income and (ii) are identical up to the addition of a term that depends only on the common price vector (Gorman, 1953).[10]

Demand function (2.10), known as the *Gorman polar form*, restricts individual differences to the intercept term $a_i(\mathbf{P})$, requiring the slope term to be common to all the consumers.[11] If the *adding up* condition, $\mathbf{Y}_{it}.\mathbf{P}_t = X_{it}$, is imposed, it follows that $a_i(\mathbf{P}_t).\mathbf{P}_t = 0$ and $b(\mathbf{P}_t).\mathbf{P}_t = 1$. Whenever individual demand equations take the Gorman polar form (2.10), the aggregate demand function can immediately be derived as

$$\mathbf{Y}_t = \sum_i a_i(\mathbf{P}_t) + b(\mathbf{P}_t)\sum_i X_{it}\,. \tag{2.11}$$

Gorman's theorem requires the individual demand functions to be linear in income, which means the proportion of income spent by a person on consumption is independent of the size of his income; he spends the same portion of his income on goods regardless of how large his income grows. The theorem also necessitates identical marginal propensities to consume. That is, the income proportion that Bill Gates spends on each good should be same as the income proportion that a

[10] For a simple statement of Gorman's proof see Brighi (1989:5).

[11] There is a vast literature on the requirements of a representative agent, including Antonelli (1886); Deaton and Muellbauer (1980); Gorman (1953); Green (1964); Heineke and Scheffrin (1990); Jorgenson, *et al.* (1982); Lau (1977, 1982); Lewbel (1989); Muellbauer (1975, 1976); Nataf (1948); and Stoker (1984, 1993).

person in a poor corner of the States spends on the good. These requirements, taken together, entail that an aggregate consumption equation of the form (2.9) exists if and only if total consumption is independent of the distribution of income in the economy. If there were two groups of households with different marginal reactions to income changes, a transfer of income from one group to the other would alter total consumption. In that case, there would be distributional effects that are not accounted for by total income.

As an illustration, consider an economy consisting of one rich family and three poor families. The rich household receives £50 per month and spends 5% of its income on food. Each poor family receives £10 per month and spends 25% of its income on food. Aggregate monthly expenditure on food in the economy is £10. A transfer of £5 from each poor household to the rich reduces total food expenditure to £7. However, if the same amount, i.e., £15, is taken from the rich and evenly distributed among the poor households, aggregate expenditure rises to £13, even though aggregate income in either case is the same. What effect does an increase of £10 in total income have on total expenditure? Again, it all depends on who gets the income. If the rich receives the extra income, total expenditure changes by 50 pence. But if any of the poor families receives the extra income, aggregate expenditure rises by £2.5. The point is that, with different marginal responses, knowledge of total income is not sufficient to determine total consumption.

Gorman (1953, 1961) also establishes the conditions under which the aggregate equation $F(\mathbf{X}_t, \mathbf{P}_t)$ is derivable from maximization of a utility function. The result draws on two technical notions. The first is the notion of homotheticity. A monotone preference relation $\geq$ on a choice set $\mathbf{X} \subseteq \mathbf{R}_+^L$ is called *homothetic* just in case $\mathbf{x} \geq \mathbf{y} \Leftrightarrow \alpha\mathbf{x} \geq \alpha\mathbf{y}$ for all $\alpha > 0$.[12] Homothetic preferences can be represented by a monotonic transformation of a homogenous of degree 1 utility function. The second is the notion of *integrability.* An individual demand function is called integrable if it can be generated by maximisation of a utility function

[12] Here, homothetic preferences mean that the agent always spends a fixed proportion of his or her income on each good (Kirman, 1989:132).

subject to a budget constraint. Having stated these preliminaries, Gorman's answer to the second question can be stated as follows:

Theorem 2.2 (Gorman 1953, Nataf, 1948): Suppose the individual demand function (2.10) is integrable; that is, it can be derived from maximization of a utility function $u(.)$. Then, aggregate demand function $F(\mathbf{X}_t, \mathbf{P}_t)$ exists and is integrable if and only if $u(.)$ is a homothetic utility function (See Shafer et al., 1982, for a proof).[13]

Market demand function can be interpreted as a consumer demand function if and only if each individual demand function f_i is derived from a homothetic utility function $u(.)$ common to all the consumers. In that case, for all *i*, $F = f_i$.[14] A failure of homotheticity makes individual marginal propensity to consume dependant on the individual income level, which in turn renders total consumption dependant on the income distribution in the society. As an illustration, following Shafer et al. (1982), consider an economy with two goods and two consumers who have identical but non-homothetic preferences represented by $u(x, y) = xy + y$. Let the prices be (1,1). An income distribution of $m_1 = £1$ and $m_2 = £1$ leads to a different demand than an income distribution of $m_1 = 2$ and $m_2 = 0$. In the first case total demand for *y* is £2 and for *x* is zero whereas in the second case total demand for *y* is £3/2 and for *x* is £0.5. Gorman also shows that if households receive strictly non-negative incomes, the homotheticity requirement can be replaced with quasi-homotheticity.[15]

2.3.1 Identical Marginal Propensity throughout Time

Theorem (2.1) and (2.2) provide the conditions for the existence of a representative agent in a static setting. As one moves to a dynamic setting, the

[13] Also see Brighi et al., (1989), Appendix 1.

[14] An assumption underlying the Gorman result is the restriction of zero expenditure at zero income.

[15] Quasi homothetic preferences generalise homothetic preferences. Homothetic preferences imply Engle curves that are linear and pass through the origin. Whereas, quasi homothetic preferences allow vertical non-zero intercepts, leading to Engle curves that do not necessarily pass through the origin. A utility function creating such Engle curves is called quasi-homothetic. Engle curves describe demand as a function of income.

existence of a representative agent calls for further conditions. To explore these conditions, note that theorem (2.1) requires the slope function $b(p)$ to be independent of the individual income level. This condition necessitates identical marginal propensity to consume over time (i.e., throughout life), regardless of whether one is young, employed, or retired. Hall introduces this condition into his model economy by assuming that people live an infinite life, which means they do not need to worry about their future income.

Clarida (1990) abandons the assumption of infinite life span for individuals, noting that the propensity to consume declines monotonically with age (1991:854). He then studies the effect this presumption would have on aggregate consumption function and its potential for explaining empirical regularities discernible in aggregate data. Specifically, he considers a simple economy in which each consumer lives for *n* periods, earns income Y_t during *m* (*m*<*n*) working periods, and receives nothing during the retirement periods (*n*-*m*). Consumption during retirement is financed by saving a portion of labour income. Individual income Y_t follows a random walk with drift g:

$$Y_t = g + Y_{t-1} + \varepsilon_t \tag{2.12}$$

Further, the interest rate is zero and, as in Hall's economy, every one acts according to the life-cycle permanent income hypothesis. In this economy, even though individual consumption is a random walk, aggregate consumption is not a random walk. In fact, if *n* is taken to be three and *m* two, average consumption change in the economy follows:

$$\Delta\overline{C}_t = \overline{g} + \alpha\overline{\varepsilon}_t + \beta\overline{\varepsilon}_{t-1} + \gamma\overline{\varepsilon}_{t-2}, \tag{2.13}$$

where the sign '$\bar{\ }$' denotes average (Deaton, 1992:169). Appendix (*A*) explains the steps from (2.12) to (2.13). In this economy in which people have a finite life span, and face different levels of income during their life, average (aggregate) consumption is no longer orthogonal to lagged innovations; both parameters

β and γ are non-zero. Nor does it respond one for one to innovations in current income. The economy exhibits a correlation between consumption change and past income (known as 'excess sensitivity'), and the variance of consumption changes is much less than the variance of income changes (known as 'excess smoothness').[16] Therefore, unless one is prepared to assume that households live forever, in a dynamic setting the representative agent methodology not only requires the households to have identical marginal propensity to consume at any time but also to have identical marginal propensity to consume throughout time. Or else, aggregation can produce relations that are not representative of relations at the individual level.

2.3.2 Identical Aggregate and Individual Income Processes

Another requirement for the existence of a representative consumer in a dynamic setting is that individual income and aggregate income follow the same stochastic process. If different processes generate individual and aggregate income, and consumers lack full knowledge of the aggregate income process, aggregation over individual consumption functions can easily create a macro consumption function entirely different than the individual functions. Pischke (1995) was the first to note this requirement. He considers an economy similar to Hall's economy but supposes that individual and aggregate income follow different processes.[17] Specifically, he assumes that the average income in the economy follows a random walk with drift, i.e.,

$$\overline{Y}_t = g + \overline{Y}_{t-1} + \varepsilon_t . \tag{2.14}$$

He, however, takes individual income to be the average income plus an idiosyncratic component that is purely transitory, represented by a white noise,

$$Y_{it} = \overline{Y}_t + u_{it} , \tag{2.15}$$

[16] Since α is less than 1. See Appendix A.

[17] The statement here has draws on Deaton (1992)'s discussion of Pischke's paper.

where the innovations ε_t and u_{it} are uncorrelated. The first difference of individual income is the first difference of the random walk, including the drift term plus the first difference of the white noise term:

$$\Delta Y_{it} = g + \varepsilon_t + u_{it} - u_{it-1}, \tag{2.16}$$

The households, Pischke notes, are not in a position to infer the contemporaneous aggregate shock ε_t. As a result, they cannot separate the macro shock from the idiosyncratic component (private shock) u_{it}. Each individual can at best estimate the sum of the terms from the data, which amounts to estimating the moving average process:

$$\Delta Y_{it} = g + \eta_{it} - \lambda \eta_{it-1}. \tag{2.17}$$

Given this result and the conditions defining Hall's model, the change in individual consumption follows $\Delta C_{it} = (1 - \lambda / 1 + r)\eta_{it}$. Individual consumption is thus a random walk,

$$C_{it} = C_{it-1} + (1 - \lambda / 1 + r)\eta_{it}. \tag{2.18}$$

In contrast, aggregate consumption is not a random walk. It follows a second order autoregressive process (Appendix B),

$$C_t = (\lambda + 1)C_{t-1} - \lambda C_{t-2} + \varsigma_t, \tag{2.19}$$

where $\varsigma_t = (1 - \lambda / 1 + r)\varepsilon_t$. The difference would disappear if the households had knowledge of the history of aggregate income including current aggregate income Y_t *and* were able to infer the aggregate income process. This would enable them to separate the common contemporaneous shock ε_t from the private shock u_{it}. In that case, the aggregate and individual consumption functions would coincide (Pischke, 1995:809).

In a dynamic setting, then, for a representative consumer to exist the processes generating the individual and aggregate incomes should be the same. Or the individuals should have complete knowledge of the history of aggregate income to infer the aggregate income process. In fact, at a closer look, full knowledge of the history of the aggregate income is not enough. It must also be assumed that individuals with the same information make the same inferences (Grossman, et al., 1982). Otherwise, even with full knowledge of the history of aggregate income, they may infer different aggregate process, which can result in a difference between the individual and aggregate functions. So, in a dynamic setting, the representative agent methodology necessitates a variant of the Harsanyi doctrine that people with the same information always form the same probabilistic beliefs. Critical analysis of the theory of objective Bayesianism has shown fundamental flows in the Harsanyi doctrine and in objective Bayesianism in general (more on this in Chapter IV).[18] Moreover, information on the current aggregate variables is hardly available. Even the interested econometricians receive such information with a delay of a quarter or more. Also, there seems to be no rationale for the individuals to obtain such information. Gathering such information is often quite costly.

2.3.3 Absence of Interaction among Economic Agents

Gorman's result demands the parameters in the individual consumption functions be independent of the explanatory variables varying across the individuals. Since the aggregate function is derived by summing over the individual functions, the same condition must necessarily hold of the parameters of the aggregate function. This necessitates absence of any interaction among decision-maker units populating the economy. Whenever there are interdependencies, the parameters in the aggregate function become dependent on the explanatory variables varying across the individuals. The dependence makes the form of the aggregate function entirely different than the individual functions, and also makes it impossible to

[18] Goodfriend (1992) assumes that the agents observe aggregate income with one lag period and use this information to guess about contemporaneous income shock. Consumption change is then shown to follow an AR(1) process.

interpret the aggregate parameters in the same way as the individual parameters. To see this, consider Hall's model again. In setting up his model, Hall regards the real interest rate r as constant, thus assuming that it is independent of the (current) consumption level. The assumption is reflected in the individual consumption function (2.5), restated here as

$$C_{it+1} = \left[\frac{(1+\delta)}{(1+r)}\right] C_{it} + \varepsilon_{it+1}, \tag{2.20}$$

In this setting, the agent takes the interest rate as given in deciding how to allocate his income between consumption and saving. This is reasonable. If he saves a little bit more or less, his action won't affect the real interest rate. But, if everyone in the economy makes a similar decision, the real interest rate moves. If everybody saves less, the real interest rate rises, pushing asset prices down. Alternatively, if everybody saves more, the real interest rate falls, pushing asset prices up.[19] Contrary to Hall's assumption, aggregate consumption and the real interest rate do not move independently. The real interest rate depends on the consumption level and vice versa. One cannot hold one of these as constant and let the other vary. So, although in modelling individual consumption the real interest rate r can be considered as independent of the individual consumption level C_{it}, in modelling aggregate consumption the assumption that the real interest rate r is independent of the aggregate consumption level is untenable. It would be conceptually wrong to write the aggregate function as

$$C_{t+1} = \left[\frac{(1+\delta)}{(1+r)}\right] C_{t} + \varepsilon_{t+1}, \tag{2.21}$$

Since the interest rate depends on the aggregate consumption level, the relation between current and future aggregate consumption is nonlinear (Hartley, 1997:156). In fact, with interaction, significant differences between the micro and

[19] A similar discussion is found in Dow (1988:8), Leijonhufvud, (1968: 210-211) and Snowdon (1994:370).

macro consumption function do not end here. If everyone decides to save less, the decision increases the real interest rate, lowering the asset prices. This increases the opportunity cost of current consumption, thus moderating the increase in current consumption actually achieved. Alternatively, if everyone decides to save more, the decision lowers the real interest rate, pushing the asset prices up. This lowers the opportunity cost of current consumption, hence moderating the reduction in the current consumption actually achieved. Such endogenous fluctuations in the real interest rate and asset prices restrain intertemporal arrangement of consumption. The inhibition can create a tighter link between the future consumption and current income than is predicted by Hall's model, which abstracts from fluctuations in the interest rate and asset prices. So, even if (2.20) were true of the individual, the aggregate consumption might still include other variables besides current aggregate consumption.

A consequence of these considerations is that in an interactive system the behaviour of an aggregate variable cannot be modelled in isolation of the mechanisms generating the (independent) variables affecting the variable. In the above setting, this means that one cannot establish an adequate theoretical model of consumption without simultaneously modelling the mechanisms generating income, asset prices, and interest rate. Since aggregate consumption also influences these variables, the interdependencies necessitate a non-recursive model to account for the feedback. So, in an interactive system, although a recursive model may accurately describe individual consumption behaviour, to describe the behaviour of aggregate consumption, one may have to adopt a non-recursive (feedback) model.[20]

To sum up, the existence of a representative individual requires that the dependent variable in the micro functions be linear in the explanatory variables, the coefficient in the micro functions (except the intercept) be the same across the

[20] Aware of the interdependencies between consumption, income, and interest rates in the economy, Michner (1984) took issue with Hall's partial equilibrium approach, arguing for a general equilibrium approach to study aggregate consumption. He showed that, in the general equilibrium setting, the permanent income hypothesis no longer implies that aggregate consumption follows a random walk process. In his simple equilibrium model, aggregate consumption change in fact turned out to be a constant function of aggregate current income.

individuals, the coefficients be constant over time, the mechanisms generating the individual and aggregate explanatory variables be the same or the agents have full knowledge of the mechanisms generating the aggregate explanatory variables, and there be no interaction among the individuals. These assumptions are incredibly strong and, even as gross approximation, are hardly true of the modern economies.

2.5 Problems with the Representative Agent Modelling Approach

The analysis has so far been concerned with the requirements of the representative agent modelling approach. It has also pointed out that the assumptions are incredible, and not, even approximately, true of the modern economies. This is not, however, the only problem with the methodology. The more serious problem is that it is fundamentally unsuitable for studying the economy. There are several reasons behind this claim, briefly stated below:

First, the assumptions underlying the representative agent modelling approach give rise to a view of the economy as a society of essentially identical individuals operating in isolated homogenous choice situations. In such a society, there is no place for money that is a means of exchange among agents operating in different decision situations, with different needs, preferences, beliefs, and attitudes towards risk (Friedman, et al., 1990:xii). Nor does such a society provide any place for the institutions created around money. Monetary institutions are for co-ordinating among differently situated agents with different needs and beliefs, who do not exist in a society of identical individuals (David Colander, 1996:62). To give another example, if people had identical preferences, had access to identical information, entertained the same beliefs, and encountered identical choice situations, there would be no trade in securities. A society of identical individuals leaves no place for security markets. These markets arise because people have access to different information, make different inferences from the same data, have different attitudes towards risk, and so forth. Any attempt at explaining the emergence of security markets, their effects on the functioning of the economy, and the role of the institutions associated with them demands abandoning the straightjacket of the representative agent-based modelling methodology and taking individual heterogeneities seriously (Arrow, 1986:212). These considerations reveal that the difficulty with the representative agent modelling approach is not that it abstracts away certain aspects of the economy. Any modelling approach proceeds with abstraction and idealisation. The fundamental difficulty is that it abstracts away the very same features that are necessary for understanding basic economic phenomena.[21]

Second, an implication of the methodology is that every proposition true of the individual is true of the economy and every proposition true of the economy is true of the individual. This implication is wrong. In general, when one moves from the individual level to the economy level the causal status of the variables affected by individual decisions changes. Coffee scarcity is exogenous to one's decision but it is the people who altogether cause coffee scarcity (Schelling, 1978:78). Economic growth is exogenous to one's decision but it is the external effects of individuals' capital accumulation that causes it (Romer, 1994). Asset prices are exogenous to one's decision but it is the individuals' saving, consumption, and investment decisions that determine the prices (Lucas, 1978). Interest rate is exogenous to one's decision but it is the individuals, saving decisions that determine it. Coffee scarcity, the interest rate, asset prices, unemployment level, economic growth, and population density should be regarded as exogenous in modelling individual behaviour. In modelling the economy, however, it is the individual decisions that have to be considered as exogenous. Thus, it is wrong to think that if a variable is exogenous to the agent it is also exogenous to the economy or if a variable is endogenous to the economy it is also endogenous to the individual. Failure to recognise this point results in fallacious conclusions about the economy.

Theoretical differences between the individual and the economy are not limited to the changes in the status of the variables affected by individual decisions. There is a multitude of other types of propositions that apply to the individual but not the economy or apply to the economy but not the individual. As an illustration,

[21] For a list of other phenomena that cannot occur in a society of identical, entirely isolated, individuals, see Stiglitz, 1991.

consider an example from Schelling (1998), which concerns the pattern of sales of best-seller novels, fictions, and biographies by new unknown authors. Sales data show that the sales of such works in a society follow a logistic path, growing exponentially at first, then passing an inflection point, and finally declining exponentially until the leftover copies are remaindered. A possible explanation for this pattern, Schelling says, is the following. "People who read the book, if they like it, *they* talk about it, some people more than others; the more people who read the book, the more people there are to talk about it. Some of the people they talk to buy the book; if they like it, they talk about it. Talk is proportionate to the number of people who have read the book; if all talk is equally effective, the number talking about it grows exponentially. But there is a limit to the number of people likely to be recruited; eventually most of those who would be interested have already heard of the book, maybe bought it, and when they want to talk about it find that there's hardly anybody left who hasn't already heard about it. If there were initially L potentially interested readers, and N have now read it and want to talk about it, and everybody who has read it meets and talks about it with n out of the L per week, there will be $N \times n \times L$ contacts per week, with $N \times n \times (L-N)$ of them potentially productive, and N will grow logistically" (Schelling. 1998:34). The logistic curve discernable in the data on the sale of bestsellers by unknown authors cannot be attributed to a single individual. The key to the emergence of the curve is the finiteness of the number of interested readers living in a society. The curve is not crucially dependent on the particular decision making mechanism driving one to buy the book.[22] Similar patterns are also likely to emerge in sales data for newly invented durable goods.

Third, a further problem relates to the suitability of the representative agent models for policy analysis. Policies are usually designed to influence the economy by changing certain distributional aspects of it. Monetary policies, for instance, often operate by reducing the consumption of those who are facing liquidity constraints. The eventual effect of any such policy surely depends on the

[22] See Schelling (1978) for other categories of propositions that are true of a closed interactive system but not true of the behaviour of each person nor even, necessarily, of any groups smaller than the whole system (see especially around 1978:49). Also see Hartley (1997:148-9) for an example from monetary economics, due to Laidler (1982).

distribution of assets in the economy (Stiglitz, 1991:26). However, this contradicts the central presumption of the representative agent models that the value of the aggregate dependent variable (here, aggregate consumption) is independent of the distribution of the explanatory variables in the economy (here, income and assets). On these models, as long as a policy shift is limited to a change in the distribution of the explanatory variables, it has no effect at all on the dependent variable. If the possibility of influencing the economy through distributional channels is granted, then one has to look for models that do not simply deal with aggregates and are sensitive to the distributional features (Martel, 1996:140). Analysis of policies calls for knowing the joint distribution of the micro variables affecting decisions, predicting how a policy affects the distribution, and determining how the distributional change affects the state of the economy as a whole. None of these issues can be settled within the representative agent modelling framework.

3 Modelling Heterogeneous Behaviour

The representative agent paradigm is inadequate as a framework for explaining macroeconomic phenomena. Essential to understanding of large-scale economic phenomena is to think of the economy as a system of interactive heterogeneous individuals. Individual heterogeneity and interaction generate considerably difficult aggregation issues, making the relation between micro and macro models extremely complicated. The interest in aggregation over interactive heterogeneous agents is relatively recent (Hansen, 1998:240-1). The reminder of this chapter studies some of the aggregation issues directly relevant to the question of whether, in the presence of heterogeneity and interaction, the correct form of the aggregate model can be derived from the micro models alone. Or inferring the correct form of the macro model necessitates a substantial amount of information concerning the structure of the economy.

This section concentrates on the aggregation problems arising from individual heterogeneity. It starts with a discussion of the fundamental theorem of *exact* aggregation, due to Lau (1982). The significance of this theorem is that it specifies the conditions that are necessary in the presence of behavioural heterogeneity in order for the micro models alone to determine the aggregate model. An analysis of the conditions and their implications enables us to assess the success of the microfoundations program.

3.1 The Fundamental Theorem of Exact Aggregation

Individuals differ in many respects that are relevant to economic decisions. They differ in their tastes, opinions, information, incomes, demographic attributes, and the environment they operate in. Such differences usually give rise to differences in preferences, making people with identical income exhibit different patterns of consumption behaviour, and thus affect aggregate consumption. Of all possible individual heterogeneities, Lau (1982) considers demographic attributes such as age and number of children. To spell out Lau's result, we need to extend the framework adopted earlier to state Gorman's theorems. In particular, the micro functions should now include additional arguments to refer to individual demographic attributes.[23] That is,

$$\mathbf{Y}_{it} = f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) \qquad i = 1, \ldots, N \tag{3.1}$$

where $\mathbf{Y}_{it}$ denotes the individual consumption vector at time *t*, X_{it} individual income, $\mathbf{A}_{it}$ vector of individual attributes, $\mathbf{P}_t$ vector of prices at time *t*, and *N* the number of households in the economy. The aggregate demand $\mathbf{Y}_t$ is given by the sum of the individual demands:

$$\mathbf{Y}_t = \sum_i^N f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) \tag{3.2}$$

Clearly, in addition to the individual demand functions, the economist also needs to know the joint distribution of income and attributes in the economy to compute

[23] For the theory of exact aggregation see Jorgenson, et al. (1982), Lau (1977, 1982), and Heineke et al. (1988).

total consumption using equation (3.2). The search for an aggregate consumption function involves finding a function that reduces the distributional information required to compute total consumption. To achieve such a reduction, the function should dispense with the need for full knowledge of the income-attributes distribution and make it possible to compute total consumption by using a limited number of statistics (indices) summarising the distribution. That is, the desired macro function should take the form:

$$\begin{aligned} \mathbf{Y}_t &= \textstyle\sum_i f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) \\ &= F(g_1(X_{1t},\ldots,X_{Nt},\mathbf{A}_{1t},\ldots,\mathbf{A}_{Nt}),\ldots,g_L(X_{1t},\ldots,X_{Nt},\mathbf{A}_{1t},\ldots,\mathbf{A}_{Nt}),\mathbf{P}_t) \end{aligned} \tag{3.3}$$

where each function $g_l(.)$, $l = 1,\ldots,L$, is an index of the income-attributes distribution such as $\sum_i^N X_{it}$ and $\sum_i^N X_{it}\mathbf{A}_{it}$. Equation (3.3) should satisfy several conditions to reduce the information necessary for correctly computing total consumption. They include:

(1) The number of statistics $g_l(.)$, $l = 1,\ldots,L$, in the equation must be smaller than the number of the micro functions (i.e., $L < N$) for any reduction to occur in the information necessary for calculating aggregate consumption.

(2) The value of a statistic, summarising some aspect of a distribution, is mathematically invariant with respect to the ordering of the units in the population. Therefore, to be a statistic, the value of each function $g_l(X_{it},\ldots,X_{nt},\mathbf{A}_{1t}\ldots,\mathbf{A}_{nt})$ must be invariant with respect to whether individual i possesses attributes $\mathbf{A}^*$ and income x or individual j possesses attributes $\mathbf{A}^*$ and income x. Swapping the income and attributes of two individuals should leave the value of the function unchanged. This means each index function $g_l(.)$ must be *symmetric* with respect to subscript i through N. The symmetry requirement, as shown in Appendix C, necessitates the individual demand functions to be identical up to the addition of a term that is independent of the individual attributes and expenditure (Jorgenson, *et al.* 1982:113). Formally,

$$f_i(X_{it},\mathbf{A}_{it},\mathbf{P}_t)=f(X_{it},\mathbf{A}_{it},\mathbf{P}_t)+k_i(\mathbf{P}_t)\,. \tag{3.4}$$

Consequently, in order for an aggregation function of the form (3.3) to exist, all the individual demand functions for the same commodity must be identical up to the addition of a function that is independent of the individual attributes and income (Lau, 1982:122).

(3) Index functions $g_l(.)$, $l=1,...,L$, must be functionally independent. Or else, some of the indices play no genuine role in reducing the necessary distributional information and can be omitted without harm.

(4) To ensure that each $g_l(.)$ plays an essential role, the aggregate function $F(g_1,g_2,...,g_L,\mathbf{P}_t)$ must also be *invertible* in the indices $g_1,...,g_L$. That is, there must be a price vector $\mathbf{P}_t$ such that $F(g_1,g_2,...,g_L,\mathbf{P}_t)$ is invertible in $g_1,...,g_L$. To see why invertibility is necessary, consider function $F(G(g_1,g_2),g_3,...,g_L,\mathbf{P}_t)$. There is no price vector $\mathbf{P}_t$ for this function such that $F(G(g_1,g_2),g_3,...,g_L,\mathbf{P}_t)$ is invertible in $g_1,...,g_L$. The difficulty is with g_1 and g_2, which are effectively a single function, namely, G (Lau, 1982:126). Taken together, functional independence and invertibility ensure that the aggregate function is represented by a minimal number of index functions $g_l(.)$'s.

Individual demand functions that can be aggregated into an aggregate function of the form (3.3) are said to be exactly aggregable. The reason behind this nomenclature is that, when there is an aggregate function like (3.3), masking some aspects of the income-attributes distribution through aggregation does not jeopardize the ability to correctly compute aggregate consumption (Heineke et al., 1988). Lau (1982) establishes a theorem, known as the fundamental theorem of exact aggregation, that defines the conditions under which individual functions (3.1) can be exactly aggregated:

The Fundamental Theorem of Exact Aggregation: Aggregate function (3.3) exists, is continuously differentiable, and satisfies conditions (1) through (4) if and only if the individual functions (3.1) can be written as

$$f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) = b_1(\mathbf{P}_t)g_1^*(X_{it}, \mathbf{A}_{it}) + ... + b_q(\mathbf{P}_t)g_L^*(X_{it}, \mathbf{A}_{it}) + a_i(\mathbf{P}_t),$$
$$i = 1,...,N, \tag{3.5}$$

that is, if and only if the individual demand functions can be represented as sums of products of separate functions of prices and individual income and attributes (Jorgenson et al., 1982:104).

Equation (3.5) imposes several restrictions on the individual demand functions. To begin with, it requires the functions to be identical up to an additive term that is independent of the variables varying across individuals. In this regard, Lau's theorem makes no departure from Gorman's result. Secondly, with identical income *and* attributes, equation (3.5) excludes heterogeneity in marginal responses. In this respect, the theorem is a significant generalisation of Gorman's result, which, with identical income, excludes heterogeneity in marginal responses. Thirdly, equation (3.5) requires the individual functions to be linear in a number of functions of individual income and attributes. However, unlike in the case of Gorman's polar form, index functions $g_l^*(X_{it}, \mathbf{A}_{it})$ are permitted to depend nonlinearly on the individual income and attributes.

When the individual functions are of the form (3.5), each index g_l in the aggregate equation (3.3) corresponds to the sum of the individual functions $g_l^*(X_{it}, \mathbf{A}_{it})$, i.e., $g_l = \sum_i g_l^*(X_{it}, \mathbf{A}_{it})$, $(l = 1,...,L)$. Therefore, a corollary of the exact aggregation theorem is that the indices in the aggregate demand function are expressible as sums of some functions, each depending only on x_{it} or $\mathbf{A}_{it}$ (Jorgenson, et al., 1982:106). As a consequence, the aggregate function can be derived from the individual equations by substituting the sum of $g_l^*(X_{it}, \mathbf{A}_{it})$ for the corresponding index function $g_l(.)$.

The class of exactly aggregable functions, defined by equation (3.5), is the only class of functions where the individual functions alone determine the aggregate function and the meaning of the individual parameters fixes the interpretation of the aggregate parameters. However, this feature of exactly aggregable functions does not imply that if the individual functions were integrable, the aggregate function would also be integrable. It is only when the individual functions can be stated in terms of two terms $g_l^*(X_{it}, \mathbf{A}_{it})$, $l = 1,2$, that the integrability of the functions guarantees the integrability of the aggregate function, and the existence of a representative agent (Muellbauer, 1975, 1976).

Lau's theorem takes a significant step in claiming the room for individual heterogeneity. But the result does not yield much support for the microfoundations project. By reflecting on the implications of the theorem, one in fact begins to see the complications that individual heterogeneity creates for the project, even in simple situations where individual functions are exactly aggregable. Recall when the conditions of Gorman's theorem are in place, computing total consumption requires no information about the income distribution. As soon as one moves away from this unrealistic situation to a situation where the conditions of exact aggregations are in place, one can no longer predict aggregate consumption from total income. Instead, one requires knowing quite a good deal about the income distribution in the economy to calculate the required statistics. To illustrate this point, consider an example adapted from Stoker (1993). Suppose an economy of two small and two large families, with different marginal propensities to consume. Let the demand function for the small families be $Y_{it} = b_0(\mathbf{P}_t)X_{it}$ and for the large families be $Y_{it} = b_1(\mathbf{P}_t)X_{it}$. Further, let the attribute vector $\mathbf{A}_{it}$ be a qualitative variable, with A_{it}=1 denoting a small family and $A_{it} = 0$ a large family. The demand function for each household can then be written as

$$Y_{it} = b_0(\mathbf{P}_t)A_{it}X_{it} + b_1(\mathbf{P}_t)(1 - A_{it})X_{it}, \tag{3.6}$$

which is of the form (3.5). The aggregate demand model can be written as

$$Y_t = b_1(\mathbf{P}_t)\sum_i X_{it} + [b_0(\mathbf{P}_t) - b_1(\mathbf{P}_t)]\sum_i A_{it}X_{it}\,. \tag{3.7}$$

Now, suppose each small family currently receives £40 as income and spends a fourth of his income on goods and each large family receives £60 as income and spends half of its income on food. The aggregate equation (3.7) predicts total food consumption to be £80. If total income is doubled, depending who receives the additional income the aggregate model yields different results. If all the income goes to the small families, the model forecasts total consumption to be £130. If all the income goes to the large families, the model forecasts total consumption to be less then £180. Other income distributions lead to different predictions of total consumption. Predicting total consumption using equation (3.7) demands information on the amount of total income going to the small or large families. In real economies, the micro parameters $b_l(\mathbf{P}_t)$ are not known and econometricians turn to aggregate data to estimate them. This practice yields useful results if the relevant aspects of the distribution of the individual explanatory variables are not masked in the data. In the present case, the data should not be so aggregated that the total income going into the small families cannot be told apart from the total income going into the large families; the income of these family groups should be kept separate (Stoker, 1993:1836). As we consider real economies, the diversity of market participants turns out to be much richer and more complex and more disaggregated information is needed for estimating the correct aggregate model. The problem is that such information is difficult to obtain (Deaton and Muellbauer, 1980).

This difficulty aside, there is also no guarantee that exactly aggregable functions can always be stated in the form of equation (3.5) using a small number of terms $g_l^*(X_{it}, \mathbf{A}_{it})$. The effort to state individual functions in the form necessary for exact aggregation may require a large number of terms $g_l^*(X_{it}, \mathbf{A}_{it})$, which results in an aggregate function with a large number of indices $g_l(.)$, again making it difficult to reliably estimate it from samples usually available in practice. In practice, to counter this complexity, the analyst may need to work with a simplified aggregate function substantially different from the exact aggregate equation. The existence of a true aggregate function is one thing and the practicality or usefulness of the function is another thing. The microfoundations thesis wrongly implies that not only a true aggregate function exists but also it is simple enough to be estimated and used in practice.

3.2 The Effect of Nonlinearity

The requirements of exact aggregation are unrealistic and must be abandoned in modelling many phenomena. In reality, a household's income must reach a certain level to enable it to afford a car, purchase a house, save, go on a holiday, send its children to private schools, move house, buy a luxury car, and so forth. The demand for such commodities is not linearly dependent on income, and, for that reason, one has to work with a nonlinear (discrete) individual consumption model. If behaviours follow a nonlinear pattern, the exact aggregate function can no longer be inferred from the individual functions alone. To correctly derive the aggregate function, it is necessary to know in advance the joint distribution of the explanatory variables (income and attributes) in the economy (Cameron, 1990:207). This necessity will remain even if there were no heterogeneity in individual functions. A simple example best illustrates the issue.

Following Stoker (1993), suppose that the concern is to study the purchase of a single unit of a product such as a car, and that we only observe whether it is bought (say $Y_{it}=1$) or not ($Y_{it}=0$). Further, suppose the value to family i of buying the product depends on the price of the product P_t and the family's overall income X_{it}. To be specific, suppose the net benefit (utility) of the product for family i is given by $1+\beta_1 \ln P_t + \beta_2 X_{it}$. In that case, an appropriate model of a family i's decision to purchase the product would be the discrete model:

$$\begin{aligned} Y_{it} &= f(X_{it}, P_t) \\ &= 1 \quad \text{if } 1+\beta_1 \ln P_t + \beta_2 X_{it} \geq 0 \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{3.8}$$

The objective is to model the aggregate demand $\bar{Y}_t = N_t^{-1}\sum Y_{it}$, which is the proportion of the families who buy the product. How is this proportion to be modelled? Surely, it demands estimating the probability that a family buys the product, namely, $p(\ 1+\beta_1 \ln P_t + \beta_2 X_{it} \geq 0)$, which, of course, requires specifying the probability distribution of income X_t in the economy. Given the income distribution, the probability that a purchase is made can be calculated and the derivation of the aggregate model will then be straightforward. If the distribution of X_t is found to be, say, lognormal with $\ln X_t$ having mean μ_t and variance υ_t^2, the aggregate model will be:

$$E_t(y) = \Phi[\frac{1}{\beta_2 \upsilon_t}(1+\beta_1 \ln P_t + \beta_2 E_t(x) - \beta_2 \frac{\upsilon_t^2}{2})], \qquad (3.9)$$

where $E_t(y)$ denotes the expected number of families purchasing the product and $\Phi(.)$ is the univariate normal *cumulative* distribution function. If there were behavioural heterogeneity, for instance, if the parameters β_1 and β_2 varied across the families, further information about the probability distribution of the households would be needed to correctly compute aggregate demand and, as a result, the aggregate consumption model would further depart from the individual consumption models.[24]

This example points to some significant differences between aggregating over linear and nonlinear models. In the former case, when the requirements of the exact aggregation theorem hold, the individual models alone determine the correct macro model. In the case of nonlinear models, even when the same model is true of every individual, the correct form of the aggregate model depends on the distribution of the explanatory micro variables, and cannot be inferred from the micro models alone. An assumption about the distribution of the micro explanatory variables is an assumption about the configuration of the society.

[24] See Stoker (1984, 1986 and 1993), and Cameron (1990) for a discussion of aggregation of nonlinear models.

Thus, in the case of nonlinear models, the microfoundations thesis, which only permits macro models that can be derived solely from purely individualistic assumptions, falters. Moreover, it is difficult to see how the distribution of the explanatory variables can be estimated in a large economy. Economic data are hardly disaggregated enough to permit estimation of the distributions required for aggregating over nonlinear choice models (Cameron, 1990:212).

3.3 The Effect of Dynamics

Lau lays down the requirements of exact aggregation for a static setting. In practice, the agent lives in an uncertain environment and, to make decisions, needs to rely on his expectations of the future values of the variables affecting the outcomes of his decisions. Ideally, he estimates the expectations based on some observable variables whose values are already known. In that case, as in Hall's study, the appropriate model of individual behaviour is a dynamic model, which further complicates aggregation issues. Specifically, aggregation over simple heterogeneous dynamic models can produce complex model that is different than the individual models and cannot be given the same interpretation attributed to the micro models. The simplest instance of this phenomenon occurs in the case of aggregating over heterogeneous first order autoregressive processes, $AR(1)$. Consider aggregation of the following two $AR(1)$ processes:

$$X_{it} = \alpha_i X_{it-1} + \varepsilon_{it}, \qquad i = 1,2 \tag{3.10}$$

where ε_{1t} and ε_{2t} are a pair of independent, zero-mean, white noise series.[25] A simple calculation shows that the aggregate variable $X_t = X_{1t} + X_{2t}$ follows an autoregressive moving average (2,1) process

$$X_t = \alpha X_{t-1} + \beta X_{t-2} + \eta_{t-1} + \eta_t. \tag{3.11}$$

[25] The stochastic process $\{Z_t, t = 1,2,3,...\}$ is said to be a white-noise process provided that (i) $E(Z_t) = 0$ and (ii) $Cov(Z_t, Z_s) = \sigma^2$ for t=s and 0 for $t \neq s$.

In general, Box and Jenkins (1970) and Granger and Morris (1976) have shown that if N heterogeneous time series are added, each obeying an $AR(1)$ model with different parameter values, their sum typically follows an $ARMA(N, N-1)$ (Appendix, D).[26] If, as in the real economies, the number of decision-making units is large, the true aggregate model contains an extremely large number of parameters, making it impossible to estimate it from ordinarily available samples, simply because the number of parameters to be estimated exceeds the sample size. Also, due to excessive complexity (high number of parameters), the true model is not of any practical use (Granger, 1980:230-1). The relation between the micro and macro parameters also turns out to be so complicated that it makes it problematic to ascribe much of behavioural interpretation to the parameters in the aggregate model (Stoker, 1993:1981).

3.4 The Effect of Heterogeneous Environments

Households and firms encounter different environments in the sense that the processes generating the variables affecting their behaviours (e.g., income) vary across decision makers. The process generating the income of a household working, for instance, in the agricultural sector differs from the process generating the income of a household working in the banking sector. When there are no behavioural heterogeneities, such environmental differences have no impact on the form of the aggregate function. However, when there are behavioural heterogeneities, environmental differences critically shape the form of the regularities emerging at the economy level. Thus, no attempt at modelling the relation between the micro and macro levels can ever afford to neglect them. On the other hand, the existence of environmental heterogeneities fundamentally aggravates the differences between the micro and macro functions. To spell out some of the problems raised by environmental heterogeneities for the microfoundations project, we return to the example we borrowed from Lippi (1988) in the last chapter.

[26] See Granger 1999:42-48 for a brief discussion.

Lippi's illustration concerns an economy consisting of two consumers with demands following the static routines:

$$Y_{it} = \Pi_i X_{it}, \qquad i = 1,2. \tag{3.12}$$

Y_{it} and X_{it} are respectively consumption and income of the *i*th agent at time *t*, and the parameters Π_i are different, i.e. $\Pi_1 \neq \Pi_2$. Moreover, the independent micro variables X_{it} follow the autoregressive process:

$$X_{it} = a_i X_{it-1} + v_{it} \qquad 0 < a_i < 1, \tag{3.13}$$

with a_i being different for each individual, and v_{it} being orthogonal white-noise processes. The variables representing the state of the economy are aggregate demand $Y_t = Y_{1t} + Y_{2t}$ and aggregate income $X_t = X_{1t} + X_{2t}$. A general formula established in Lippi (1988) shows that the demand function for this economy is given by (Appendix, *E*):

$$Y_t = \alpha Y_{t-1} + \beta X_t + \gamma X_{t-1} + u_t\ . \tag{3.14}$$

The error term u_t is a white-noise process. The parameters of the aggregate model are defined as

$$\alpha = \frac{(\Pi_1 - k)a_1 + (k - \Pi_2)a_2}{(\Pi_1 - \Pi_2)}\ , \tag{3.15a}$$

$$\beta = k = \frac{Cov(\Pi_{1t} v_{1t} + \Pi_2 v_{2t}, v_{1t} + v_{2t})}{Var(v_{1t} + v_{2t})} \tag{3.15b}$$

$$\gamma = \frac{(\Pi_1 - k)\Pi_2 a_1 + (k - \Pi_2)\Pi_1 a_2}{(\Pi_1 - \Pi_2)}. \tag{3.15c}$$

Noticeably, aggregate demand equation (3.14) differs from the micro demand functions (3.12), containing variables that are absent in the micro functions. The aggregate parameters, more importantly, relate in a very complicated manner to both the parameters of the micro consumption functions (3.12) and those of the environmental functions (3.13), which represent the processes generating individual incomes.

Lippi's example is simple but reveals some key points regarding the micro and macro relation. To begin with, it illustrates that when people encounter heterogeneous choice situations and behave differently the correct aggregate model cannot be derived from behavioural equations alone. In addition to the behavioural equations, some knowledge of the structure of the economy, such as the (causal) process generating income in the banking system, is also necessary for deriving the regularities true of the economy. Moreover, as the number of heterogeneous individuals increases, and the complexity of the behavioural functions and the processes generating the independent micro variables rises, the complexity of the true aggregate equation increases beyond control. In fact, with both individual and environmental heterogeneities, even if the number of decision-makers does not exceed a single digit, the complexity of the correct aggregate equation exceeds the complexity of most aggregate equation used in practice. Finally, the example reveals that in an economy of heterogeneous decision makers, each operating in a different situation, there is no simple link between the aggregate parameters and the behavioural parameters. Both the parameters of the behavioural equations and those of the functions of the micro explanatory variables contribute in a complex manner to the aggregate parameters. It is not then appropriate to ascribe any behavioural meaning to the aggregate parameters. Inspecting the equalities (3.15a) through (3.15c), one wonders what interpretation can be given to the parameters in the aggregate equation (3.14) except that they are by-product of aggregation. In fact, the very existence of an aggregate (demand) function, which meaningfully relate to the micro functions, is in doubt. Economists have rarely come to grip with the issues arising from an attempt to aggregate over heterogeneous individuals operating in different dynamic situations. But those who have come to realise the severity of the complications have felt bound to abandon the nomenclature of a true aggregate function. Referring to Theil (1954)'s work on aggregation, Zellner (1969) writes:

> His [Theil] main result that the mathematical expectation of macro-coefficient estimators will in general depend on a complicated combination of corresponding and noncorresponding micro-coefficients was so disturbing to him that he seriously considered the following question in his concluding chapter (1954:180) "Should not we abolish these [macro] models altogether?" (Zellner, 1969:365).

3.5 Heterogeneity and Policy Evaluation

The analysis of Lippi's example has an important implication for the usefulness of aggregate models for policy analysis. Economic policies often deliberately seek to influence the economy by altering the mechanisms generating individual explanatory variables such as income. When decision makers operate in different choice situations and there are behavioural heterogeneities, the aggregate model is partly defined by the mechanisms generating the individual explanatory variables. And so, introduction of a new policy that affects these mechanisms can invalidate the aggregate model true of the economy prior to the intervention. Therefore, in the presence of individual and environmental heterogeneities, aggregate models *correctly* derived from individual models can yield a very wrong prediction of the effects of policies. To illustrate, we consider a simple model discussed in Geweke (1986).

Geweke's model is concerned with an industry in a small country that produces a single output Y_t, ultimately sold competitively in a world market. The production technology is the same for all the firms in the industry. To be specific, for the *i*th firm at time *t*,

$$Y_{it} = aX_{it} + dX_{it}^2, \qquad a > 0,\ d < 0, \tag{3.16}$$

where X_{it} is an input factor used to produce Y_t. Firms are distributed throughout the country and, as a consequence, the price for the output of each firm P_{it} varys, say, with access to deep-water ports. Output price varies through time but relative output prices across firms never vary.[27] The output price for the *i*th firm may be stated as

$$P_{it} = P_t P_i . \tag{3.17}$$

Input price r_t is the same across the country. Equation (3.16) is exactly aggregable, making it possible to estimate parameters *a* and *b* using time series data on the aggregate (average) input factor X_t and aggregate (average) output Y_t. Equation (3.16) and (3.17) give rise to average supply function,

$$Y_t = \frac{-a^2}{4d} + \frac{r_t^2}{4dP_t^2 E(P_i^2)}, \tag{3.18}$$

where $E(.)$ stands for average.. Suppose the concern is to predict the effect on production of an *ad rem* subsidy that amounts to substituting (3.17) with the new price regime

$$P_{it} = P_t P_i + u , \qquad u > 0. \tag{3.19}$$

Supply function (3.18) predicts the effect of the subsidy on average supply to be

$$[(r_t^2 / 4dP_i^2 E(P_i^2))((1 + u / P_t)^{-2} - 1)] .$$

However, when the new price regime (3.19) is in place, the actual average supply function is

$$Y_t = \frac{-a^2}{4d} + \frac{r_t^2}{4dP_t^2} E(P_i + \frac{u}{P_t})^{-2}, \tag{3.20}$$

and the actual effect of the subsidy on average supply is

[27] This, for example, would be the case if price differentials were due to different transportation costs and the relative prices of output and transportation were unchanging.

$$[(r_t^2 / 4dP_t^2)E[(P_i + u / P_t)^{-2} - P_i^{-2}],$$

which is different from the predicted change. The change in the mechanism generating individual prices invalidates the aggregate production function true prior to the intervention. The function cannot then be used to evaluate policies.[28] New classical economists argue for establishing aggregate models on features of human behaviour such as tastes that in their view, unlike expectations, are not affected by policy shifts. Geweke (1985) warns that the effect of ignoring aggregation problems caused by policy shifts may not be less than the effect of ignoring shifts in expectations.

The analysis of aggregation over heterogeneous individuals shows how individual heterogeneity limits the circumstances under which the microfoundations project can be accomplished. It is only when the micro functions are identical and (intrinsically) linear, and identical processes generate explanatory micro variables, that the individual functions alone determine the macro model.[29] If any of these conditions fails, substantial information regarding the structural features of the economy, including the processes generating the micro explanatory variables, would be needed to determine the macro model. This necessity of relying on macroeconomic phenomena (i.e., the processes generating micro explanatory variables) to model other macroeconomic phenomena undermines the central thesis of the microfoundations project that "the economist should start at the level of isolated individual" (Kirman, 1989:138; Rizvi, 1994:372). Modelling the economy requires beginning with some knowledge of the structure of the economy.

4 Modelling Interaction

The analysis of the representative agent modelling approach showed that, for studying the economy, one could not take a single individual as the unit of analysis. Explaining macroeconomic phenomena requires viewing the economy

[28] Kupiec and Sharpe (1991) provide another example.

[29] A relation that is not linear but can be made linear by taking the logarithm of each side of the equation is called intrinsically linear.

as an interactive system of heterogeneous decision-making units, which means one has to take 'a collection of interactive heterogeneous individuals' as the unit of analysis. We have studied some of the implications of individual heterogeneity for the microfoundations project. We now look at some of the issues arising from the presence of interaction in the economy.

4.1 Market Interactions

The earliest model of economic interaction is the theory of Walrasian general equilibrium, which is still a basic model of the market in economics (Ackerman, 1999). New classical economists usually interpret the call for microfoundations as the call for deducing the laws of aggregates from the theory of general equilibrium joined with the rational expectations hypothesis (Lucas and Sargent, 1979). The basic idea of the general equilibrium theory is that one cannot model a sector of the economy such as the consumption sector while treating the influences impinging on the sector by the rest of the economy as constant. Various sectors of the economy are interdependent and must be modelled simultaneously. The nuts and bolts of the Walrasian theory can be explained by considering an economy that, in addition to the consumption sector, has a single production sector. Specifically, consider an economy consisting of n consumers who own nonnegative initial endowments of capital goods and labour and consume q goods, and m firms producing the q goods using as input labour and capital services provided by the consumers. The theory introduces several basic assumptions about the consumers and firms of the economy. I rely on Leigh Tesfatsion's notes on macroeconomics (2003:2) to state these assumptions, while adding to her list the assumption of rational expectations:[30]

A1: Consumers are (subjective) expected utility maximisers.

A2: Firms are (subjective) expected profit optimisers.

A3: "The preferences of each consumer are exogenously given."

[30] Exact quotations are placed inside commas.

A4: "The income of consumers comes from dividends and from the sale of capital services and labour services."

A5: "Market for services and consumption goods are *complete*. That is, for each valued service and consumption good, there is a market price at which it can be bought or sold."

A6: "Consumers, taking expected good prices, wages, rental rates, and dividends as given, choose demand for consumption goods and supplies of capital and labour services to maximise their utility subject to a budget constraint and physical feasibility conditions (non-negativity and endowment constraints)."

A7: "Firms, taking expected good prices, wages, and rental rates as given, choose supplies of goods and demands for capital and labour services to maximise expected profits subject to technological feasibility conditions."

A8: "All purchase and sale agreements are costlessly enforced."

A9: Expectations are rational.

In addition to these assumptions, the theory introduces certain technical restrictions regarding the utility functions as well as production functions including continuity, convexity, and monotonicity of preferences. These are to ensure the existence of a Walrasian equilibrium, which is a set of relative prices and corresponding demand and supply quantities at which all consumers are maximising their expected utility conditional on their expected prices and dividends, all producers are optimising their profits conditional on their expected prices, and markets for all goods clear. Altogether, these assumptions entail that the economy is in equilibrium, prices fully reflect all the relevant information, and there is no conflict across business plans. In a Walrasian world, a decision maker has no need to communicate with others or adjust his decisions to those of others in the market. He only needs to consider prices to decide on his optimal course of action. Since in such a world all interactions take place through prices, the Walrasian economic model is referred to as a model of *market* or *indirect* interaction.

The call to establish the laws of the aggregates on the general equilibrium theory is an attempt to derive the laws of the economy from the above assumptions about individual behaviour, firm behaviour, tastes, technologies, and endowments as well as the postulates necessary for the existence of an equilibrium (Rizvi, 1994).

A question taken up by Sonnenschein (1972), Mantel (1976), and Debreu (1974) (henceforth, SMD) is whether the Walrasian assumptions impose any restrictions on the regularities emerging at the economy level. These theorists, to be precise, have inquired if the conditions imply any restrictions for the excess demand curve of the economy. The authors have found that, given the Walrasian conditions, *only* three properties carry over from the individual's excess demand curves to the aggregate excess demand curve. They are (i) "continuity, (ii) that the value of total excess demand must equal zero at all positive prices, i.e., that the budget constraint for the economy as a whole be satisfied (Walras' law), and (iii) the excess demand is homogenous of degree zero (only relative prices count)" (Kirman, 1992:122).[31] A Walrasian economy can exhibit any aggregate excess demand curve that satisfies these three requirements (Appendix *F* provides a simple statement of the SMD theorem).

The roots of the SMD theorem can be traced to the analysis of the subjective expected utility theory in the second chapter. The expected utility theory is a method for solving a decision problem. Almost any observed behaviour can be rationalised by varying the specification of the problem that the agent is trying to solve. Substantive implications attributed to the theory originate from the exogenous assumptions introduced to specify the agent's model of his choice situation and definition of his decision problem. The assumption that consumers (or firms) are subjective expected utility maximisers imposes little restriction on behaviour. The other assumption in the Walrasian theory, possibly restricting behaviour, is the market clearing condition. Again, as explained in chapter 2, when expectations of endogenous variables are involved, which is almost always the case in economics, the condition is not adequate to pin down any particular behaviour. Infinitely many price vectors usually clear the market.

Also, in the presence of behavioural heterogeneities, including heterogeneities in choice situations, regardless of the form of the individual functions, the aggregate relations can take unlimited forms. This means it is never possible to derive the relations emerging in an economy from the thin rationality assumptions of the

[31] Rizvi (1994), Kirman (1989) and (1992) offer accessible discussions of the SMD result.

equilibrium theory, which takes no notice of distributional features, how people model their choice situations, define their decision problems, and interact with each other (Kirman, 1989:128). Therefore, even if the Walrasian assumptions were plausible, they would still be inadequate for furnishing a foundation for the regularities true of economic aggregates. It is wrong to think that "significant results could be obtained by starting from very general hypotheses about the behaviour of economic agents", (Ingrao, et al., 1990:316).

The Walrasian theory also provides no explanation of who sets the equilibrium prices. It simply assumes that the economy is in equilibrium, implying that prices are exogenous to the economy. Moreover, by supposing that all business plans are costlessly enforced, the theory rules out the existence of transaction costs, and, hence, money, which is a means for facilitating the coordination of the entire economy, finds no room in the theory (Debreu, 1959). Finally, by supposing that prices convey all the information relevant for decisions and thereby ruling out any direct interaction among market participants, the theory excludes the possibility of coordination failures. It thus fails to make any room for central macroeconomic phenomena that arise from the inability of market participants in a decentralised economy to coordinate their actions. Explanation of the process of price formation, market crashes, depressions, convergence to equilibrium, the role of money, and economic institutions calls for questioning the Walrasian view and allowing direct interaction into the models of the economy.

4.2 Non-Market Interactions

Attempts at modelling the phenomena not explainable within the Walrasian setting have led a growing number of economists to start thinking of the economy as a society of *directly* interacting heterogeneous individuals. A consequence of this change of attitude has been the development of formal models which allow for the state of a person (i.e., strategies, preferences, and expectations) to directly depend on the states (i.e., strategies, preferences, and expectations) of other participants in the economy (Glaeser, et al., 2001). The models are based on rival principles and are still highly simple. Nevertheless, they well serve to bring to the fore some basic lessons about the relation between the individual and aggregate levels in a heterogeneous and interactive system. To illustrate the type of models that can possibly account for the phenomena left out by the Walrasian theory, of the many approaches to modelling non-market interactions, this section concentrates on the more familiar approach provided by the Game theory, which views the economy as a many-person game or as a collection of interdependent teams (Bryant, 1996). [32] After describing a typical game theoretic model of a macroeconomic phenomenon, the section derives some general implications of this alternative view of the economy for the microfoundations project.

We consider a generalisation of the stag hunt game used by some New Keynesian economists to study certain aspects of the economy that appear puzzling from the Walrasian perspective. The description is based on the stag hunt production game given in Bryant (1994), the coordination game in Cooper (1999), and the model of involuntary unemployment in Tesfatsion (2003). As in any game theoretic model, the optimal strategy of each market participant in these models depends on the strategies of every other market participant. Thus, what is rational for a person to choose depends on his beliefs about the preferences, expectations, and strategies of others in the market.

Consider an economy consisting of N agents (or, N groups of individuals) indexed $i = 1,2,...,N$, living on N separate locations. Each agent is endowed with L units of leisure and likes to consume two goods, leisure C_1 and another commodity C_2, called bread. Each agent has a strictly increasing and concave differentiable utility function $u(C_1, C_2)$. The agents first work (i.e., sacrifices leisure) to produce grains, and grains are effortlessly carried out to a location and combined to produce bread. N different types of grains are needed to produce bread, each being produced by a different individual. Also, one unit of leisure produces one unit of grain and one unit of bread is produced by N units of grain – one unit for each type of grain. Bread production thus follows the relation,

[32] Non-market interactions "are interactions between individuals, which are not regulated by the price mechanism. Glaeser, et al. (2001:1). For a survey of the literature on Non-market interactions see Glaeser et al (2001), section 2.

$$Q(g_i,\ldots,g_N) = N.\min\{g_1,\ldots,g_N\}, \tag{4.1}$$

where g_i is the amount of grain produced by the *i*th agent, and a surplus of any of the grains is costlessly discarded as waste. The bread is *equally* distributed among all individuals:

$$\frac{Q(g_i,\ldots,g_N)}{N} = \min\{g_1,\ldots,g_N\}. \tag{4.2}$$

Each agent is assumed to know the common leisure endowment *L*, the common utility function $u(C_1,C_2)$, the production function (4.1), the distribution rule (4.2), and that every one is rational, as well as that every one has common knowledge of these rules and parameters.

Each individual should decide how many hours of leisure to sacrifice for producing grain. Because of the nature of the production function and the distribution rule, a player's optimal decision depends on the other players' strategies. Let e_i be the effort that agent *i* devotes to grain production g_i and $\bar{e}$ be the vector of the efforts of the other agents. So, g_i depends on $\bar{e}$, i.e., $g_i = f(e_i,\bar{e})$. Suppose the more leisure is sacrificed the less pleasant it is but if all individuals equally sacrifice leisure to produce grain, the additional output produced by the increased effort more than compensates for the added pain of the sacrifice. This means all individuals are better off if all exert the maximum effort possible. Let us denote the payoff of agent *i* from action e_i when all other agents take action $\bar{e}$ by $\Pi(e_i,\bar{e})$ and let $\hat{e}_i(\bar{e})$ be the optimal response of agent *i* when other agents take action $\bar{e}$. Since any effort made by agent *i* above the minimum effort made by some other agent *j* is wasted, if other agents are choosing action *e*, it is in the interest of agent *i* to select *e* as well. That is, $\hat{e}_i(e) = e$. The game thus has a continuum of (symmetric) Nash equilibria defined by

$$\mathrm{S} = \{e \in [0,L] \mid \Pi_1(e,e) = 0\}, \tag{4.3}$$

where the subscript in Π_1 denotes a partial derivative. The set S includes an optimal equilibrium corresponding to the case when everyone devotes maximal effort level to production. Let denote the optimal equilibrium with $s*$. The continuum of Nash equilibria in S is Pareto ranked as

$$0 \leq s \leq s*. \tag{4.4}$$

Any Nash equilibrium s below $s*$ is a coordination failure.[33] This is in contrast to general equilibrium models where all the equilibria are efficient.

This game theoretic model surely abstracts away a great deal of complexities of the real world. Yet, by capturing the notion of strategic uncertainty and allowing for coordination failure, it provides a general framework for thinking about important issues including why economies often go into recessions without an observable external shock, why there are business cycles, why there is involuntary unemployment, why there are legal and financial institutions, and many others matters. Any theory aiming to address these issues should view the economy as an interactive system, and refer to the notion of coordination failure, which the team production game adequately helps capture (Bryant, 1996; Mankiw, 1993; Tefastion, 2003).

Two aggregation issues arise in the above model economy. The first concerns the existence of a production function $Q = G(\mathrm{L})$ that correctly maps aggregate leisure L to aggregate production Q such that $G(\mathrm{L}) = N.\min\{g_i,...,g_N\}$. The second issue concerns the connection between the aggregate function and the micro production functions.

To address the first question, note that the game has many solutions. Even when the players' beliefs are consistent, there is a continuum of Nash equilibria. And

[33] " A *coordination failure* is said to occur when mutual gains, potentially attainable from a feasible all-around change in agent behaviour (strategies) are not realised because no *individual* agent has an incentive to deviate from his [sic] current behaviour" (Leigh Tesfatsion, 1994). Quoted in Bryant (1996:157).

each solution gives rise to a particular level of aggregate output. As a consequence,. aggregate output is not solely a function of aggregate inputs in the economy (total leisure available). In fact, depending on what everyone thinks about the strategy of everyone else, almost any aggregate output is possible. So, there can theoretically be no function G(L) that correctly predicts aggregate output Q (bread) from aggregate input $\boldsymbol{L}$. The existence of multiple solutions calls "into question the very meaningfulness of aggregate production functions, and of aggregate inputs" (Bryant, 1996:168).

This point can be traced to a criticism of Klein's (1946) treatment of the aggregation problem. Klein argued that "there are certain equations in microeconomics that are independent of the equilibrium conditions and we should expect the corresponding equations of macroeconomics will also be independent of the equilibrium conditions. The principal equations that have this independence property [...] are the technological production functions. The aggregate production function should not depend upon profit maximisation, but purely on technological factors" (1946b:303). Consequently, he rejected using the entire micro model with the profit maximisation assumption to derive the production function of the economy.

May (1947) criticised Klein's position by arguing that even the production function of a firm is not a purely technical relationship, since it results from a decision-making process. He concluded that the aggregate production function is also a fictitious entity in the sense that there is no global decision-maker, who allocates recourses optimally:

> ... The aggregate production function is dependent on all the functions of the micromodel, including the behaviour equations such as profit-maximisation conditions, as well as upon all exogenous variables and parameters. This is the mathematical expression of the fact that the productive possibilities of an economy are dependent not only upon the productive possibilities of the individual firms (reflected in production functions) but on the manner in which these technological possibilities are utilized, as determined by the socio-economic framework (reflected in behaviour equations and institutional parameters). Thus the fact that our aggregate production function is not purely technological corresponds to the social character of aggregate production (May, 1947: 63).

In general, because of the dependence of aggregate output on equilibrium conditions, if there is a function correctly relating aggregate output to aggregate inputs, it theoretically includes among its arguments a variable or variables that refer to the conditions. Led by similar thoughts, Colander (1986) rejects the conventional aggregate production function $Q = f(K,L)$, which defines aggregate output Q as a function of total labour supply L and total capital K. As an alternative, he proposes an aggregate production function that schematically takes the form $Q = f(K,L,C)$, where C represents the degree of coordination in the economy.[34] A similar consideration, of course, applies to the consumption sector of the economy, as what one consumes can also critically depend on what other people consume. Without any exception, the theoretical functions describing the state of an interactive economy contain variable or variables that refer to the level of interdependencies in the economy.

This remark also embodies the answer to the second question, which has to do with the relation between the aggregate and individual functions in an interactive economy. As it should be clear by now, to be able to derive an aggregate function from individual functions, every variable in the aggregate function must somehow be defined by aggregating over the variables in the individual functions. But variables such as C, which refers to the level of coordination or the web of perceived interdependencies in the economy, are not aggregates of any individual functions. And so, the functions describing an interactive system cannot be derived by aggregating over the individual functions. In fact, to describe individual behaviour in an interactive system, the individual model, as in the stag hunt game, should include variables referring to the state of the economy. So, to be precise, the individual models are not individualistic either; they are of a social character.

It is appropriate to close this section with two further points: In theory, it makes sense to include in the aggregate functions variables such as C that refer to the level of coordination. It is, however, difficult to envision how such variables can

[34] This function is quoted in Bryant (1996), where he refers to Colander (1986) but does not mention the reference. The underlying idea, though, is vividly explicit in Colander (1996).

be operationalized. In his later writings, Colander acknowledges the difficulties with his schematic aggregate production function, and has become inclined towards a purely statistical approach to the study of aggregate data, akin to the view of macroeconomics put forward in Basmann (1972) and Sims (1980) (Colander, 1996:66). Sims, as we learnt earlier, regards models of economic aggregates as efficient summaries of data with no theoretical link to the processes at the individual level.

Finally, the game theoretic assumption that the state of every individual depends on the state of every other individual is not necessary for multiple solutions. Even if the game theoretic framework is weakened, and less extreme forms of interdependencies are considered, the multiplicity phenomenon still persists. For multiple solutions, it is enough that the states of some of the decision-making units depend on the states of some other units (Glaeser et al., 2001).

5 Conclusion

This chapter studied the second issue at the heart of any attempt at theorising about the economy, i.e., the nature of the link between the micro and macro levels. It started with an investigation into the representative agent modelling approach that sees no difference between the laws of a single individual and the economy. Even though the approach is still quite prevalent in theoretical economic modelling, the conditions under which the behaviours of a collection of individuals can be modelled as the behaviour of a single individual are extremely restrictive. More importantly, the approach is fallacious. Variables like prices, economic growth, interest rate, unemployment level, and inflation, which enter in the decision model of a single individual as exogenous variables, are determined within the economy. It is thus in principle wrong to extend causal implications of individual models to the economy. Also, since from the perspective of the representative-agent modelling, individual differences are entirely irrelevant, representative agent models are not suitable for the analysis of policies designed to work by manipulating some distributional features of the economy.

These problems demonstrate why the study of the economy cannot be based on an analysis of the behaviour of a single individual. One ought to take as unit of analysis the behaviour of a group of heterogeneous interactive individuals to study the economy. This necessity leads to complications that fundamentally blur the relation between the individual and aggregate levels. When there is individual heterogeneity, the correct form of the aggregate relations not only depends on the form of the individual behavioural models but also on the joint distribution of the independent micro variables in the economy and on the mechanisms generating them. So, assumptions regarding the joint distribution of the independent micro variables and the causal processes generating the variables become an integral part of a correctly specified model of the aggregates. As a result, there will be no resemblance between the laws of the individuals and those of the aggregates. Also, since the parameters in the aggregate functions depend on the parameters defining the processes generating the independent micro variables, they do not admit any behavioural interpretation.

Additionally, economic policies often seek to influence the economy by changing the distribution of variables such as income. In the presence of individual heterogeneity, the aggregate model is sensitive to changes in the distributional configuration of the economy and the mechanisms generating the individual exogenous variables. Thus, a distributional policy change can invalidate the model fitted to the economy prior to the intervention. Policy analysis, in theory, requires knowing how a policy affects the distributional configuration of the economy, and how the distributional change affects the fitted model so as to derive the model that would consequently be true of the economy. Information on the distributional configuration of the economy, or the mechanisms generating the exogenous variables entering individual models, is difficult to obtain, making the task of establishing models useful for policy analysis difficult.

Complications arising from individual heterogeneity undermine the aim of deriving a macroeconomic model from individual models alone. Yet, the complications arising from direct interaction among market participants are more detrimental to the microfoundations project. Modelling individual behaviour in an interactive environment necessitates introducing into the micro model variables referring to preferences, expectations, and strategies of other decision makers. Such variables are not aggregable. Moreover, the existence of multiple equilibria in an interactive system excludes the very existence of a true aggregate 'function' linking explanatory aggregate variables (say, capital and labour) to the aggregate dependent variable (say, output). If there is a true model involving aggregate variables, it must involve a variable or variables referring to the interdependencies in the economy. It is though difficult to see how such a variable or variables could be operationalized. Nor are they aggregate of any micro variables.

These reflections on the connection between the micro and macro levels in an interactive heterogeneous system do not rule out the emergence of regular patterns at the economy level that can be modelled statistically. What they reject is the claim that the emerging patterns are in any straightforward manner related to the processes at the individual level or that they can be given any behavioural meaning. The analysis of aggregation issues, particularly those related to the effects of interaction, supports the view of macroeconomic models put forward by atheoretical econometricians, like Sims (1980), who view large-scale economic models simply as efficient summaries of data.

Finally, if the view of the economy as an interactive system is taken seriously, the existence of causal relations among economic aggregates such as aggregate capital will be in doubt. In that case, there seems to be no point in applying structural modelling tools to aggregate economic data. There will be no causal relations for them to discover.

Appendices

Appendix A: Clarida's Life Cycle Model

As in the text, we state the simplest possible case of Clarida's model, which is also discussed in Deaton (1992). The case is built around an economy with the following features:

Assumption 1: Each worker lives for three periods, working in the first two periods of his or her life and retiring in the third. It is assumed that only one person is born in each period.

Assumption 2: In period t, each person receives an identical amount of labour income Y_t while working, but zero during retirement. Consumption during retirement is financed from assets accumulated during the working periods.

Assumptions 3: Y_t follows a random walk with drift,

$$Y_t = g + Y_{t-1} + \varepsilon_t \, . \qquad \text{(A1)}$$

The per capita labour income also follows a random walk:

$$\frac{2Y_t}{3} = \frac{2g}{3} + \frac{2Y_{t-1}}{3} + \frac{2\varepsilon_t}{3} \, .$$

Assumption 4: Interest rate is zero and each person decides to leave no asset behind.

Assumption 5: Everyone is a pure permanent income life-cycler.[1]

Note that labour income received by each individual does not follow a random walk; by assumption labour income is zero with probability one during retirement. Assuming rational expectations, each individual best forecast of labour income during the next working period is the current labour income plus the cumulative drift, i.e., $g + Y_1$.

A person who is born in period t thus consumes $\frac{2Y_t + g}{3}$ during his first working period and $\frac{2Y_t + g}{3} + \frac{\varepsilon_{t+1}}{2}$ during the second working period and retirement period.

[1] As in the text, permanent income is defined as the annuity value of current financial and human wealth, and consumption is accordingly set equal to permanent income.

The table below shows the individual consumptions during the first five periods of the life of the economy.

Period 1	Period 2	Period 3	Period 4	Period 5	Period 6
$\frac{2Y_1+g}{3}$	$\frac{2Y_1+g}{3}+\frac{\varepsilon_2}{2}$	$\frac{2Y_1+g}{3}+\frac{\varepsilon_2}{2}$	Dead		
	$\frac{2Y_2+g}{3}$	$\frac{2Y_2+g}{3}+\frac{\varepsilon_3}{2}$	$\frac{2Y_2+g}{3}+\frac{\varepsilon_3}{2}$	Dead	
		$\frac{2Y_3+g}{3}$	$\frac{2Y_3+g}{3}+\frac{\varepsilon_4}{2}$	$\frac{2Y_3+g}{3}+\frac{\varepsilon_4}{2}$	Dead
			$\frac{2Y_4+g}{3}$	$\frac{2Y_4+g}{3}+\frac{\varepsilon_5}{2}$	...

Now consider total consumption change between period 4 and period 3.

$$\Delta C_4 = C_4 - C_3 = \frac{2Y_4+g}{3}+\frac{\varepsilon_4}{2}-\frac{2Y_1+g}{3}-\frac{\varepsilon_2}{2}. \qquad \text{(A2)}$$

Writing Y_4 in terms of Y_1 yields

$$Y_4 = 3g + Y_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4. \qquad \text{(A3)}$$

Substituting (A3) for Y_4 in (A2) yields

$$\Delta C_4 = 2g + \frac{7\varepsilon_4}{6} + \frac{2\varepsilon_3}{3} + \frac{\varepsilon_2}{6}. \tag{A4}$$

If we consider total consumption change at time t in general, rather than at period 4, then total consumption change for the economy is given by

$$\Delta C_t = 2g + \frac{7\varepsilon_t}{6} + \frac{2\varepsilon_{t-1}}{3} + \frac{\varepsilon_{t-2}}{6} , \tag{A5}$$

which is of the same form as the result stated in the text. Average consumption change follows:

$$\Delta \overline{C}_t = \frac{2}{3}g + \frac{7}{18}\overline{\varepsilon}_t + \frac{2}{9}\overline{\varepsilon}_{t-1} + \frac{\overline{\varepsilon}_{t-2}}{18} . \tag{A6}$$

Appendix B: Pischke's Incomplete Information Model

The following assumptions characterize Pischke's economy.

Assumption I: Average income follows a random walk with drift.[2]

Let Y_t stand for average income and g for the drift term. Then, average income is given by

$$Y_t = g + Y_{t-1} + \varepsilon_t . \tag{B1}$$

Assumption II: Each consumer income is the average income plus an idiosyncratic component that is purely transitory, represented by a white noise

$$Y_{it} = Y_t + u_{i_t} . \tag{B2}$$

The first difference of individual income is given by

$$\begin{aligned} \Delta Y_{it} &= Y_t + u_{i_t} - Y_{t-1} - u_{i_{t-1}} \\ \Delta Y_{it} &= g + Y_{t-1} + \varepsilon_t + u_{i_t} - Y_{t-1} - u_{i_{t-1}} \\ \Delta Y_{it} &= g + \varepsilon_t + u_{i_t} - u_{i_{t-1}} . \end{aligned} \tag{B3}$$

Assumption III: Each person only observes the sum of the contemporaneous macro and private shocks and cannot separate them. He only estimates the moving average process

[2] This appendix is based on Deaton (1992: 171-173)

$$\Delta Y_{it} = g + \eta_{it} - \lambda \eta_{it-1} . \tag{B4}$$

Assumption IV: Every household satisfies the infinite-life permanent income model (Hall's model). Individual consumption therefore follows random walk:

$$\Delta C_{it} = (1 - \frac{\lambda}{1+r}) \eta_{it} . \tag{B5}$$

The change in average consumption C_t is obtained by averaging over (B5):

$$\Delta C_t = (1 - \frac{\lambda}{1+r}) \eta_t . \tag{B6}$$

Now, since the real first difference of individual income is

$$\Delta Y_{it} = g + \varepsilon_t + u_{it} + u_{it-1}$$

and

$$\Delta Y_t = \sum \Delta Y_{it} = g + \varepsilon_t + \sum u_{it} / N + \sum u_{it-1} / N ,$$

we have

$$\Delta Y_t = g + \varepsilon_t \tag{B7}$$

(because u_{it} is white noise, $\sum u_{it} / N$ and $\sum u_{it-1} / N$ are equal to zero; in other words, the idiosyncratic components by assumption have zero means over the population).

On the other hand, since the derived first difference of individual income is

$$\Delta Y_{it} = g + \eta_{it} + \lambda \eta_{it-1} ,$$

and

$$\Delta Y_t = \sum \Delta Y_{it} = g + \sum \eta_{it} / N + \sum \lambda \eta_{it-1} / N ,$$

we have

$$\Delta Y_t = g + \eta_t - \lambda \eta_{t-1} . \tag{B8}$$

From (B7) and (B8) we have

$$\varepsilon_t = \eta_t - \lambda\eta_{t-1}$$

and

$$\eta_t = \varepsilon_t + \lambda\eta_{t-1}\,. \tag{B9}$$

Combining (B6) and (B9) yields

$$\Delta C_t = (1 - \frac{\lambda}{1+r})(\varepsilon_t + \lambda\eta_{t-1}),$$

$$\Delta C_t = (1 - \frac{\lambda}{1+r})\varepsilon_t + (1 - \frac{\lambda}{1+r})\lambda\eta_{t-1},$$

$$\Delta C_t = (1 - \frac{\lambda}{1+r})\lambda\eta_{t-1} + (1 - \frac{\lambda}{1+r})\varepsilon_t\,.$$

From (B6) we have

$$\Delta C_t = \lambda\Delta C_{t-1} + (1 - \frac{\lambda}{1+r})\varepsilon_t,$$

which yields the average consumption function as

$$C_t = (\lambda + 1)C_{t-1} - \lambda C_{t-2} + (1 - \frac{\lambda}{1+r})\varepsilon_t\,. \tag{B10}$$

Appendix C: Lau's Theorem (1982)

The individual functions $f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t)$ are of the type

$$f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) = f(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) + k_i(\mathbf{P}_t), \tag{C1}$$

only if the index functions $g_l(.)$ are symmetric.

Suppose $g_l(.)$ are not symmetric. In that case, exchanging the income X_{rt} and attributes $\mathbf{A}_{rt}$ of agent r with those of agent s changes the value of $g_l(.)$. Hence,

$$\sum f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) \neq \sum\nolimits_{i \neq s, i \neq r} f_i(X_{it}, \mathbf{A}_{it}, \mathbf{P}_t) + f_s(X_{rt}, \mathbf{A}_{rt}, \mathbf{P}_t) + f_r(X_{st}, \mathbf{A}_{st}, \mathbf{P}_t). \tag{C2}$$

After eliminating the identical terms and reordering, we obtain

$$f_s(X_{rt}, \mathbf{A}_{rt}, \mathbf{P}_t) - f_r(X_{rt}, \mathbf{A}_{rt}, \mathbf{P}_t) \neq f_r(X_{st}, \mathbf{A}_{st}, \mathbf{P}_t) - f_s(X_{st}, \mathbf{A}_{st}, \mathbf{P}_t), \tag{C3}$$

which only holds if the individual functions cannot be stated as (C1). Therefore, in order for (C1) to hold, the index functions $g_t(.)$ must be symmetric.

Appendix D: Aggregation over Heterogeneous Time Series

Suppose that X_{1t} and X_{2t} are a pair of series generated by

$$X_{1t} = \alpha_1 X_{1t-1} + \varepsilon_{1t} \tag{D1a}$$
$$X_{21t} = \alpha_2 X_{2t-1} + \varepsilon_{2t} \tag{D1b}$$

where ε_{1t} and ε_{2t} are a pair of independent, zero-mean white noise series. Equation system (D1) can be written as:

$$(1-\alpha_1 L)X_{1t} = \varepsilon_{1t} \tag{D2a}$$
$$(1-\alpha_2 L)X_{2t} = \varepsilon_{2t}.^3 \tag{D2b}$$

Or,

$$X_{1t} = \varepsilon_{1t}/(1-\alpha_1 L) \tag{D3a}$$
$$X_{2t} = \varepsilon_{2t}/(1-\alpha_2 L). \tag{D3b}$$

Let $X_t = X_{1t} + X_{2t}$. It follows that

$$(1-\alpha_1 L)X_t = (1-\alpha_1 L)X_{1t} + (1-\alpha_1 L)X_{2t}$$
$$(1-\alpha_2 L)(1-\alpha_1 L)X_t = (1-\alpha_2 L)(1-\alpha_1 L)X_{1t} + (1-\alpha_2 L)(1-\alpha_1 L)X_{2t}. \tag{D4}$$

Using (D3a) and (D3b), aggregate equation (D4) can be restated as:

$$(1-\alpha_2 L)(1-\alpha_1 L)X_t = (1-\alpha_2 L)\varepsilon_{1t} + (1-\alpha_1 L)\varepsilon_{2t} \tag{D5}$$

Based on the definition of ε_{1t} and ε_{2t}, the right hand side of (D5) is equivalent to:

$$(1-\alpha L)\varepsilon_t = (1-\alpha_2 L)\varepsilon_{1t} + (1-\alpha_1 L)\varepsilon_{2t}. \tag{D6}$$

[3] The polynomials are usually written as $\alpha_i(L)$ but the sake of simplicity is written here as $\alpha_i L$

Combining (D5) and (D6) gives the desired result:

$$(1-\alpha_2 L)(1-\alpha_1 L)X_t = (1-\alpha L)\varepsilon_t, \qquad \text{(D7)}$$

which is an ARMA (2,1). This exercise is an example of a general theorem proved by Granger and Morris (1976) and Box and Jenkins (1970).

Appendix E: Lippi's Simple Economy[4]

We work with the two-consumer economy described in the text. Let Y_{it} denote the consumption of the ith agent and X_{it} the income of the ith agent, where $i=1,2$. Suppose individual consumptions follow the static rules:

$$\begin{cases} Y_{1t} = \Pi_1 X_{1t} \\ Y_{2t} = \Pi_2 X_{2t} \end{cases} \qquad \Pi_1 \neq \Pi_2 \qquad \text{(E1)}$$

while the process generating individual incomes are given by:

$$\begin{cases} X_{1t} = \alpha_1 X_{1t-1} + v_{1t} \\ X_{2t} = \alpha_2 X_{2t-1} + v_{2t} \end{cases} \qquad \alpha_1 \neq \alpha_2 . \qquad \text{(E2)}$$

v_{it} s are white-noise process. Also, for the sake of simplicity, assume that v_{1t} and v_{2t} are independent. Aggregate consumption Y_t and aggregate income X_t are defined as

$$\begin{cases} Y_t = Y_{1t} + Y_{2t} \\ X_t = X_{1t} + X_{2t} \end{cases} . \qquad \text{(E3)}$$

The concern is to infer aggregate consumption function $Y_t = f(X_t)$. Equation (E2) can be restated as

$$\begin{cases} (1-\alpha_1 L)X_{1t} = v_{1t} \\ (1-\alpha_2 L)X_{2t} = v_{2t} \end{cases} \qquad \text{(E4)}$$

where $\alpha_i L$ s are polynomials in the lag operator L and $\alpha_i(0)=1$. Then

[4] The proof to follow is a restatement of the proof in Lippi (1988).

$$\begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix} = \begin{pmatrix} \frac{1}{1-\alpha_1 L} & 0 \\ 0 & \frac{1}{1-\alpha_2 L} \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix}. \tag{E5}$$

From (E1) and (E3), for vector (Y_t, X_t) we have

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} \Pi_1 & \Pi_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} X_{1t} \\ X_{2t} \end{pmatrix}. \tag{E6}$$

Combining (E5) and (E6) yields

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} \Pi_1 & \Pi_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{1-\alpha_1 L} & 0 \\ 0 & \frac{1}{1-\alpha_2 L} \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix}. \tag{E7}$$

Represent (E7) as

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} \Pi_1 & \Pi_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{1-\alpha_1 L} & 0 \\ 0 & \frac{1}{1-\alpha_2 L} \end{pmatrix} \begin{pmatrix} \Pi_1 & \Pi_2 \\ 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} \Pi_1 & \Pi_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix}. \tag{E8}$$

And let

$$\begin{pmatrix} W_{1t} \\ W_{2t} \end{pmatrix} = \begin{pmatrix} \Pi_1 & \Pi_2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_{1t} \\ v_{2t} \end{pmatrix}.$$

Like v_{it}, W_{it} are also white-noise processes. Then, we have

$$\begin{pmatrix} Y_t \\ X_t \end{pmatrix} = \begin{pmatrix} \frac{\Pi_1}{1-\alpha_1 L} & \frac{\Pi_2}{1-\alpha_2 L} \\ \frac{1}{1-\alpha_1 L} & \frac{1}{1-\alpha_2 L} \end{pmatrix} \begin{pmatrix} \frac{1}{\Pi_1 - \Pi_2} & \frac{-\Pi_2}{\Pi_1 - \Pi_2} \\ \frac{-1}{\Pi_1 - \Pi_2} & \frac{1}{\Pi_1 - \Pi_2} \end{pmatrix} \begin{pmatrix} W_{1t} \\ W_{2t} \end{pmatrix}. \tag{E9}$$

To simplify matters, let

$$A=(1-\alpha_1 L)$$
$$B=(1-\alpha_2 L)$$
$$C=\Pi_1-\Pi_2$$
$$E=ABC.$$

Equation system (E9) can be written as

$$\begin{pmatrix} Y_t \\ \\ X_t \end{pmatrix}=\begin{pmatrix} \dfrac{\Pi_1 B-\Pi_2 A}{E} & \dfrac{(A-B)\Pi_1\Pi_2}{E} \\ \dfrac{B-A}{E} & \dfrac{A\Pi_1-B\Pi_2}{E} \end{pmatrix}\begin{pmatrix} W_{1t} \\ \\ W_{2t} \end{pmatrix}. \tag{E10}$$

Still to further simplify necessary calculations let rewrite the fist matrix on the right hand side of (E10) as

$$\begin{pmatrix} F & G \\ H & I \end{pmatrix}$$

and call it M. Multiplying both side of (E10) by the adjoint of M yields

$$\begin{pmatrix} I & -G \\ -H & F \end{pmatrix}\begin{pmatrix} Y_t \\ X_t \end{pmatrix}=\begin{pmatrix} FI-GH & 0 \\ 0 & FI-GH \end{pmatrix}\begin{pmatrix} W_{1t} \\ W_{2t} \end{pmatrix}. \tag{E11}$$

From (E11) we have

$$\begin{cases} -IY_t-GX_t=(FI-GH)W_{1t} \\ -HY_t+FX_t=(FI-GH)W_{2t} \end{cases}. \tag{E12}$$

Let multiply the second equation in (E12) by a scalar k and subtract it from the first one. This yields

$$(I+kH)Y_t=(G+kF)X_t+(FI-GH)(W_{1t}-kW_{2t}). \tag{E13}$$

After substituting the definitions of F, G, H, and I into (E13), we need only some elementary algebra to derive equation

$$\left(1-\frac{\Gamma_1\alpha_1 L+\Gamma_2\alpha_2 L}{\Pi_1-\Pi_2}\right)Y_t=\left(k-\frac{\Gamma_1\Pi_2\alpha_1 L+\Gamma_2\Pi_1\alpha_2 L}{\Pi_1-\Pi_2}\right)X_t+u_t, \tag{E14}$$

where

$$\Gamma_1 = \Pi_1 - k$$
$$\Gamma_2 = k - \Pi_2$$
$$u_t = W_{1t} - kW_{2t}$$
$$k = \frac{Cov(W_{1t}, W_{2t})}{Var(W_{2t})} = \frac{Cov(\Pi_{1t}v_{1t} + \Pi_2 v_{2t}, v_{1t} + v_{2t})}{Var(v_{1t} + v_{2t})}.$$

The aggregate consumption function of the economy is given by

$$Y_t = \left(\frac{\Gamma_1\alpha_1 L + \Gamma_2\alpha_2 L}{\Pi_1 - \Pi_2}\right)Y_{t-1} + kX_t - \left(\frac{\Gamma_1\Pi_2\alpha_1 L + \Gamma_2\Pi_1\alpha_2 L}{\Pi_1 - \Pi_2}\right)X_{t-1} + u_t. \qquad \text{(E15)}$$

Appendix F: The SMD Theorem

There are several variants of the SMD theorem available. To state the theorem in its simplest form, following Kirman (1989), consider an exchange economy in which there are a finite number l of goods and N consumers. Define the following notations:

$e(a)$: a positive bundle of initial endowment of all goods for individual a;
$\phi(a, \mathbf{p})$: a demand function for individual a derived from a strictly convex monotone utility function, with $\boldsymbol{P}$ being the price vector;
$Z(a, \mathbf{p}) = \phi(a, \mathbf{p}) - e(a)$: the excess demand for individual a;
$Z(\mathbf{p}) = \sum Z(a, \mathbf{p})$: the aggregate excess demand function of the economy obtained by summing over the excess demands of the N individuals.

For this economy, the SMD theorem reads as follows (Kirman, 1989:129):

Theorem: Given a continuous function $f : \mathbf{p} \to R^l$ satisfying Walras' Law, i.e., $\mathbf{p}f(\mathbf{p}) = 0$ for $\boldsymbol{p}$ in $\boldsymbol{P}$, then for any positive ε there is an economy ε with consumers with strictly convex monotone preferences such that

$$f(\mathbf{p}) = Z_\varepsilon(\mathbf{p}), \text{ for all } p \text{ in } \Delta_\varepsilon.$$

Here $Z_\varepsilon(.)$ is the excess demand of the economy ε and Δ_ε is the price simplex with prices above ε, i.e.,

$$\{\mathbf{p} \mid \textstyle\sum_i p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i\}.$$

The result states that, for any arbitrary function $f : \Delta_\varepsilon \to R^l$ satisfying the Walras Law, there is an economy with N consumers with strictly convex monotone preferences whose excess demand function for prices in Δ_ε coincides with the arbitrary function. This means the aggregate excess demand function of an economy of N consumers with strictly convex monotone preferences can be any arbitrary continuous function satisfying the Walras Law. The standard restrictions on the preferences do not further restrict the class of functions to which the excess demand function belongs.

Finale

> The moral ... is this: if you put very little in, you get very little out (Sonnenschein, 1973:405).

This thesis has studied some general issues at the heart of the theoretical approach to macroeconomics. The issues relate to the possibility of establishing an explanatory and predictive microeconomic theory, *and* transforming it into a theory of the economy as a whole using aggregation methods. It is now time to bring together the results of the analysis:

Early in the thesis, we showed that the proposal that *homo economicus* behaves like a decision scientist, understood in terms of one or another expected utility theory, contributes very little to the understanding of behavioural matters and hence economic phenomena. These theories take as given how the agent specifies his choice situation and defines his decision problem. They only state how he solves an already well-structured decision problem. But accurate prediction and explanation of behaviour depend critically on how the agent models his choice situation and defines his decision problem, rather than on how he solves an already well-structured decision problem. To predict how an agent models his choice situation, and defines his decision problem, one needs a theory of how he processes information, models the causal structure of his choice situation, adapts goals, forms preferences, and modifies them as a result of subsequent experiences or new information. Without such a theory, there is no prospect for accurately predicting or explaining behaviour in a dynamic and changing environment.

The proposal to model *homo economicus* as an intuitive econometrician is an intriguing and substantive step towards understanding how the agent models his choice situation and modifies it in response to new information. The trouble is that there is no 'tight enough' theory of statistical learning capable of fully, and accurately, explaining the central phases of learning from data – in particular model formulation and re-specification. Reflection on nonparametric inference reveals that there can be no algorithm that receives an ordinarily sized sample and yields the model that, given the data, best approximates the underlying data generating mechanism. The choice of a model at a deep level requires various subjective judgements. With ordinarily sized samples, even nonparametric learning of an interpretable model of few variables, representing a simple choice situation, is theoretically impossible. In real-life inference situations, learning of an interpretable model of several variables calls for starting with a parametric model.

However, any statistical theory of parametric learning necessarily presupposes a reservoir of models or, more precisely, a reservoir of basic probabilistic assumptions that can be used for creating models. It also requires knowledge of the pre-estimation implications of the models, and methods for exploring their post-estimation consequences. None of these can be explained within a statistical theory of parametric inference. Therefore, within the framework of the intuitive statistician hypothesis, any explanation of how the agent comes to model his choice situation is necessarily bound to be incomplete.

Statistical theories of causal inference are also of limited power. Because of the possibility of selection bias, mistaking concomitants for genuine causes, taking barren proxies for real causes, aggregation over heterogeneous units, and so on, the class of explanations possible in general for a statistical dependency or independency is larger than the class of causal explanations. As a result, an essential step in drawing causal inferences from observational data is to first exclude non-causal explanations. Only then do statistical tools become relevant for inferring causal conclusions. Even at this stage, statistical analysis can at best infer a class of statistically indistinguishable models, which in practice usually have little or nothing in common. Selecting a causal model calls for substantive causal background information at every level. However, for the reasons mentioned above, this information cannot come from a theory of statistical learning.

One outcome of this analysis is that the description level at which the econometrician (statistician) works is inappropriate for establishing a predictive model of human learning. To specify how a person processes data, conceptualizes his environment, models his choice situation, defines his decision problem, and learns from experience, it is necessary to work at a far deeper, and more refined, level of description. One, in particular, needs to establish a theory of cognition, object representation, pattern recognition, and even preference formation, as well as a detailed history of the person's experiences (Arthur, 1994). A precise theory of human cognition and decision-making may eventually arise. However, because of the level of description the theory is defined for, the theory may not be of much use for economic analysis. The basic problem in establishing a predictive theory of economic behaviour is of mismatched levels – a useful theory of behaviour may require working at a description level useless to economics.

The connection between the individual and aggregate levels is also highly complex. To explain large-scale economic phenomena it is necessary to view the economy as a society of interactive, and heterogeneous, agents. However, the regularities that emerge at the aggregate level in an interactive and heterogeneous economy are not directly related to the laws operating at the micro level. The regularities are the joint outcome of individual interactions *and* the processes characterising the physical and institutional environment. In light of this, modelling the emergent regularities requires starting with a great deal of information about the structure of the economy. It is, therefore, wrong to attribute any purely behavioural interpretation to the regularities. Moreover, due to the ubiquitous existence of multiple equilibria in models of interactions, the relations that emerge among economic aggregates are statistical. They are not causal.

Bibliography

Abrevaya, J. and W. Jiang (2004). A Nonparametric Approach to Measuring and Testing Curvature, Columbia University, Graduate School of Business. **2004**.

Aigner, D. and S. Goldfeld (1974). Estimation and Prediction from Aggregate Data when Aggregates are Measured more Accurately than Their Components. *Econometrica* **42**: 113-34.

Akaike, H. (1970). Statistical Predictor Information. *Annals of the Institute of Statistical Mathematics* **22**: 203-17.

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **AC-19**: 716-23.

Akerman, F. (1999). Still Dead after all These Years: Interpreting the Failure of General Equilibrium Theory, Global Development and Environment Institute, Tufts University.

Alderich, J. (1989). Autonomy. *History and Methodology of Econometrics*. N. de Marchi and C. L. Gilbert. Oxford: 15-35.

Allais, M. (1953). Le Comportement de l'Homme Rationnel devant le Risque: Critique des Postulates et Axioms de l'Ecole Americanine. *Econometrica* **21**: 503-546.

Allais, M. (1997). An Outline of My Main Contributions to Economic Science. *The American Economic Review* **87**: 3-12.

Altaman, N. S. (1992). An Introduction to Kernel and Nearest-neighbour Nonparametric Regression. *American Statistician* **46**: 175-185.

Amemiya, T. (1980). Selection of Regressors. *International Economic Review* **21**: 331-354.

Anand, P. (1993). *Foundations of Rational Choice Under Risk*. Oxford, Clarendon Press.

Anscombe, F. J. (1963). Tests of Goodness of Fit. *Journal of the Royal Statistical Society Series B* **25**: 81-94.

Anscombe, F. J. and R. J. Aumann (1963). A Definition of Subjective Probability. *Annals of Mathematical Statistics* **34**: 199-205.

Antonelli, G. B. (1986, 1971). Sulla Teoria Matematica dell'Economia Politica, Pisa; English translation in. *Preferences, Utility and Demand: A Minnesota Symposium*. J. S. Chipman, L. Hurwicz, M. K. Richter and H. F. Sonnenschein. New York, Harcourt Brace Jovanovich: 333-360.

Armendt, B. (1980). Is There a Dutch Book Argument for Probability Kinematics? *Philosophy of Science* **47**: 583-588.

Armendt, B. (1992). Dutch Strategies for Diachronic Rules: When Believers See the Sure Loss Coming. *Proceedings of the Biennial Meetings of the Philosophy of Science Association*. **1**: 217-229.

Arntzenius, F. (1999). Reichenbach's Common Cause Principle. *Stanford Encyclopaedia of Philosophy (on the World Wide Web).*

Arrow, K. (1968). Economic Equilibrium. *International Encyclopaedia of the Social Sciences.* London, Macmillan. **4:** 376-89.

Arrow, K. (1982). Risk Perception in Psychology and Economics. *Economic Inquiry* **20**: 1-9.

Arrow, K. (1986). Rationality of Self and Others in an Economic System. *The Journal of Business* **59**: pp.S385-S399.

Arrow, K. (1987). Rationality of Self and Others in an Economic System. *Rational Choice: the Contrast between Economics and Psychology.* R. Hogarth and M. Reder. Chicago, London, University of Chicago Press**:** 201-215.

Arrow, K. (1994). Methodological Individualism and Social Knowledge. *American Economic Review* **84**: 1-9.

Arthur, B. (2000). Cognition: The Black Box of Economics. *The Complexity Vision and the Teaching of Economics.* D. Colander. Northampton, Mass, Edward Elgar Publishing**:** Chapter 3.

Arthur, W. B. (1991). Designing Economic Agents that Act like Human Agents: A Behavioural Approach to Bounded Rationality. *Learning and Adaptive Economic Behaviour* **81(2)**: 353-9.

Arthur, W. B. (1993). On Designing Economic Agents that Behave like Human Agents. *Journal of Evolutionary Economics* **3**: 1-22.

Arthur, W. B. (1994). Inductive Reasoning and Bounded Rationality. *American Economic Review* **84**: 406-411.

Arthur, W. B., J. H. Holland, et al. (1997). Asset Pricing under Endogenous Expectations in an Artificial Stock Market. *Economic Notes* **26**: 297-330.

Aumann, R. J. (1987). Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica* **55**: 1-18.

Barker, T. S. and M. H. Pesaran, Eds. (1990). *Disaggregation in Econometric Modelling.* London, Rutledge.

Barnard, G. A. (1962). Prepared Contribution and Discussion. *Foundations of Statistical Inference.* L. J. Savage. London, Methuen**:** 39-49.

Barnett, V. (1999). *Comparative Statistical Inference.* New York, Wiley.

Basmann, R. L. (1972). The Brookings Quarterly Econometric Model: Science or Number Mysticism? *Problems and Issues in Current Econometric Practice.* K. Brunner. Columbus, Ohio, College of Administrative Science, Ohio State University**:** Chapter 1.

Bayarri, M. J. and J. O. Berger (1999). Quantifying Surprise in the Data and Model Verification (with discussion). *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting.* J. M. Bernardo, J. O. Berger, A. P. Dawid and A. Smith. Oxford, Oxford University Press. **6:** 53-82.

Bayes, T. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society* **53**: 370-418.

Becker, G. S. (1962). Irrational Behaviour and Economic Theory. *Journal of Political Economy* **70**: 1-13.

Becker, G. S. (1976). *The Economic Approach to Human Behaviour.* Chicago ; London, University of Chicago Press.

Becker, G. S. (1981). *A Treatise on the Family.* Cambridge, MA, Harvard University Press.

Begg, D. K. H. (1982). *The Rational Expectations Revolution in Macroeconomics.* Oxford, Philip Alan.

Berger, J. (1980). *Statistical Decision Theory: Foundations, Concepts, and Methods.* New York, Springer.

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, Springer.

Berger, J. O. (1999). Bayes Factors. *Encyclopaedia of Statistical Sciences.* S. Kotz, B. R. Campbell and D. L. Banks. New York, Wiley. **3**: 20-29.

Berger, J. O. and M. Delampady (1987). Testing Precise Hypotheses (with discussion). *Statisitcal Science* **2**: 317-352.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian theory.* New York; Chichester, John Wiley.

Bernheim, B. D. (1984). Rationalizable Strategic Behaviour. *Econometrica* **52**: 1007-1028.

Bernheim, B. D. (1986). Axiomatic Characterizations of Rational Choice in Strategic Environments. *Scand. Journal of Economics* **88**: 473-488.

Bewley, T. F., Ed. (1987). *Advances in Econometrics: Fifth World Congress.* Cambridge, Cambridge University Press.

Bicchieri, C. (1987). Rationality and Predictability in Economics. *The British Journal for the Philosophy of Science* **38**: 501-513.

Bicchieri, C., R. Jeffrey, et al., Eds. (1997). *The Dynamics of Norms.* Cambridge, New York, Cambridge University Press.

Bierens, H. J. (1987). *Kernel Estimators of Regression Functions.* Cambridge, Cambridge University Press.

Birnbaum, A. (1962). On the Foundations of Statistical Inference (with discussion). *Journal of the American Statistical Association* **57**: 269-306.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition.* Oxford, Oxford University Press.

Black, F. (1982). The Trouble with Econometric Models. *Financial Analysis Journal* **38**(March-April): 29-37.

Blalock, H. M. (1964). *Causal Inferences in Nonexperimental Research.* Chapel Hill, The University of North Carolina Press.

Blalock, H. M. (1968). Theory Building and Causal Inference. *Methodology in Social Research.* H. M. J. Blalock and H. M. J. Blalock. London, McGraw-Hill.

Blalock, H. M. (1972). *Causal Models in the Social Sciences.* London, Macmillan.

Blalock, H. M. (1972). Four-variable Causal Models and Partial Correlations. *Causal Models in the Social Sciences.* H. M. Blalock. London, Macmillan.

Blalock, H. M. J. and A. B. Blalock, Eds. (1968). *Methodology in Social Research.* London, McGraw-Hill.

Blanchard, O. and S. Fisher (1989). *Lectures on Macroeconomics.* Cambridge, Mass., MIT Press.

Blanchard, O. J. (1979). Backward and Forward Solutions for Economics with Rational Expectations. *American Economic Review* **69**: 114-8.

Blanchard, O. J. and M. Watson (1982). Bubbles, Rational Expectations, and Financial Markets. *Crises in the Economic and Financial Structure: Bubbles, Bursts, and Shocks.* P. Wachtel. Lexington, Mass, Lexington Books: 295-315.

Blume, L. E. and D. Easley (1995). What Has the Rational Learning Literature Taught Us? *Learning and Rationality in Economics.* A. Kirman and M. Salmon. Oxford, Blackwell: 13-39.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York, John Wiley and Sons.

Boudon, R. (1968). A New Look at Correlation Analysis. *Methodology in Social Research.* H. M. J. Blalock and A. B. Blalock. London, McGraw-Hill: 199-235.

Bowles, S. (1998). Endogenous Preferences: The Cultural Consequences of Markets and other Economic Institutions. *Journal of Economic Literature* **XXXVI**: 75-111.

Bowles, S. and H. Gintis (2000). Walrasian Economics in Perspective. Amherst, Massachusetts, Department of Economics, University of Massachusetts.

Bowman, A. and A. Azzalini (1997). Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations. London, Oxford University Press.

Box, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness (with discussion). *Journal of the Royal Statistical Society, A.* **143**: 383-430.

Box, G. E. P. (1983). An Apology for the Ecumenism in Statistics. *Scientific Inference, Data Analysis, and Robustness.* G. E. P. Box, T. Leonard and C.-F. Wu. New York, Academic Press: 51-84.

Box, G. E. P. (1994). Statistics and Quality Improvement. *Journal of the Royal Statistical Society, Series A* **157**: 209-229.

Box, G. E. P. and G. M. Jenkins (1976). *Time Series Analysis: Forecasting and Control.* London, Holden Day.

Box, G. E. P., T. Leonard, et al., Eds. (1983). *Scientific Inference, Data Analysis, and Robustness*. New York, Academic Press.

Bradley, J. V. (1968). *Distribution-Free Statistical Tests*. Englewood Cliffs, New Jersey, Prentice-Hall.

Bray, M. (1982). Learning, Estimation, and Stability of Rational Expectations. *Journal of Economic Theory* **26**: 318-339.

Bray, M. (1983). Convergence to Rational Expectations Equilibrium. *Individual Forecasts and Aggregate Outcomes*. R. Frydman and E. S. Phelps, Cambridge University Press.

Bray, M. (1989). Rational Expectations, Information, and Asset Markets. *The Economics of Missing Markets, Information, and Games*. F. Hahn. Oxford, Clarendon Press: 243-278.

Bray, M. and N. Savin (1986). Rational Expectations Equilibria, Learning, and Model Specification. *Econometrica* **54**: 1129-1160.

Breckler, S. (1990). Applications of Covariance Structure Modelling in Psychology: Cause for Concern? *Psychological Bulletin* **107**: 260-273.

Breiman, L. (1996). Heuristics of Instability and Stabilisation in Model Selection. *Annals of Statistics* **24**: 2350-83.

Breiman, L. and D. Freedman (1983). How Many Variables Should Be Entered in a Regression Equation? *Journal of the American Statistical Association*: 131-136.

Breiman, L., J. H. Friedman, et al. (1984). *Classifcation and Regression Trees*, Wadsworth International Group.

Breiman, L. and P. Spector (1992). Sub-model Selection and Evaluation in Regression: The *x*-Random Case. *International Statistical Review* **60**: 291-319.

Brighi, L. and M. Forni (1989). Aggregation across Agents in Demand Systems. Badia Fiesolana, San Domenico, European University Institute: 38.

Browne, M. W. (2000). Cross-validation Methods. *Journal of Mathematical Psychology* **44**: 108-132.

Bryant, J. (1994). Coordination Theory, The Stag Hunt and Macroeconomics. *Problems of Coordination in Economic Activity*. J. W. Friedman. Norwell, MA, Kluwer Academic Publisher: 207-227.

Bryant, J. (1996). Team Coordination Problems and Macroeconomic Models. *Beyond Microfoundations: Post Walrasian Macroeconomics*. D. Colander. Cambridge, Cambridge University Press: 157-171.

Bullard, J. (1994). Learning Equilibria. *Journal of Economic Theory* **64**: 468-85.

Burns, A. F. and W. C. Mitchell (1964). *Measuring Business Cycles*. New York.

Buse, A. (1992). Aggregation, Distribution and Dynamics in the Linear and Quadratic Expenditure Systems. *Review of Economic Statistics* **74**: 45-53.

Busemeyer, J. R. and W. Yi-Min (2000). Model Comparison and Model Selections Based on Generalization Criterion Methodology. *Journal of Mathematical Psychology* **44**: 171-189.

Bush, R. R. and F. Moseteller (1955). *Stochastic Models for Learning*. New York, Wiley.

Cagan, P. (1956). The Monetary Dynamics of Hyperinflation. *Studies in Quantity Theory of Money*. M. Friedman University of Chicago Press: 25-120.

Calvo, C. A. (1979). On Models of Money and Perfect Foresight. *International Economic Review,* **20**: 83-103.

Camerer, C. (1995). Individual Decision Making. *The Handbook of Experimental Economics*. J. K. Kagel, Roth, A.E. Princeton, Princeton University Press: 587-703.

Cameron, A. C. (1990). Aggregation in Discrete Choice Models: An Illustration of Nonlinear Aggregation. *Disaggregation in Economic Modelling*. T. S. Barker and M. H. Pesaran. London, Rutledge: 206-234.

Carlin, B. P. and T. A. Louis (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London; New York, Chapman & Hall.

Cartwright, N. (1989). *Nature's Capacities and Their Measurement*. Oxford, Oxford University Press.

Cartwright, N. (1995). Probabilities and Experiments. *Journal of Econometrics* **67**: 47-59.

Cartwright, N. (1997). What is the Causal Structure. *Causality In Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. V. R. McKim and S. P. Turner. Notre Dame, University of Notre Dame Press: 343-357.

Cartwright, N. (1999). *The Dappled World: Essays on the Perimeter of Science*. New York, Cambridge University Press.

Cartwright, N. (2001). What Is Wrong with Bayes Nets? *The Monist* **84**: 242-64.

Cartwright, N. (2002). Against Modularity, the Causal Markov Condition and Any Link between the Two: Comments on Hausman and Woodward. *British Journal for the Philosophy of Science* **53**: 411-453.

Chari, V. V. (1999). Nobel Laureate Robert E. Lucas, Jr.: Architect of Modern Macroeconomics. *Federal Reserve Bank of Minneapolis Quarterly Review* **23**: 2-12.

Chatfield, C. (1985). The Initial Examination of Data (with discussion). *Journal of the Royal Statistical Society, Series A* **148**: 214-253.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical Inference. *Journal of the Royal Statistical Society, Series A* **158**: 419-466.

Chen, X. and H. White (1998). Nonparametric Adaptive Learning with Feedback. *Journal of Economic Theory* **82**: 190-222.

Cheng, B. and D. M. Titterington (1994). Neural Networks: A Review from a Statistical Perspective. *Statistical Science* **9**: 2-54.

Cheng, P. W. and K. J. Holyoake (1995). Complex Adaptive Systems as Intuitive Statisticians: Causality, Contingency, and Prediction. *Comparative Approaches to Cognition.* J. A. Meyer and H. Roitblat. Cambridge, MA, MIT Press: 271-302.

Cherkassky, V., J. H. Friedman, et al., Eds. (1994). *From Statistics to Neural Networks, Theory and Pattern Recognition Applications.* Nato ASI Series, Springer-Verlag 1.

Chickering, D. (1995). A Transformational Characterization of Bayesian Network Structures. *Uncertainty in Artificial Intelligence.* P. Besnard and S. Hanks. San Franscisco, Morgan Kaufmann. **II:** 87-98.

Cho, I.-K. (1994). Bounded Rationality, Neural Network and Folk Theorem in Repeated Games with Discounting. *Economic Theory* **4**: 935-957.

Christensen, D. (1991). Clever Bookies and Coherent Beliefs. *Philosophical Review* **100**: 229-47.

Christensen, D. (1996). Dutch Books Depragmatized: Epistemic Consistency for Partial Believers. *The Journal of Philosophy* **93**: 450-79.

Clarida, R. H. (1991). Aggregate Stochastic Implications of The Life Cycle Hypothesis. *The Quarterly Journal of Economics* **106**: 851-69.

Clogg, C. C. and A. Haritou (1997). The Regression Method of Causal Inference and a Dilemma Confronting This Method. *Causality in Crisis.* V. McKim and S. Turner, University of Notre Dame Press: 83-112.

Colander, D. (1992). New Keynesian Economics in Perspective. *Eastern Economic Journal* **18**: 438-48.

Colander, D. (1993). The Macrofoundations of Micro. *Eastern Economic Journal* **19**: 447-57.

Colander, D. (1996). *Beyond Microfoundations: Post Walrasian Macroeconomics.* Cambridge , New York, Cambridge University Press.

Coleman, J. S. (1990). *Foundations of Social Theory.* Cambridge, Mass., Harvard University Press.

Collingwood, R. G. (1948). *An Essay on Metaphysics.* Oxford, Clarendon Press.

Congdon, P. (2001). *Bayesian Statistical Modelling.* Chichester, New York, John Wiley.

Conslik, J. (1996). Why Bounded Rationality? *Journal of Economic Literature* **34**: 669-700.

Cooley, T. F. and S. LeRoy (1985). Atheoretical Macroeconomics: A Critique. *Journal of Monetary Economics* **16**: 283-308.

Cooley, T. F., S. F. LeRoy, et al. (1984). Economic Policy Evaluation: Note. *The American Economic Review* **74**: 467-470.

Cooper, G. (1995). Causal Discovery from Data in the Presence of Selection Bias. *Proceedings of the Workshop on Artificial Intelligence and Statistics*: 140-150.

Cooper, G. (2000). A Bayesian Method for Causal Modeling and Discovery Under Selection. *Proceedings of the Conference on Uncertainty in Artificial Intelligence (2000)*: 98-106.

Cooper, R. W. (1999). *Coordination Games: Complementarities and Macroeconomics.* Cambridge, Cambridge University Press.

Cox, D. R. (1958). *The Planning of Experiments.* New York, John Wiley and Sons.

Cox, D. R. (1958). Some Problems Connected with Statistical Inference. *The Annals of Mathematical Statistics* **29**: 357-372.

Cox, D. R. (1977). The Role of Significance Tests (with Discussion). *Scandinavian Journal of Statistics* **4**: 49-70.

Cox, D. R. (1992). Causality: Some Statistical Aspects. *Journal of the Royal Statistical Society, Series A* **155**: 291-301.

Cox, D. R. and N. J. H. Small (1978). Testing Multivariate Normality. *Biometrika* **65**: 263-272.

Cox, D. R. and A. Stuart (1955). Some Quick Sign Tests for Trend in Location and Dispersion. *Biometrika* **42**: 80-95.

Craven, P. and G. Whaba (1979). Smoothing Noisy Data with Spline Functions. *Numerische Mathematik* **31**: 377-403.

Cyert, R. and M. H. DeGroot (1974). Rational Expectations and Bayesian Analysis. *Journal of Political Economy* **82**: 521-536.

D'Agostino, R. B. (1986). Graphical Analysis. *Goodness-of-Fit Techniques.* R. B. D'Agostino and M. A. Stephenes. New York, Marcel Dekker, INC: 7-62.

D'Agostino, R. B. and M. A. Stephenes, Eds. (1986). *Goodness-of-Fit Techniques.* New York and Basel, Marcel Dekker.

Darnell, A. C. and J. L. Evans (1990). *The Limits of Econometrics.* Aldershot, U.K., Elgar.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series A* **41**: 1-31.

Dawid, A. P. (1982). Intersubjective Statistical Models. *Exchangeability in Probability and Statistics.* G. Koch and F. Spizzichino. Amsterdam, North-Holland: 217-232.

Dawid, A. P. (1983). Invariant Prior Distributions. *Encyclopaedia of Statistical Sciences.* S. Kotz and N. L. Johnson. New York, John Wiley: 228-236.

Dawid, A. P. (1986). Comment. *Statistical Science* **1**: 488-492.

Dawid, A. P. (2002). Probability, Causality and Empirical World: A Bayes-de Finetti-Popper-Borel Synthesis. Technical Report, University College London.

De Finetti, B. (1972). *Probability, Induction, and Statistics.* New York, Wiley.

De Finetti, B. (1980). Foresight: Its Logical Laws, Its Subjective Sources. *Studies in Subjective Probability.* H. E. Kyburg, Jr. and H. E. Smokler. New York, Krieger: 51-131.

Deaton, A. (1992). *Understanding Consumption.* Oxford and New York, Oxford University Press.

Deaton, A. and J. Muellbauer (1980). *Economics and Consumer Behaviour.* Cambridge, Cambridge University Press.

Deaton, A. S. (1982). Model Selection Procedures, or, Does the Consumption Function Exist? *Evaluating the Reliability of Macroeconometric Models.* G. C. Chow and P. Corsi. New York, Wiley: 43-69.

Debreu, G. (1959). *The Theory of Value.* New York, Wiley.

Debreu, G. (1974). Excess Demand Functions. *Journal of Mathematical Economics* **1**: 15-23.

DeGroot, M. H. (1982). Comment. *Journal of the American Statistical Association* **77**: 336-339.

Demiralp, S. and K. D. Hoover (2003). Searching for the Causal Structure of a Vector Autoregression. *Oxford Bulletin of Economics and Statistics* **65, Supplement**: 745-767.

Dempster, A. P. (1971). Model Searching and Estimation in the Logic of Inference (with discussion). *Foundations of Statistical Inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ont., 1970).* Toronto, Ont., Holt, Rinehart and Winston of Canada: 56-81.

Dempster, A. P. (1980). Bayesian Inference in Applied Statistics. *Bayesian Statistics.* J. M. Bernardo, M. DeGroot and D. V. Lindley. Valencia, University press: 255-79 (with discussion 279-91).

Dempster, A. P. (1983). Purposes and Limitations of Data Analysis. *Scientific Inference, Data Analysis, and Robustness.* G. E. P. Box, T. Leonard and C.-F. Wu. New York, Academic: 117-133.

Dempster, A. P., N. M. Laird, et al. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**: 1-38.

Dhrymes, P. J., et al. (1972). Criteria for Evaluation of Econometric Models. *Annals of Economic and Social Measurement* **1**: 291-324.

Diaconis, P. and M. Shahshahani (1984). On Nonlinear Functions of Linear Combinations. *SIAM Journal of Scientific and Statistical Computing* **5**: 175-191.

Diaconis, P. and S. L. Zabell (1982). Updating Subjective Probability. *Journal of the American Statistical Association* **77**: 822-30.

Diaconis, P. and S. L. Zabell (1985). Some alternatives to Bayes' rule. *Information and Group Decision Making, Proc. Second Univ. of Calif. Irvine Conf. Political Economy.* B. Grofman, G. Owen. Greenwich, CT, Jai Press: 25-38.

Dow, S. (1988). Post Keynesian Economics: Conceptual Underpinnings. *British Review of Economic Issues* **10**: 1-18.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with Discussion). *Journal of the Royal Statistical Society, Series B* **57**: 45-97.

Draper, D. (1996). Utility, Sensitivity Analysis, and Cross-validation in Bayesian Model-checking. Discussion of "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies," by A. Gleman et al. *Statistica Sinica* **6**: 28-35.

Draper, D., J. S. Hodges, et al. (1993). Exchangeability and Data Analysis (with discussion). *Journal of the Royal Statistical Scoiety, Series A* **156**: 9-37.

Dreze, J., H. (1987). *Essays on Economic Decisions under Uncertainty.* Cambridge, Cambridge University Press.

du Toit S.H.C, A. G. W Steyn, et al. (1986). *Graphical Exploratory Data Analysis.* Verlag, Springer.

Duncan, O. (1975). *Introduction to Structural Equation Models.* New York, Academic Press.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory.* Cambridge, MA, MIT Press.

Edgeworth, E. Y. (1881). *Mathematical Physics.* London, Rutledge.

Edwards, W., H. Lindman, et al. (1963). Bayesian Statistical Inference for Psychological Research. *Psychological Review* **70**: 193-242.

Efron, B. (1983). Estimating the Error Rate of a Prediction Rule: Some Improvement on Cross Validation. *Journal of the American Statistical Association* **78**: 316-33.

Efron, B. (1984). Comparing Non-nested Linear Models. *Journal of the American Statistical Association* **79**: 791-803.

Efron, B. (1986). How Biased is the Apparent Error Rate of a Prediction Rule? *Journal of the American Statistical Association* **81**: 461-470.

Efron, B. and G. Gong (1983). A Leisurely look at the Bootstrap, the Jacknife, and Cross-validation. *American Statistician* **37**: 36-48.

Efron, B. and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* New York, London, Chapman and Hall.

Efron, B. and R. J. Tibshirani (1997). Improvements on Cross-Validation: The .632+ Bootstrap Method. *Journal of the American Statistical Association* **92**: 548-560.

Ellery, E. (1982). *Rational Decision and Causality.* Cambridge, Cambridge University Press.

Engle, R., D. Hendry, et al. (1983). Exogeneity. *Econometrica* **51**: 277-304.

Epstein, R. J. ·(1987). *A History of Econometrics.* Amsterdam, Elsevier Science Publishers.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* New York and Basel, Marcel Dekker, INC.

Evans, G. W. and S. Honkapohja (2001). *Learning and Expectations in Macroeconomics.* Princeton, N.J., Princeton University Press.

Fair, R. C. (1978). The Effect of Economic Events on Votes for President. *The Review of Economics and Statistics* **LX**: 159-173.

Fair, R. C. (1987). Macroeconomic Models. *The New Palgrave: A Dictionary of Economics*. J. Eatwell. London, Macmillan. **3:** 269-273.

Fair, R. C. (1994). *Testing Macroeconometric Models.* Cambridge and London, Harvard University Press.
Faraway, J. (1998). Data Splitting Strategies for Reducing the Effect of Model Selection on Inference. *Computing Science and Statistics* **30**: 332-341.

Felipe, J. and F. M. Fisher (2003). Aggregation in Production Functions: What Applied Economists Should Know. *Macroeconomica* **54**: 208-262.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications.* New York; Chichester, Wiley.

Fine, T. L. (1973). *Theories of Probability.* New York, Academic Press.

Fischhoff, P. C. (1983). Predicting Frames. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **9**: 103-16.

Fishburn, P. C. (1970). *Utility Theory for Decision Making.* New York, Wiley.

Fishburn, P. C. (1981). Subjective Expected Utility: A Review of Normative Theories. *Theory and Decision* **13**: 139-199.

Fisher, F. M. (1987). Aggregation Problem. *The New Palgrave, A Dictionary of Economics.* New York, Stockton Press. **1:** 53-5.

Fisher, F. M. (1989). Games Economists Play: A Noncooperative View. *RAND Journal of Economics* **20**(113-23).

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society of London A* **222**: 309-368.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* **7**: 179-188.

Flavin, M. A. (1981). The Adjustment of Consumption to Changing Expectations about Future Income. *Journal of Political Economy* **89**: 1020-1037.

Fox, J. (1984). *Linear Statistical Models and Related Methods.* New York, John Wiley.

Fox, J. (1990). Describing Univariate Distributions. *Modern Methods of Data Analysis.* J. Fox and J. S. Long. London, Sage Publications: 59-125.

Fox, J. and J. S. Long, Eds. (1990). *Modern Methods of Data Analysis.* London, Sage Publications.

Freedman, D. A. (1981). Some Pitfalls in Large Econometric Models: A Case Study. *Journal of Business* **54**: 497-500.

Freedman, D. A. (1983). A Note on Screening Regression Equations. *The American Statistician* **37**: 152-155.

Freedman, D. A. (1987). As Others See Us: A Case Study in Path Analysis, (with Discussion). *Journal of Educational Statistics* **12**: 101-223.

Freedman, D. A. (1991). Statistical Models and Shoe Leather (with discussion). *Sociological Methodology* **21**: 291-358.

Freedman, D. A. (1997). From Association to Causation via Regression. *Causality in Crisis? Statistical Methods and Search for Causal Knowledge in the Social Sciences.* V. R. McKim and S. P. Turner. Indiana, University of Notre Dame Press: 113-82.

Friedman, B. M. and F. H. Hahn (1990). Preface to the Handbook. *Handbook of Monetary Economics.* B. M. Friedman and F. H. Hand. Amsterdam, Elsevier Science Publishers. **1**.

Friedman, D., D. W. Massaro, et al. (1995). A Comparison of Learning Models. *Journal of Mathematical Psychology* **39**: 164-178.

Friedman, J. H. (1994). An Overview of Predictive Learning and Function Approximation. *From Statistics to Neural Networks, Theory and Pattern Recognition Applications.* V. Cherkassky, J. H. Friedman and H. Wechsler. Berlin, Springer: 1-61.

Friedman, J. H. and W. Stuelzle (1981). Projection Pursuit Regression. *Journal of the American Statistical Association* **76**: 817-823.

Friedman, J. W. (1994). Problems of Coordination in Economic Activity. Boston, Dordrecht, London, Kluwer Academic Publishers.

Friedman, M. (1953). *Essays in Positive Economics.* Chicago and London, Chicago University Press.

Friedman , M. (1974). Explanation and Scientific Understanding. *Journal of Philosophy* **71**: 5-19.

Friedman , M. and J. Savage (1952). The Expected Utility hypothesis and the Measurability of Utility. *Journal of Political Economy* **60**: 464-474.

Friedman, M. and A. J. Schwartz (1963). *A Monetary History of the United States 1867-1960,* Princeton University Press for the National Bureau of Economic Research.

Frydenberg, M. (1990). The Chain Graph Markov Property. *Scandinavian Journal of Statistics* **17**: 333-353.

Frydman, R. and E. S. Phelps (1984). *Individual Forecasting and Aggregate Outcomes: "Rational Expectations" Examined.* Cambridge Cambridgeshire; New York, Cambridge University Press.

Fudenberg, D. and J. Tirole (1991). *Game Theory.* Cambridge, Mass., MIT Press.

Fuller, W. (1976). *Introduction to Stochastic Time Series.* New York, Wiley.

Galambos, J. (1982). Characterizations of Distributions. *Encyclopaedia of Statistical Sciences.* New York, Wiley. **1**: 422-428.

Gasking, D. (1955). Causation and Recipes. *Mind* **64**: 479-87.

Geiger, D., Paz, A., and Pearl, J. (1990). Learning Causal Trees from Dependence Information. *Proceedings, AAAI-90,*. Boston, MA.: 770-776.

Geiger, D. and J. Pearl (1988). On the Logic of Causal Models. *Uncertainty in Artificial Intelligence*. L. Kanal. Amsterdam, North-Holland Publishing Co.: 3-14.

Geiger, D., T. S. Verma, et al. (1990). Identifying Independence in Bayesian Networks. *Networks* **20**: 507-534.

Geisser, S. (1975). The Predictive Sample Reuse Method with Application. *Journal of the American Statistical Association* **10**: 320-328.
Geisser, S. and W. Eddy (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association* **74**: 153-160.

Gelfand, A. and D. Dey (1994). Bayesian Model Choice: Asymptotics and Exact Calculation. *Journal of the Royal Statistical Society, B.* **56**: 501-514.

Gelfand, A. E., D. K. Dey, et al. (1992). Model Determination Using Predictive Distributions, with Implementation via Sampling-Based Methods (with discussion). *Bayesian Statistics*. J. M. Bernardo, et al. Oxford, Oxford University Press. **4**: 147-167.

Gelman, A. (2002). Exploratory Data Analysis for Complex Models. New York, Columbia University: 1-26.

Gelman, A., J. B. Carlin, et al. (1995). *Bayesian Data Analysis*. London New York, Chapman & Hall.

Gelman, A., X. L. Meng, et al. (1996). Posterior Predictive Assessment of Model Fitness via Realized Discrepancies. *Statistica Sinica* **6**: 733-807.

Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **4**: 1-58.

Geweke, J. (1985). Microeconomic Modelling and the Theory of Representative Agent. *American Economic Review* **75**: 206-10.

Geweke, J. (1999). Simulation Methods for Model Criticism and Robustness Analysis. *Bayesian Statistics*. J. M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M Smith. Oxford, Oxford University Press. **6**: 53-82.

Geweke, J. and W. McCausland (2001). Bayesian Specification Analysis in Econometrics. *American Journal of Agricultural Economics* **83**: 1181-1186.

Ghosh, J. K. and R. Mukerjee (1992). Non-informative Priors. *Bayesian Statistics*. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. Oxford, Clarendon Press. **4**: 195-210.

Gilchrist, W. (1984). *Statistical Modelling*. New York, John Wiley & Sons.

Gilks, W. R., S. Richardson, et al. (1996). *Markov Chain Monte Carlo in Practice*. London; New York, Chapman & Hall.

Girr, N. C. (1977). *Multivariate Statistical Inference*. New York, Academic.

Glaeser, E. L. and J. A. Schinkman (2001). Non-market Interaction. Cambridge, Massachusetts, Harvard University.

Glymour, C. (1997a). A Review of Recent Work on the Foundations of Causal Inference. *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the*

Social Sciences. V. McKim and S. Turner. Notre Dame, University of Notre Dame Press: 210-248.

Glymour, C. (1997b). Representations and Misrepresentations, Reply to Humphreys and Woodward. *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences.* V. McKim and S. Turner: 317-322.

Glymour, C. (1999). Rabbit Hunting. *Synthese* **121**: 55-78.

Glymour, C.; D. Madigan, et al. (1996). Statistical Inference and Data Mining. *Communications of ACM* **39**: 35-41.

Glymour, C. and P. Spirtes (1993). Comment: Conditional Independence and Causal Inference. *Statistical Science* **8**: 250-257.

Glymour, C. and P. Spirtes (1994). Selecting Variables and Getting to the Truth. *Grue! The New Riddle of Induction.* D. Stalker. Open Court.

Glymour, C., P. Spirtes, et al. (1999). Response to Rejoinder. *Computation, Causation, and Discovery.* C. Glymour and G. Cooper. Cambridge, MA, AAAI Press: 343-345.

Glymour, C., P. Spirtes, et al. (1991). Causal Inference. *Erkenntnis* **35**: 151-189.

Glymour, C., P. Spirtes, et al. (1994). In Place of Regression. *Patrick Suppes: Scientific Philosopher.* P. Humphreys. Dordrecht, Holland, Kluwer Academic Publishers.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations.* New York, John Wiley.

Goldberger, A. S. (1971). Discerning a Causal Pattern Among Data on Voting Behaviour. *Causal Models in the Social Sciences.* H. M. Blalock: 33-49.

Goldberger, A. S. (1989). Economic and Mechanical Models of Intergenerational Transmission. *American Economic Review* **79**: 504-13.

Goldberger, A. S. (1992). Models of Substance; Comments on N. Wermuth, 'On Block-Recursive Linear Regression Equations. *Brazilian Journal of Probability and Statistics* **6**: 1-56.

Good, I. J. (1968). Corroboration, Explanation, Evolving Probability, Simplicity, and a Sharpened Razor. *British Journal for the Philosophy of Science* **19**: 123-143.

Good, I. J. (1983). The Philosophy of Exploratory Data Analysis. *Philosophy of Science* **50**: 283-295.

Good, I. J. (1988). Surprise Index. *Encyclopaedia of Statistical Sciences.* S. Kotz, N. L. Johnson and C. B. Reid. **7**: 104-19.

Goodfriend, M. (1992). Information-Aggregation Bias. *American Economic Review* **82**: 508-19.

Goodman, N. (1955). *Fact, Fiction, and Forecast.* Cambridge, Harvard University Press.

Gorman, W. M. (1953). Community Preference Fields. *Econometrica* **21**: 63-80.

Gorman, W. M. (1961). On a Class of Preference Fields. *Macroeconomica* **13**: 53-56.

Granger, C. (1980). Testing Causality: A Personal Viewpoint. *Journal of Economic Dynamics and Control* **2**: 329-352.

Granger, C. (1988). Causality Testing in a Decision Science. *Causation in Decision, Belief Change, and Statistics I.* W. Harper, and B. Skyrms, Kluwer Academic Press. **I:** 1-20.

Granger, C., M. King, L., et al. (1995). Comments on Testing Economic Theories and the Use of Model Selection Criteria. *Journal of Econometrics* **67**: 173-187.
Granger, C.-W. (1988). Some Comments on Econometrics Methodology. *Economic Record* **64**: 327-30.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Methods and Cross-Spectral Methods. *Econometrica* **34**: 424-438.

Granger, C. W. J. (1990). Where Are the Controversies in Econometric Methodology? *Modelling Economic Series: Readings in Econometrics.* C. W. J. Granger. Oxford, Oxford University Press: 1-28.

Granger, C. W. J., Ed. (1990a). *Modelling Economic Series: Reading in Economic Methodology.* Oxford, Oxford University Press.

Granger, C. W. J. (1999). *Empirical Modelling in Economics: Specification and Evaluation.* Cambridge, Cambridge University Press.

Granger, C. W. J. and M. Morris (1976). Time Series Modeling and Interpretation. *Journal of the Royal Statistical Society A* **38**: 246-257.

Granger, W. (1980). Long Memory Relationships and the Aggregation of Dynamic Models. *Journal of Econometrics* **14**: 227-38.

Granger, W. (1987). Implications of Aggregation with Common Factors. *Econometric Theory* **3**: 208-22.

Granger, W. (1990). Aggregation of Time Series Variables. *Disaggregation in Econometric modelling.* I. T. Barker and M. H. Pesaran. London, Rutledge: 17-34.

Green, H. A. J. (1964). *Aggregation in Economic Analysis, an Introductory Survey.* Princeton, N.J., Princeton University Press.

Green, H. A. J. (1977). Aggregation Problems of Macroeconomics. *The Microeconomic Foundations of Macroeconomics.* C. G. Harcourt. Boulder, Colorado, Westview Press: 179-94.

Green, P. and B. Silverman (1994). *Nonparametric Regression and Generalised Linear Models: A Roughness Penalty Approach, Monographs on Statistics and Applied Probability.* New York, Chapman and Hall.

Greene, W. H. (1990). *Econometric analysis.* New York, Macmillan.

Greenwald, B. and J. Stiglitz (1993). New and Old Keynesians. *Journal of Economic Perspectives* **7**: 23-44.

Grether, D. M. and C. R. Plott (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. *American Economic Review* **69**: 623-38.

Griffits, T. L., E. R. Baraff, et al. (2004). Using Physical Theories to infer Hidden Causal Structure. *To appear in Proceedings of the 26th Annual Conference of the Cognitive Science Society*, http://cog.brown.edu/~gruffydd/papers/hidden.pdf. **2004**.

Griliches, Z. and M. D. Intriligator (1984). *Handbook of Econometrics*. Amsterdam ; Oxford, North Holland.

Grossman, S. and R. J. Shiller (1982). Consumption Correlatedness and Risk Measurement in Economics with Non-Traded Assets and Heterogeneous Information. *Journal of Financial Economics* **10**: 195-210.

Grunfeld, J. and Z. Griliches (1960). Is Aggregation Necessarily Bad. *Review of Economic Statistics* **42**: 1-13.

Guttman, I. (1967). The Use of the Concept of a Future Observation in Goodness-of-Fit Problems. *Journal of the Royal Statistical Society, Series B* **29**: 83-100.

Haavelmo, T. (1943). The Statistical Implications of a System of Simultaneous Equations. *Econometrica* **11**: 1-12.

Haavelmo, T. (1944). The Probability Approach in Econometrics. *Econometrica* **12 (Supplement)**.

Hacking, I. (1967). Slightly More Realistic Personal Probability. *Philosophy of Science* **34**: 311-25.

Hahn, F. H. (1973). *On the Notion of Equilibrium in Economics*. Cambridge, Cambridge University Press.

Hall, P. (1983). Large Sample Optimality of Least Squares Cross-Validation in Density Estimation. *The Annals of Statistics* **11**: 1156-1174.

Hall, P. (1989). On Convergence Rates of Nonparametric Problems. *International Statistical Review* **57**: 45-58.

Hall, R. E. (1978). Stochastic Implications of the Life Cycle-Permanent Income Hypothesis: Theory and Evidence. *Journal of Political Economy* **86**: 971-87.

Hall, R. E. (1989). Consumption. *Modern Business Cycle Theory*. R. Barro. Oxford, Basil Blackwell and Harvard University Press.

Hansen, L. P. (1998). New Approaches to Macroeconomic Modeling: Evolutionary Stochastic Dynamics, Multiple Equilibria, and Externalities as Field Effects (book review). *Journal of Economic Literature* **36**: 239-241.

Hanushek, E. A. and J. E. Jackson (1977). *Statistical Methods for Social Scientists*. New York ; London, Academic Press.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge, Cambridge University Press.

Härdle, W. (1993). Applied Nonparametric Methods. *Handbook of Econometrics*. R. Engle, McFadden, D., North-Holland: Chapter 38.

Harper, W. L. and B. Skyrms (1988). *Causation in Decision, Belief Change, and Statistics : Proceedings of the Irvine Conference on Probability and Causation*. Dordrecht, Kluwer Academic Publishers.

Harper, W. L. and B. Skyrms (1988). *Causation, Chance, and Credence : Proceedings of the Irvine Conference on Probability and Causation.* Dordrecht, Kluwer Academic Publishers.

Harsanyi, J. C. (1965). Bargaining and Conflict Situations in the Light of a New Approach to Game Theory. *The American Economic Review* **55**: 447-457.

Harsanyi, J. C. (1966). A General Theory of Rational Behaviour in Game Situations. *Econometrica* **34**(613-634).

Harsanyi, J. C. (1977). On the Rationale of the Bayesian Approach: Comments on Professor Watkins. *Foundational Problems in the Special Sciences.* Butts and Hîntikka, D. Reidel Publishing Company. **381-392.**

Hartigan, J. A. (1964). Invariant Prior Distributions. *Annals of Mathematical Statistics* **35**: 836-845.

Hartley, J. (1997). *The Representative Agent in Macroeconomics.* London, Rutledge.

Hastie, T. J. and R. J. Tibshirani (1990). *Generalized Additive Models*, Chapman & Hall.

Hastie, T. J. and R. J. Tibshirani (1994). Nonparametric Regression and Classification. *From Statistics to Neural Networks, Theory and Pattern Recognition Applications.* V. Cherkassky, Springer-Verlag 1.

Hausman, D. (1984). Causal Priority. *Nous* **18**: 261-279.

Hausman, D. (1991). On Dogmatism in Economics: The Case of Preference Reversals. *Journal of Socio-Economics* **20**: 205-225.

Hausman, D. M. (1992). *Essays on Philosophy and Economic Methodology.* Cambridge; New York, Cambridge University Press.

Hausman, D. M. (1992). *The Inexact and Separate Science of Economics.* Cambridge, Cambridge University Press.

Hausman, D. M. (1998). *Causal Asymmetries.* Cambridge, U.K.; New York, Cambridge University Press.

Hausman, D. M. and J. Woodward (1999). Independence, Invariance, and the Causal Markov Condition. *British Journal for the Philosophy of Science* **50**: 521-583.

Hausman, J., B. H. Hall, et al. (1984). Econometric Models For Count Data with an Application to the Patents-R&D Relationship. *Econometrica* **52**: 909-938.

Hayek, F. A. (1979). *The Counter-Revolution in Science: Studies in the Abuse of Reason.* Indianapolis, Liberty Press.

Hedström, P. and R. Swedberg (1998). *Social Mechanisms: an Analytical Approach to Social Theory.* Cambridge, Cambridge University Press.

Heineke, J. M. and H. Scheffrin (1988). Exact Aggregation and the Finite Basis Property. *International Economic Review* **29**,: 525-538.

Hempel, C. (1965). *Aspects of Scientific Explanation and other Essays in the Philosophy of Science.* New York, Free Press.

Hendry, D. and M. Morgan (1995). *The Foundations of Econometric Analysis.* Cambridge; New York and Melbourne, Cambridge University Press.

Hendry, D. F. (1987). Econometric Methodology: a Personal Perspective. *Advances in Econometrics Fifth World Congress.* T. Bewley. Cambridge, Cambridge University Press. **II:** 29-48.

Hendry, D. F. (1993). *Econometrics: Alchemy or Science? Essays in Econometric Methodology.* Cambridge and Oxford, Blackwell.

Hendry, D. F., E. E. Leamer, et al. (1990). The ET Dialogue: A Conversation on Econometric Methodology. *Econometric Theory* **6**: 171-261.

Hershberger, S. L. (1994). The Specification of Equivalent Models before Collection of Data. *Latent Variables Analysis.* A. Von Eye and C. C. Clogg. Thousand Oaks, CA, Sage: 68-108.

Hesslow, G. (1976). Discussion: Two Notes on the Probabilistic Approach to Causality. *Philosophy of Science* **43**: 290-92.

Hicks, J. R. (1939). *Value and Capital: an Inquiry into Some Fundamental Principles of Economic Theory.* Oxford, Oxford University Press.

Hicks, J. R. (1956). *A Revision of Demand Theory.* Oxford, Oxford University Press.

Hicks, J. R. (1979). *Causality in Economics.* Oxford, Basil Blackwell.

Hildenbrand, W. (1985). A Problem in Demand Aggregation: Per Capita Demand as a Function of Per Capita Expenditure. *EUI Working papers.* Florence.

Hill, B., M., (1986). Some Subjective Considerations in the Selection of Models (with discussion). *Econometric Review* **1**: 191-288.

Hill, B., M., (1990). A Theory of Bayesian Data Analysis. *Bayesian and Likelihood Methods in Statistics and Econometrics.* S. Geisser, J. S. Hodges, S. J. Press and A. Zellner. North-Holland, Elsevier Science Publishers B. V.: 40-73.

Hoaglin, D. C. (1980). A Poissoness Plot. *The American Statistician* **34**: 146-149.

Hoaglin, D. C. and J. W. Tukey (1985). Checking the Shape of Discrete Distributions. *Exploring Data Tables, Trends, and Shapes.* D. C. Hoaglin, F. Moseteller and J. W. Tukey. New York, Wiley.

Hodges, J. S. (1987). Uncertainty, Policy Analysis and Statistics (with discussion). *Statistical Science* **2**: 259-291.

Hodges, J. S. (1990). Can / May Bayesians Do Pure Tests of Significance? *Bayesian and Likelihood Methods in Statistics and Econometrics.* S. Geisser, J. S. Hodges and A. Zellner. North-Holland, Elsevier Science Publishers: 75-90.

Honkapohja, S. (1995). Bounded Rationality in Macroeconomics: A Review Essay. *Journal of Monetary Economics* **35**: 509-518.

Hoover, K. (2003). Nonstationary Time Series, Cointegration, and the Principle of the Common Cause. *British Journal for the Philosophy of Science* **54**: 527-551.

Hoover, K. D. (2001). *Causality in Macroeconomics.* Cambridge, New York, Cambridge University Press.

Hoover, K. D. and S. J. Perez (1999). Data Minining Reconsidered: Encompassing and the General-to-Specific Approach to Specification Search. *Econometrics Journal* **2**: 167-191.

Howitt, P. (1986). Conversations with Economists: A Review Essay. *Journal of Monetary Economics* **18**: 103-118.

Howitt, P. (1987). Macroeconomics: Relations with Microeconomics. *The New Palgrave: A Dictionary of Economics.*

Howson, C. (1993). Dutch Books and Consistency. *PSA.* M. F. D. Hull, and K. Okruhlik: 161-8.

Howson, C. (1995). Theories of Probability. *British Journal for the Philosophy of Science* **46**: 1-32.

Howson, C. (1997). Bayesian Rules of Updating. *Erkenntnis* **45**: 195-208.

Howson, C. (2000). *Induction : Hume's problem.* Oxford, Clarendon.

Howson, C. (2004). Chapter3: The Laws of Probability (part of an unpublished manuscript). London, London School of Economics and Political Sciences: 31 pages.

Howson, C. and P. Urbach (1993). *Scientific Reasoning: The Bayesian Approach.* Chicago, Open Court.

Huber, P. (1985). Projection Pursuit (with discussion). *Ann. Statist* **13**: 135-175.

Humphreys, P. (1990). *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences.* Princeton, Princeton University Press.

Humphreys, P. N. (1997). A Critical Appraisal of Causal Discovery Algorithms. *Causality in Crisis?* V. R. McKim and S. T. Turner, University of Notre Dame Press: 249-63.

Humphreys, P. N., Freedman, D. (1996). The Grand Leap. *British Journal for the Philosophy of Science* **47**: 113-123.

Hurwicz, L. (1962). On the Structural Form of Interdependent Systems. *Logic, Methodology, and the Philosophy of Science.* E. Nagel, P. Suppes and A. Tarski. California, Stanford University Press: 232-239.

Hylleberg, S. and M. Paldam, Eds. (1991). *New Approaches to Empirical Macroeconomics.* Oxford, Blackwell Publishers.

Ingrao, B. and G. Israel (1990). *The Invisible Hand: Economic Equilibrium in the History of Science.* Massachusetts, MIT Press.

Irzik, G. (1996). Can Causes Be Reduced to Correlations? *British Journal for the Philosophy of Science* **47**: 249-270.

Irzik, G. and E. Meyer (1987). Causal Modelling: New Directions for Statistical Explanation. *Philosophy of Science* **54**: 495-514.

Jacobs, D. P., E. Kalai, et al., Eds. (1998). *Frontiers of Research in Economic Theory.* Econometric Society Monographs. Cambridge, Cambridge University Press.

Janssen, M. C. W. (1993). *Microfoundations: A Critical Inquiry.* London, Rutledge.

Jaynes, E. (1983). *Papers on Probability, Statistics, and Statistical Physics,.* Dordrecht, Reidel Publishing Co.

Jaynes, E. T. (1968). Prior Probabilities. *Papers on Probability, Statistics and Statistical Physics,.* R. Rosenkranntz. Dordrecht, Reidel: 116-130.

Jeffrey, R. (1968). Probable Knowledge. *The Problem of Inductive Inference.* I. Lakatos. Amsterdam, North-Holland: 166-180.

Jeffrey, R. (1988). Conditioning, Kinematics, and Exchangeability'. *Causation, Chance, and Credence.* B. Skyrms and H. W. Dordrecht, Kluwer: 221-225.

Jeffreys, H. (1946). An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society, A* **186**: 453-461.

Jeffreys, H. (1955). The Present Position of Probability Theory. *The British Journal for the Philosophy of Science* **5**: 275-89.

Jeffreys, H. (1961). *Theory of Probability.* Oxford, The Clarendon Press (First published in 1939).

Jeffreys, H. (1973). *Scientific Inference.* London, Cambridge University Press.

Jerison (1994). Optimal Income Distribution Rules and Representative Consumers. *Review of Economic Studies* **61**: 731-77.

Jevons, W. S. (1965 [1871]). *The Theory of Political Economy.* New York, A. M. Kelley.

Johnson, N. L., S. Kotz, et al. (1997). *Discrete Multivariate Distributions.* New York; Chichester, John Wiley.

Johnson, N. L., S. Kotz, et al. (1994). *Continuous Univariate Distributions.* New York, John Wiley.

Johnston, J. (1984). *Econometric Methods.* New York, McGraw - Hill.

Jorgenson, D. W., L. J. Lau, et al. (1982). The Transcendental Logarithmic Model of Aggregate Consumer Behaviour. *Advances in Econometrics.* R. L. Basmann and G. Rhodes. Greenwich Conn, JAI Press: 97-238.

Jorgeskog, K. and D. Sorbom (1990). Model Search with Tetrad and LISREL. *Sociological Methods and Research* **19**: 93-106.

Judge, G. G., W. E. Griffiths, et al. (1985). *The Theory and Practice of Econometrics.* New York, John Wiley.

Kadane, B. J., J. M. Dickey, et al. (1980). Interactive Elicitation of Opinion for a Normal Linear Model. *Journal of the American Statistical Association* **75**: 845-854.

Kadane, B. J. and P. D. Larkey (1982). Subjective Probability and the Theory of Games. *Management Science* **28**: 113-120.

Kadane, J. B. (1980). Predictive and Structural Methods for Eliciting Prior Distributions. *Bayesian Analysis in Econometrics and Statistics.* A. Zellner, North-Holland Publishing Company: 89-93.

Kadane, J. B. and L. J. Wolfson (1998). Experiences in Elicitation. *Journal of the Royal Statistical Society Series D* **47**: 3-19.

Kahneman, D. (1996). New Challenges to the Rationality Assumption. *The Rational Foundations of Economic Behaviour,.* K. Arrow, Mark Perlman, Christian Schmidt. London, Macmillan Press Ltd: 203-219.

Kahneman, D. (2003). Maps of Bounded Rationality: Psychology for Behavioral Economics. *American Economic Review* **93**: 1449-75.

Kahneman, D. and A. Tversky (1973). On the Psychology of Prediction. *Psychological Review* **80**: 237-251.

Kalai, E. and E. Lehrer (1993). Rational Learning Leads to Nash Equilibrium. *Econometrica* **61**: 1019-45.

Kass, R. E. (1993). Bayes Factors in Practice. *Statistician* **42**: 551-560.

Kass, R. E. and L. Wasserman (1996). Comments. *Statistica Sinica* **6**: 774-779.

Kass, R. E. and L. Wasserman (1996). The Selection of Prior Distribution by Formal Rules. *Journal of the American Statistical Association* **91**: 1343-70.

Kenny, D. A. (1979). *Correlation and Causality.* New York, John Wiley.

Keynes, J. M. (1936). *The General Theory of Employment, Interest and Money.* London, Macmillan.

Kiefer, M. N. and Y. Nyarko (1995). Savage Bayesian Models of Economics. *Essays in Learning and Rationality in Economics and Games.* A. Kirman and M. Salmon, Basil Blackwell Press: 42-62.

Kiefer, N. M. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature* **XXVI**: 646-679.

Kiiveri, H. T. and T. P. Speed (1982). Structural Analysis of Multivariate Data: A Review. *Sociological Methodology.* S. Leinhardt. San Francisco, Jossey Bass: 209-289.

Kiiveri, H. T., T. P. Speed, et al. (1984). Recursive Causal Models. *Journal of the Australian Mathematical Society* **36**: 30-52.

Kirman, A. (1989). The Intrinsic Limits of Modern Economic Theory: The Emperor Has No Clothes. *Economic Journal* **99**: Conference:126-39.

Kirman, A. (1992). Whom or What Does the Representative Individual Represent? *Journal of Economic Perspectives* **6**: 117-136.

Kirman, A. (1997). The Economy as an Evolving Network. *Journal of Evolutionary Economics* **7**: 339-53.

Kirman, A. (1997). Microfoundations - Built on Sand? A Review of Maarten Janssen's Microfoundations. *Economics and Philosophy*: 322-33.

Kirman, A. (1997). Some Observations on Interaction in Economics, http://www.cpm.mmu.ac.uk/pub/workshop/kirman.html.

Kirman, A. P. and M. Salmon (1995). *Learning and Rationality in Economics.* Oxford, UK; Cambridge, Mass., Blackwell.

Klein, L. R. (1946a). Macroeconomics and the Theory of Rational Behaviour. *Econometrica* **14**: 93-108.

Klein, L. R. (1946b). Remarks on the Theory of Aggregation. *Econometrica* **14**: 303-312.

Kmenta, J. (1986). *Elements of Econometrics.* New York, Macmillan Publishing Company.

Knight, F. H. (1964 [1921]). *Risk, Uncertainty, Profit.* New York, Augustus M. Kelley.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Assessment and Model Selection. *IJCAI*: 1137-1145.

Koopmans, I. C. (1947). Identification Problems in Economic Model Construction. *Econometrica* **17**: 125-144.

Koopmans, I. C. (1947b). Measurement without Theory. *The Review of Economic Statistics* **29**(3): 161-172.

Koopmans, T. (1979). Economics among the Sciences. *American Economic Review* **69**: 1-13.

Koopmans, T. and O. Reiersol (1950). The Identification of Structural Characteristics. *Annals of Mathematical Statistics* **21**: 151-181.

Koopmans, T. C. (1949). The Econometric Approach to Business Fluctuations. *American Economic Review* **39**: 64-73.

Koster, J. (1999). On the Validity of the Markov Interpretation of Path Diagrams of Gaussian Structural Equation Systems of Simultaneous Equations. *Scandinavian Journal of Statistics* **26**: 413-431.

Kreps, D. (1988). *Notes on the Theory of Choice.* Boulder and London, Westview Press.

Kreps, D. (1990). *Game Theory and Economic Modelling.* Oxford, Oxford University Press.

Kupiec, P. H. and S. A. Sharpe (1991). Animal Spirits, Margin Requirements, and Stock Price Volatility. *Journal of Finance* **46**: 717-31.

Kyburg, H. E., Jr. and H.E. Smokler, Ed. (1980). *Studies in Subjective Probability.* New York, Krieger.

Laidler, D. E. W. (1982). *Monetarist Perspectives.* Oxford, Philip Allan.

Lam, D. (1988). Marriage Markets and Assortive Mating with Household Public Goods: Theoretical Results and Empirical Implications. *Journal of Human Resources* **23**: 462-487.

Lane, D. (1986). Comments. *Econometric Review* **4**: 253-258.

Lane, D. (1993). Artificial Worlds and Economics, Part I. *Journal of Evolutionary Economics* **3**: 89-107.

Lane, D. (1993). Artificial Worlds and Economics, Part II. *Journal of Evolutionary Economics* **3**: 177-97.

Lane, D., F. Marlerba, et al. (1996). Choice and Action. *Journal of Evolutionary Economics* **6**: 43-76.

Lau, L. J. (1977). Existence Conditions for Aggregate Demand Functions: The Case of Multiple Indexes, Technical Report No.249, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Lau, L. J. (1982). A Note on the Fundamental Theorem of Exact Aggregation. *Economic Letters* **9**: 119-26.

Lau, L. J. (1986). Functional Forms in Econometric Model Building. *Handbook of Econometrics*. Z. Griliches and M. D. Intriligator. Amsterdam:, North Holland. **III:** ch.25.

Leamer, E. E. (1978). *Specification Searches: Ad hoc Inference with Nonexperimental Data*. New York, Wiley.

Leamer, E. E. (1983). Let's Take the Con out of Econometrics. *American Economic Review* **79**: 31-43.

Leamer, E. E. (1983). Model Choice and Specification Analysis. *Handbook of Econometrics*. Z. Griliches and M. D. Intriligator. Amsterdam, North Holland. **I:** ch.5.

Leamer, E. E. (1985). Vector Autoregressions for Causal Inference. *Understanding Monetary Regimes*. K. Brunner and A. H. Meltzer. Amsterdam, North- Holland**:** 255-304.

Leamer, E. E. (1991). A Bayesian Perspective on Inference from Macroeconomic Data. *The Scandinavian Journal of Economics* **93**: 225-248.

Lee, P. M. (1997). *Bayesian Statistics: An Introduction*. London, Arnold.

Lee, S. and S. L. Hershberger (1990). A Simple Rule for Generating Equivalent Models in Covariance Structure Modeling. *Multivariate Behavioral Research* **25**: 313-334.

Lehmann, E. L. (1990). Model Specification: The Views of Fisher and Neyman, and Later Developments. *Statistical Science* **5**: 160-168.

Lehmann, E. L. and H. J. M. D'Abrera (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco, Holden-Day.

Leijohnufvud, A. (1968). *On Keynesian Economics and the Economics of Keynes*. New York, Oxford University Press.

Lewbel, A. (1989). Exact Aggregation and a Representative Consumer. *Econometrica* **57**: 701-06.

Lewbel, A. (1994). Aggregation and Simple Dynamics. *American Economic Review* **84**: 905-18.

Lewis, D. (1986). Causal Explanation. *Philosophical Papers*. New York, Oxford, Oxford University Press. **II:** 214-241.

Lichtenstein, S. and P. Slovic (1971). Reversals of Preferences between Bids and Choices in Gambling Decisions. *Journal of Experimental Psychology* **89**: 46-55.

Lindley, D. V. (1968). The Choice of Variables in Multiple Regression (with discussion). *Journal of the royal Statistical Society Series B* **30**: 31-66.

Lindley, D. V. (1982). The Bayesian Approach to Statistics. *Some Recent Advances in Statistics*. J. Taago de Oliveria, and Benjamin Epstein. New York, Academic Press: 65-87.

Lindley, D. V. (1983). Theory and Practice of Bayesian Statistics. *The Statistician* **32**: 1-11.

Lindley, D. V. (1990). The 1988 Wald Memorial Lectures: The Present Position in Bayesian Statistics. *Statistical Science* **5**: 44-89.

Lindley, D. V. and L. D. Phillips (1976). Inference for a Bernoulli Process (a Bayesian View). *American Statistician* **30**: 112-119.

Lindman, H. R. (1971). Inconsistent Preferences among Gambles. *Journal of Experimental Psychology* **89**: 390-97.

Linhart, H. and W. Zucchini (1986). *Model Selection*. New York, John Wiley & Sons.

Lippi, M. (1988). On the Dynamics of Aggregate Macroequations: from Simple Microbehaviours to Complex Macrorelations. *Technical Change and Economic Theory*. G. Dosi, C. R. Freeman, G. Silvergerg and L. Soete. London, Printer: 170-196.

Lippi, M. (1992). Microfoundations of Dynamic Macroequations. *Themes in Modern Macroeconomics*. H. Brink. Hampshire and London, The Macmillan Press LTD: 35-49.

Lippi, M. and M. Forni (1990). On the dynamic specification of Aggregate Models. I. T. Barker and M. H. Pesaran.

Litterman, R., B. (1986). A statistical Approach to Economic Forecasting. *Journal of Business and Economic Statistic* **4**(1).

Liu, T. C. (1960). Underidentification, Structural Estimation, and Forecasting. *Econometrica* **28**: 855-865.

Lucas, R. (1973). Some International Evidence on Output-Inflation Tradeoffs. *American Economic Review* **63**: 326-334.

Lucas, R. (1978). Asset Prices in an Exchange Economy. *Econometrica* **46**: 1429-45.

Lucas, R. and T. Sargent (1979 (1981)). After Keynesian Macroeconomics. *Rational Expectations and Econometric Practice*. R. Lucas and T. Sargent. Minneapolis, Minn, University of Minnesota Press: 295-319.

Lucas, R. and T. Sargent, Eds. (1981). *Rational Expectations and Econometric Practice*. Minneapolis, Minn, University of Minnesota Press.

Lucas, R. E. (1976). Econometric Policy Evaluation: A Critique. *The Phillips Curve and Labour Market*. K. Brunner and A. H. Meltzer. Amsterdam, North-Holland. **1**: 19-46.

Lucas, R. E. (1980 (1981)). Methods and Problems in Business Cycle theory. *Studies in Business-Cycle Theory*. R. E. Lucas. Cambridge, Massachusetts, The MIT Press: 271-296.

Lucas, R. E. (1981). *Studies in Business-cycle Theory*. Oxford, Basil Blackwell.

Lucas, R. E. (1987). *Models of business cycles*. Oxford, Basil Blackwell.

Lucas, R. E., Jr. (1977 (1981)). Understanding Business Cycles. *Studies in Business-Cycle Theory*. R. E. Lucas. Cambridge, Massachusetts, The MIT Press: 215-239.

Lucas, R. E., Jr., (1986). Adaptive Behaviour and Economic Theory. *Journal of Business* **59**: 5401-26.

Luce, R. D. and P. Suppes (1965). Preference, Utility, and Subjective Probability. *Handbook of Mathematical Psychology III*. R. Luce, R. Bush and E. Galanter, Wiley.

Luijben, T. C. W. (1991). .Equivalent Models in Covariance Structure Analysis. *Psychometrika* **56**: 653-666.

MacCallum, B. T. (1983). On Non-Uniqueness in Rational Expectations Models. *Journal of Monetary Economics* **11**: 139-168.

MacCallum, R., D. Wegener, et al. (1993). The Problem of Equivalent Models in Applications of Covariance Structure Analysis. *Psychological Bulletin* **114**: 185-199.

MacCrimmon, K. (1968). Descriptive and Normative Implications of Decision Theory Postulates. *Risk and Uncertainty*. K. Borch and J. Mossin. New York, Macmillan: 3-23.

Machina, M. J. (1987). Choice under Uncertainty: Problems Solved and Unsolved. *Economic Perspectives* **1**: 121-154.

MacNeill and G. J. Umphrey, Eds. (1987). *Foundations of Statistical Inference*. Boston, Reidel.

Magat, W. A., W. K. Viscusi, et al. (1988). Paired Comparison and Contingent Valuation Approaches to Morbidity Risk Evaluation. *Journal of Experimental Economics and Management* **15**: 395-411.

Maher, P. (1993). *Betting on Theories*. Cambridge, Cambridge University Press.

Malinvaud, E. (1993). A Framework for Aggregation Theories. *Ricerche-Economiche* **47**(2): 107-35.

Mallow, C. L. (1970). Some Comments on Bayesian Methods. *Bayesian Statistics*. D. L. Meyer and R. O. J. Collier. Itasca, IL, Peacock: 71-84.

Mankiw, G., N. (1993). New Keynesian Economics. *Entry in the On-line Concise Encyclopedia of Economics*.

Manski, C. F. (1991). Regression. *Journal of Economic Literature* **29**: 34-50.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge, Mass, Harvard University Press.

Mantel (1976). Homothetic Preferences and Community Excess Demand Functions. *Journal of Economic Theory* **12**: 197-201.

Marcoulides, A. G. and R. E. Schumacker, Eds. (1996). *Advanced Structural Equation Modeling*. Mahwah, New Jersy, Lawrence Erlbaum Associates, Publishers.

Mardia, K. V. (1970). *Families of Bivariate Distributions*. London,.

Marrimon, R. (1997). Learning from Learning in Economics. *Advances in Economics and Econometrics: Theory and Applications*. D. Kreps, M. and K. F. Wallis. Cambridge, Cambridge University Press: 278-315.

Marron, J. (1996). A Personal View of Smoothing and Statistics. *Statistical Theory and Computational Aspects of Smoothing*. W. Hürdle and M. Schimek. Heidelberg, Germany, Physika Verlag: 1-9.

Marron, J. a. P. H. (1991). Local Minima in Cross-Validation Functions. *Journal of Royal Statistical Society, Series B* **53**: 245-252.

Marschak, J. (1953). Econometric Measurements for Policy and Prediction. *Economic Information, Decision, and Prediction*. Marschak. **1(1974)**.

Marshall, A. (1890 [1961]). *Principles of Economics, (9th ed.)*. New York, Macmillan.

Matsuyama, K. (1987). Current Account Dynamics in a Finite Horizon Model,. *Journal of International Economics* **23**: 299-313.

May, K. (1947). Technological Change and Aggregation. *Econometrica* **15**: 51-63.

McAleer, M. (1987). Specification Tests for Separate Models: a Survey. *Specification Analysis in the Linear Model*. M. King and D. Giles. London, Rutledge and Kagan Paul: ch.9.

McCann, C. R. (1994). *Probability Foundations of Economic Theory*. London; New York, Rutledge.

McCullagh, P. (1995). Discussion of Papers by Reid and Zeger and Liang. *Statistical Science* **10**: 177-179.

McFadden, D. (1999). Rationality for Economics. *Journal of Risk and Uncertainty* **19**: 73-105.

McKim, V. and S. Turner, Eds. (1997). *Causality in Crisis?: Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Notre Dame, Ind., University of Notre Dame Press.

Meece, J. L., P. C. Blumenfeld, et al. (1988). Students' Goal Orientations and Cognitive Engagement in Classroom Activities. *Journal of Educational Psychology* **80**: 514-523.

Meek, C. (1995). Causal Inference and Causal Explanation with Background Knowledge. *Uncertainty in Artificial Intelligence*. P. Besnard, S. Hanks. San Francisco, Morgan Kaufmann. **II**: 403-410.

Meek, C. and C. Glymour (1994). Conditioning and Intervening. *British Journal for the Philosophy of Science* **45**: 1001-1021.

Meng, X. L. (1994). Posterior Predictive *p*-values. *The Annals of Statistics* **22**: 1142-1160.

Menzies, P. (1989). Probabilistic Causation and Causal Processes: A Critique of Lewis. *Philosophy of Science* **59**: 642-663.

Michener, R. (1984). Permanent Income in General Equilibrium. *Journal of Monetary Economics* **13**: 297-305.

Milgrom, P. and N. Stokey (1982). Inflation, Trade, and Common Knowledge. *Journal of Economic Theory* **26**: 17-27.

Mill, J. S. (1974(1987)). *A System of Logic*. Toronto, University of Toronto Press.

Mill, J. S. (1990). *Principles of Political Economy*. New York, The Colonial Press (First published 1948).

Modigliani, F. (1977). The Monetarist Controversy, or Should We Forsake Stabilization Policies? *American Economic Review* **67**: 1-19.

Moody, J. (1994). Prediction Risk and Architecture Selection for Neural Networks. *From Statistics to Neural Networks: Theory and Pattern Recognition Applications*. V. Cherkassky, J. H. Friedman and H. Wechsler, Springer.

Morgenstern, O. (1963). *On the Accuracy of Economic Observations*, Princeton University.

Morris, S. (1995). The Common Prior Assumption in Economic Theory. *Economic and Philosophy* **11**: 227-253.

Mosteller, F. and P. Nogee (1951). An Experimental Measurement of Utility. *Journal of Political Economy* **59**: 371-404.

Moulin, H. (1986). *Game Theory for the Social Sciences*. New York, New York University Press.

Muellbauer, J. (1975). Aggregation, Income Distribution and Consumer Demand. *Review of Economic Studies* **42**: 525-43.

Muellbauer, J. (1976). Community Preferences and the Representative Consumer. *Econometrica* **44**: 979-99.

Muth, J. F. (1961). Rational Expectations and the Theory of Price Movements. *Econometrica* **29**: 315-335.

Nataf, A. (1948). Sur la Possibilité de Construction de Certains Macromodèles. *Econometrica* **16**: 232-44.

Nelson, A. (1984). Some Issues Surrounding the Reduction from Macroeconomics to Microeconomics. *Philosophy of Science* **51**: 573-94.

Nerlove, M. (1972). On Lags in Economic Behavior. *Econometrica* **40**: 221-51.

Nyarko (1997). Savage-Bayesians Play a Repeated Game. *The Dynamics of Norms*. R. J. C. Bicchieri, and B. Skyrms. Cambridge University Press.

Nyarko, Y. (1991). Learning in Mis-Specified Models and the Possibility of Cycles. *Journal of Economic Theory* **55**: 416-427.

Nyarko, Y., N. Yannelis, et al. (1994). Bounded Rationality and Learning. *Economic Theory* **4**: 811-820.

Oakes, M. (1980). *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. New York, John Wiley and Sons.

O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics. Vol. 2B, Bayesian Inference*. London, Edward Arnold.

O'Hagan, A. (1998). Eliciting Expert Beliefs in Substantial Practical Applications. *The Statistician* **47**: 21-35.

Ord, J. K. (1967). Graphical Methods for a Class of Discrete Distributions. *Journal of the Royal Statistical Society Series A* **130**: 232-238.

Pagan, A. (1987). Three Econometric Methodologies: A Critical Appraisal. *Journal of Economic Surveys* **1**: 3-24.

Payne, J., W. and J. R. Bettman (1992). Behavioral Decision Research: A Constructive Processing Perspective. *Annual Review of Psychology* **43**: 87-131.

Pearce, D. (1984). Rationalizable Strategic Behaviour and the Problem of Perfection. *Econometrica* **52**: 1029-1050.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Nateo, Morgan and Kaufman.

Pearl, J. (1993). *Aspects of Graphical Models Connected with Causality*. Proceedings of the 49th Session of the International Statistical Institute, Italy, Tome LV, Florence.

Pearl, J. (1993a). On the Statistical Interpretation of Structural Equations, Technical Report (no. R-200), Computer Science Department, University of California, Los Angeles, On line at <http://www.cs.ucla.edu/~judea/>.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrica* **82**: 669-710.

Pearl, J. (1996). Structural and Probabilistic Causality. *The Psychology of Learning and Motivation*. D. R. Shanks, K. J. Holyoak and O. L. Medin. San Diago, California, Academic press: 393-435.

Pearl, J. (1998). Graphs, Causality, and Structural Equation Models. *Sociological Methods and Research* **27**: 226-284.

Pearl, J. (1998). Tetrad and SEM. *Multivariate Behavioral Research* **33**: 119-128.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, Cambridge University Press.

Pearl, J. and T. Verma (1991). A Theory of Inferred Causation. *Principles of Knowledge, Representation and Reasoning: Proceedings of the Second International Conference*. J. A. Allen, R. Filkes and E. Sandewall. San Mateo, CA, Morgan Kaufmann: 441-452.

Peltzman, S. (1991). The Handbook of Industrial Organization: a Review Article. *Journal of Political Economy* **99**: 201-17.

Pesaran, M. H. (1974). On the General Problem of Model Selection. *Review of Economic Studies* **4**: 153-171.

Pesaran, M. H. (1987). *The Limits to Rational Expectations.* Oxford, Basil Blackwell.

Pesaran, M. H. and R. P. Smith (1985). Evaluation of Macroeconomic Models. *Economic Modeling* **2**: 125-134.

Phillips, A. W. (1958). The Relation between Unemployment and the Rate of Change of Money Wages in the United Kingdom, 1861-1957. *Economica* **25**: 283-299.

Pischke, J. S. (1995). Individual Income, Incomplete Information, and Aggregate Consumption. *Econometrica* **63**: 805-40.

Poirier, D. J. (1988). Frequentist and Subjectivist Perspectives on the Problems of Model Building In Economics (with discussion). *Journal of Economic Perspectives* **2**: 121-144.

Pollak, R. A. (2002). Gary Becker's Contributions to Family and Household Economics. *NBER Working Paper No. w9232 Issued in September 2002.*

Pratt, J. W. (1965). Bayesian Interpretation of Standard Inference Statements (with discussion). *Journal of the Royal Statistical Scoiety, Series B* **27**: 169-203.

Pratt, J. W., H. Raiffia, et al. (1964). The Foundations of Decisions under Uncertainty: an Elementary Exposition'. *Journal of the American Statistical Association* **59**: 353-75.

Pratt, J. W. and R. Schlaifer (1988). On the Interpretation and Observation of Laws. *Journal of Econometrics* **39**: 23-52.

Press, S. J. (1989). *Bayesian Statistics: Principles, Models and Applications.* Chichester, Wiley.

Ramsey, F. P. (1926 [1980]). Truth and Probability. *Studies in Subjective Probability.* H. E. J. Kyburg and H. E. Smokler. New York, Krieger: 25-52.

Ramsey, J. B. (1983). Perspective and Comment. *Econometric Reviews* **2**: 241-8.

Raykov, T. (1997). Equivalent Structural Equation Models and Group Equality Constriants. *Multivariate Behavioral Research* **32**: 94-104.

Raykov, T. and S. Penev (1999). On Structural Equation Model Equivalence. *Multivariate Behavioral Research* **34**: 199-244.

Raykov, T. and S. Penev (2001). The Problem of Equivalent Structural Equation Models: An Individual Residual Perspective. *New Developments and Techniques in Structural Equation Modeling.* G. A. Marcoulides and R. E. Schumacker. Mahwah, NJ, Lawrence Erlbaum: 297-321.

Reaume, D. M. (1996). Walras, Complexity, and Post Walrasian Macroeconomics. *Beyond Microfoundations: Post Walrasian Macroeconomics.* D. Colander. Cambridge, New York, Cambridge University Press: 145-156.

Rice, J. A. (1984). Bandwidth Choice for Nonparametric Regression. *Annals of Statistics* **12**: 1215-30.

Richardson, T. (1996). A Polynomial-Time Algorithm for Deciding Markov Equivalence of Directed Cyclic Graphical Models. *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence, Portland, Oregon.* E. Horvitz and F. Jensen. San Francisco, CA., Morgan Kaufmann.

Richardson, T. and P. Spirtes (1999). Automated Discovery of Linear Feedback Models. *Computation, Causation, and Discovery*. G. Cooper and C. Glymour. Cambridge, MA, MIT Press: 253-304.

Richenbach, H. (1956). *The Direction of Time*. Berkeley, University of Los Angeles Press.

Richter, M. K. (1971). Rational Choice. *Rational Choice, Preferences, Utility and Demand*. J. S. Chipman, L. Hurwicz, M. K. Richter and H. F. Sonnenschein. New York, Harcourt: 29-58.

Ripley, B. D. (1993). Statistical Aspects of Neural Networks. *Networks and Chaos - Statistical and Probabilistic Aspects*. O. E. Barndorff-Nielsen, F. L. Jensen and W. S. Kendall, Chapman and Hall: 40-123.

Ripley, B. D. (1995). Statistical Ideas for Selecting Network Architecture. *Neural Networks: Artificial Intelligence and Industrial Applications*. B. Kappen and S. Gielen, Springer: 183-190.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, New York, Cambridge University Press.

Rizvi, S. A. T. (1994). The Microfoundations Project in General Equilibrium Theory. *Cambridge Journal of Economics* **18**: 357-77.

Roberts, D. L. and S. Nord (1985). Causality Tests and Functional Form Sensitivity. *Applied Economics* **17**: 135-41.

Robins, J. M. (1995). Discussion of "Causal Diagrams for Empirical Research" by J. Pearl. *Biometrika* **82**: 695-698.

Robins, J. M. (1997). Causal Inference from Complex Longitudinal Data. *Latent Variable Modeling and Applications to Causality,*. M. Berkane. New York, Springer. **120:** 69-117.

Robins, J. M. (2003). General Methodological Considerations. *Journal of Econometrics* **112**: 89-106.

Robins, J. M. and L. Wasserman (1999). On the Impossibility of Inferring Causation from Association without Background Knowledge. *Computation, Causation, and Discovery*. C. Glymour and G. Cooper. Menlo Park, CA, Cambridge, MA, AAAI Press/The MIT Press: 305-321.

Robinson, P. M. (1986). Non-Parametric Methods in Specification. *Economic Journal* **96 Supplement**: 134-41.

Romer, P. M. (1994). The Origins of Endogenous Growth. *Journal of Economic Perspectives* **8**: 3-22.

Rubin, D. B. (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics* **12**: 1151-72.

Rubinstein, A. (1991). Comments on the Interpretation of Game Theory. *Econometrica* **59**: 909-924.

Salmon, M. (1995). Bounded Rationality and Learning: Procedural Learning. *Learning and Rationality in Economics*. A. Kirman and M. Salmon. Oxford, Blackwell.

Salmon, W. (1998). *Causality and Explanation.*

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World.* Princeton, N.J., Princeton University Press.

Salmon, W. C. (1989). *Four Decades of Scientific Explanation.* Minneapolis, University of Minnesota Press.

Salmon, W. C., R. C. Jeffrey, et al. (1971). *Statistical Explanation & Statistical Relevance.*

Sargent, T. J. (1981). Interpreting Economic Time Series. *Journal of Political Economy* **89**: 213-248.

Sargent, T. J. (1984). Autoregressions, Expectations, and Advice. *American Economic Review,* **74**: 408-15.

Sargent, T. J. (1987). *Dynamic Macroeconomic Theory.* Cambridge, Mass., Harvard University Press.

Sargent, T. J. (1993). *Bounded Rationality in Macroeconomics: the Arne Ryde Memorial Lectures.* Oxford, Clarendon Press.

Savage, L. J. (1954 [1972]). *The Foundations of Statistics.* New York, Wiley.

Savage, L. J. (1962). Bayesian Statistics. *Recent Developments in Information and Decision Processes.* R. E. Machol and P. Gray. New York, Macmillan and Co.

Savage, L. J. (1967). Difficulties in the Theory of Personal Probability. *Philosophy of Science* **34**: 305-310.

Savage, L. J. (1967). Implications of Personal Probability for Induction. *Journal of Philosophy* **64**: 593-607.

Savage, L. J. (1971). Elicitation of Personal Probabilities and Expectations. *Journal of the American Statistical Association* **66**: 783-801.

Savage, L. J. (1971). Letter from Leonard Savage to Robert Aumann. *Essays on Economic Decisions under Uncertainty.* J. H. Dreze: 78-81.

Savage, L. J. (1977). The Shifting Foundations of Statistics. *Logic, Laws, and Life.* R. Colodny. Pittsburgh, University of Pittsburgh Press: 3-18.

Scharfstein, D. O., M. J. Daniels, et al. (2003). Incorporating Prior Beliefs about Selection Bias into the Analysis of Randomized Trials with Missing Outcomes. *Biostatistics* **4**: 495-512.

Scheines, R. (1994). Inferring Causal Structure among Unmeasured Variables. *Selecting Models from Data: AI and Statistics IV.* P. Chessman and R. W. Oldford, Springer-Verlag: 197-204.

Scheines, R. (1997). An Introduction to Causal Inference. *Causality in Crisis?* V. McKim and S. Turner, University of Notre Dame Press: 185-200.

Scheines, R., P. Spirtes, et al. (1998). The TETRAD Project: Constraint Based Aids to Causal Model Specification. *Multivariate Behavioral Research* **33**: 65-118.

Schelling, T. C. (1978). *Micromotives and Macrobehavior.* New York, Norton.

Schelling, T. C. (1998). Social Mechanisms and Social Dynamics. *Social Mechanisms: an Analytical Approach to Social Theory.* P. Hedstrm. Cambridge, UK, Cambridge University Press: 32-34.

Schervish, M. and T. Seidenfeld (1990). An Approach to Consensus and Certainty with Increasing Evidence. *Journal of Planning and Statistical Inference* **25**: 401-414.

Schick, F. (1986). Dutch Bookies and Money Pumps. *Journal of Philosophy* **83**: 112-119.

Schumpeter, J. A. (1954). *History of Economic Analysis.* Oxford, Oxford University Press.

Scott, D. (1992). *Multivariate Density Estimation, Theory, Practice, and Visualization.* New York, John Wiley and Sons.

Seidenfeld, T. (1979). Why I am not an Objective Bayesian: Some Reflections Promoted by Rosenkrantz. *Theory and Decision* **11**: 413-440.

Selten, R. (1990). Bounded Rationality,. *Journal of Institutional Economics* **146**: 649-658.

Sen, A. (1987). Rational Behaviour. *The New Palgrave: A Dictionary of Economics.* J. Eatwell, M. Milgate, P. Newman. **4:** 68-74.

Sen, A. (1993). Internal Consistency of Choice. *Econometrica* **61**: 495-521.

Shafer, G. (1982). Lindley's Paradox. *Journal of the American Statistical Association* **77**: 325-351.

Shafer, G. (1985). Conditional Probability. *International Statistical Review* **53**: 261-277.

Shafer, G. (1986). Savage Revisited. *Statistical Science* **1**: 463-501.

Shafer, G. (1998). Lindley's paradox. *Encyclopedia of Biostatistics,.* P. Armitage and T. Colton, Wiley. 1998. **3:** 2257-8.

Shafer, W. and H. Sonnenschein (1982). Market Demand and Excess Demand Functions. *Handbook of Mathematical Economics.* K. J. Arrow and M. D. Intriligator. Amsterdam, North Holland. **2:** 670-93.

Shanks, D. R. (1995). *The Psychology of Associative Learning.* Cambridge, Cambridge University Press.

Shibata, R. (1981). An Optimal Selection of Regression Variables. *Biometrica* **68**: 45-54.

Shiller, R. J. (1987). Rational Expectations and the Dynamic Structure of Macroeconomics Models. *Journal of Monetary Economics* **4**: 1-44.

Sidanius, J. (1988). Political Sophistication and Political Deviance: A Structural Equation Examination of Context Theory. *Journal of Personality and Social Psychology* **55:** 37-51.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* London, Chapman and Hall.

Simkins, S. (1999). Measurement and Theory in Macroeconomics, Department of Economics, Merrick Hall, North Carolina A&T State University. **2005**.

Simon, H. A. (1953). Causal Ordering and Identifiability. *Studies in Econometric Method,*. W. C. Hood and T.C. Koopmans. New York, John Wiley and Sons: 49-74.

Simon, H. A. (1954). Spurious Correlation: A Causal Interpretation. *Journal of the American Statistical Association* **49**: 467-479.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics* **69**: 99-118.

Simon, H. A. (1956). Rational Choice and the Structure of the Environment. *Psychological Review* **63**: 129-138.

Simon, H. A. (1960). *The New Science of Management Decision*, Harper, H. Hamilton.

Simon, H. A. (1973). Does Scientific Discovery Have a Logic? *Philosophy of Science* **40**: 471-480.

Simon, H. A. (1984). On the Behavioral and Rational Foundations of Economic Dynamics. *Journal of Economic Behaviour and Organization* **5**: 35-55.

Simon, H. A. (1986). Rationality in Psychology and Economics. *Journal of Business* **59**: S209-24.

Simon, H. A. (1990). Invariants of Human Behavior. *Annu. Rev. Psychol.* **41**: 1-19.

Simonson, I. and A. Tversky (1992). Choice in Context: Tradeoff Contrast and Extremeness Aversion. *Journal of Marketing Research* **29**: 281-215.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, B,* **13**: 238-241.

Sims, C. A. (1972). Money, Income, and Causality. *American Economic Review* **62**: 540-52.

Sims, C. A. (1977). Exogeneity and Causal Ordering in Macroeconomic Models. *New Methods in Business Cycle research.* C. A. Sims. Federal Reserve Bank of Minneapolis, MN: 23-43.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica* **48**: 1-48.

Sims, C. A. (1981). What Kind of Science is Economics. *Journal of Political Economy* **89**: 578-583.

Sims, C. A. (1982a). Policy Analysis with Econometric Models. *Brooking Papers on Economics Activity* **1**: 107-164.

Sims, C. A. (1982b). Scientific Standards in Econometric Modelling. *Current Developments in the Interface: Economics, Econometrics, Mathematics.* Dordrecht, Boston, and London, D. Reidel: 317-37.

Sims, C. A. (1986a). Are Forecasting Models Usable for Policy Analysis? *Federal Reserve Bank of Minneapolis Quarterly Review* **10**(Winter): 2-15.

Sims, C. A. (1987). Making Economics Credible. *Advances in Econometrics: Fifth World Congress.* T. Bewley. Cambridge: 49-61.

Sims, C. A. (1987b). A Rational Expectation Framework for Short-run Policy Analysis. *New Approaches to Monetary Economics*. W. Barnett and K. J. Singleton. Cambridge, Cambridge University Press: 293-308.

Sims, C. A. (1989). Models and Their Uses. *American Journal of Agricultural Economics* **71**: 489-494.

Sims, C. A. (1991). Empirical Analysis of Macroeconomic Time Series: VAR and Structural Models: Comments. *European Economic Review* **34**: 922-32.

Sims, C. A. (1992). Interpreting the Macroeconomic Time Series Facts. *European Economic Review* **36**: 975-1000.

Sims, C. A. (1996). Macroeconomics and Methodology. *Journal of Economic Perspectives,* **10**: 105-20.

Sims, C. A. (2004). An Interview with Christopher A. Sims. *Macroeconomic Dynamics* **8**: 273-94.

Skyrms, B. (1980). *Causal Necessity: a Pragmatic Investigation of the Necessity of Laws.* New Haven; London, Yale University Press.

Skyrms, B. (1984). *Pragmatics and Empiricism.* New Haven, Yale University Press.

Skyrms, B. (1986). *Choice and Chance: an Introduction to Inductive Logic.* Belmont, Calif, Wadsworth Pub. Co.

Skyrms, B. (1987). Dynamic Coherence and Probability Kinematics. *Philosophy of Science* **54**: 1-20.

Skyrms, B. (2004). *The Stag Hunt and the Evolution of Social Structure*, Cambridge University Press.

Slovic, P. (1995). The Construction of Preference. *American Psychologist* **50**: 364-371.

Slovic, P., D. Griffin, et al. (1990). Compatibility Effects in Judgment and Choice. *Insights in Decision Making: A Tribute to Hillel J. Einhhorn.* R. Hogarth. Chicago, University of Chicago Press: 5-27.

Slovic, P. and S. Lichtenstein (1968). The Relative Importance of Probabilities and Payoffs in Risk-Taking. *Journal of Experimental Psychology, Monograph Supplement* **78**: 1-18.

Slovic, P. and A. Tversky (1974). Who Accepts Savage's Axioms? *Behavioral Science* **19**: 368-373.

Smith, A. (1976). *An Inquiry into the Nature and Causes of the Wealth of Nations.* Oxford, Clarendon.

Smith, A. F. M. (1984). Bayesian Statistics, Present Position and Potential Developments: Some Personal Views. *Journal of the Royal Statistical Society A* **147**: 245-259.

Smith, A. F. M. (1986). Some Bayesian Thoughts on Modelling and Model Choice. *The Statistician* **35**: 97-102.

Snowdon, B., H. Vane, et al. (1994). *A Modern Guide to Macroeconomics.* Aldershot, Hampshire, Edward Elgar.

Snowdon, B. and H. R. Vane (1997). *A Macroeconomics Reader*. London, Rutledge.

Sobel, J. (2000). Economists' Models of Learning. *Journal of Economic Theory* **94**: 241-261.

Sobel, M., E (1990). Effect Analysis and Causation in Linear Structural Equation Models. *Psychometrika* **55**: 495-515.

Sobel, M. E. (1995). Causal Inference in the Social and Behavioral Sciences. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. G. Arminger, C. Clogg, and M.E. Sobel. New York, Plenum Press: 1-38.

Sobel, M. E. (1998). Causal Inference in Statistical Models of the Process of Socioeconomic Achievement. *Sociological Methods and Research* **27**(2): 318-348.

Sobel, M. E. (2000). Causal Inference in the Social Sciences. *Journal of the American Statistical Association* **95**: 647-51.

Sober, E. (1987). The Principle of the Common Cause. *Probability and Causation: Essays in Honor of Wesley Salmon*. J. Fetzer. Dordrecht, Reidel: 211-228.

Sober, E. (2001). Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause. *British Journal for the Philosophy of Science* **52**: 331-346.

Sonnenschein, H. (1972). Market Excess Demand Functions. *Econometrica* **40**: 549-63.

Sonnenschein, H. (1973). The Utility Hypothesis and Market Demand Theory. *Western Economic Journal* **11**: 404-10.

Spanos, A. (1986). *Statistical Foundations of Econometric Modelling*. Cambridge, Cambridge University Press.

Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge, New York, Cambridge University Press.

Spanos, A. (2000). Revisiting Data Mining: 'Hunting' with or without a Licence. *Journal of Economic Methodology* **7**(2): 231-262.

Spanos, A. (2000). Where Do Statistical Models Come From? Revisiting the Problem of Specification. Blacksburg, Department of Economics, Virginia Tech.

Spanos, A. and A. McGuirk (2001). Econometric Methodologies for the Model Specification Problem: Addressing Old Problems in the New Century: The Model Specification Problem from a Probabilistic Reduction Perspective. *American Journal of Agricultural Economics* **83**: 1168-1176.

Spiegelhalter, D. J. (1995). Discussion of "Assessment and Propagation of Model Uncertainty" by D. Draper. *Journal of the Royal Statistical Society Series B* **57**: 45-97.

Spirtes, P. (1995). Directed Cyclic Graphical Representations of Feedback Models. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. P. Besnard and S. Hanks. San Mateo, CA, Morgan Kaufmann: 491-8.

Spirtes, P. (1997). Limits on Causal Inference from Statistical Data. presented at American Economics Association Meeting, http://www.hss.cmu.edu/philosophy/people/directory/Peter_Spirtes.html. **1999**.

Spirtes, P., C. Glymour, et al. (1991). From Probability to Causality. *Philosophical Studies* **64**: 1-36.

Spirtes, P., C. Glymour, et al. (1993). *Causation, Prediction, and Search.* New York, Springer-Verlag.

Spirtes, P., C. Glymour, et al. (1997). Reply to Humphrey's and Freedman's Review of Causation, Prediction, and Search. *British Journal for the Philosophy of Science* **48**: 555-568.

Spirtes, P., C. Meek, et al. (1996). Causal inference in the Presence of Latent Variables and Selection Bias. **II**.

Spirtes, P., R. Richardson, et al. (1998). Using Path Diagrams as a Structural Equation Modeling Tool. *Sociological Methods and Research* **27**: 148-181.

Spirtes, P. and T. Richardson (1996). A Polynomial Time Algorithm for Determining DAG Equivalence in the Presence of Latent. *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics (January 4-7, Fort Lauderdale, FL).*

Spirtes, P., T. Richardson, et al. (1997). Using Path Diagrams as a Structural Equation Modelling Tool. *Sociological Methods and Research* **27**: 182-225.

Spirtes, P., T. Richardson, et al. (1996a). Using D-separation to Calculate Zero Partial Correlation in Linear Models with Correlated Errors. Pittsburgh, PA, Department of Philosophy, Carnegie Mellon University.

Spirtes, P. and R. Scheines (1997). Reply to Freedman. *Causality in Crisis?* S. Turner and V. McKim, University of Notre Dame Press.

Spirtes, P., R. Scheines, et al. (1990). Reply to Comments. *Sociological Methods and Research* **19**: 107-121.

Spirtes, P. and T. S. Verma (1992). Equivalence of Causal Models with Latent Variables,Technical Report CMU-PHIL-33, Department of Philosophy. Pittsburgh, Carnegie Mellon University.

Spohn, W. (1980). Stochastic Independence, Causal Independence, and Shieldability. *Journal of Philosophical Logic* **9**: 73-99.

Spohn, W. (1982). How to Make Sense of Game Theory, Studies in Contemporary Economics. *Philosophy of Economics.* W. Stemuller, W. Balzer and W. Spohn. Berlin, Springer. **2**: 239-270.

Stelzl, I. (1986). Changing a Causal Hypothesis without Changing the Fit: Some Rules for Generating Equivalent Path Models,. *Multivariate Behavioral Research* **21**: 309-331.

Stigler, G. and G. Becker (1977). De Gustibus Non Est Disputandum. *American Economic Review* **67**: 76-90.

Stiglitz, J. E. (1991). Alternative Approaches to Macroeconomics: Methodological Issues and the New Keynesian Economics. Cambridge, MA, NBER Working Papers Series.

Stoker, T. M. (1984). Completeness, Distribution Restrictions, and the Form of Aggregate Functions. *Econometrica* **52**: 887-907.

Stoker, T. M. (1986). Simple Tests of Distributional Effects on Macroeconomic Equations. *Journal of Political Economy* **94**: 763-95.

Stoker, T. M. (1993). Empirical Approaches to the Problem of Aggregation over Individuals. *Journal of Economic Literature* **XXXI**: 1827-1874.

Stone, M. (1974). Cross-Validatory Choice of and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, Series B* **36**: 111-133.

Stone, M. (1977). On Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaik's Criterion. *Journal of the Royal Statistical Society, Series B* **39**: 44-47.

Stone, M. (1979). Comments on Model Selection Criteria of Akaike and Schwarz. *Journal of Royal Statistical Society B* **41**: 276-278.

Stone, R. (1993). The Assumptions on Which Causal Inference Rest. *Journal of the Royal Statistical Society, Series B* **55**: 455-66.

Summers, L. H. (1991). The Scientific Illusion in Empirical Macroeconomics. *The Scandinavian Journal of Economics* **93**: 129-148.

Suppes, p. (1961). The Philosophical Relevance of Decision Theory. *The Journal of Philosophy* **58**: 605-614.

Suppes, P. (1969). *Studies in the Methodology and Foundations of Science: Selected Papers from 1951 to 1969.* Dordrecht, D. Reidel.

Suppes, P. (1970). *A Probabilistic Theory of Causality.* Amsterdam, North-Holland Publishing Co.

Swanson, N. R. and C. W. J. Granger (1997). Impulse Response Functions Based on a Causal Approach to Residual Orthogonalisation in Vector Autoregressions. *Journal of American Statistical Association* **92**: 357-367.

Taylor, J. B. (1977). Conditions for Unique Solutions in Stochastic Macroeconomic Models with Rational Expectations. *Econometrica* **45**: 1377-1385.

Teller, P. (1973). Conditionalization and Observation. *Syntheses* **26**: 218-258.

Tesfatsion, L. (2003). Non-Walrasian Equilibrium: Illustrative Examples, http://www.econ.iastate.edu/classes/econ606/tesfatsion/syl606t.htm#Intro.

Theil, H. (1954). *Linear Aggregation of Economic Relations.* Amsterdam, North-Holland.

Theil, H. (1962). Alternative Approaches to The Aggregation Problem. *Logic, Methodology and Philosophy of Science.* E. Nagel, P. Suppes and A. Tarski: 507-527.

Thrall, R. M. (1954). Applications of Multidimensional Utility Theory. *Decision Processes.* R. M. Thrall, C.H. Coombs and R. L. Davis, John Wiley: 181-186.

Tinbergen, J. (1939). *Statistical Testing of Business Cycle Theories 2.vols.* Geneva, League of Nations.

Tirole, J. (1982). On the Possibility of Speculation under Rational Expectations. *Econometrica* **50**: 1163-81.

Tversky, A. (1969). Intransitivity of Preferences. *Psychological Review* **76**: 31-48.

Tversky, A. (1975). A Critique of Expected Utility Theory: Descriptive and Normative Considerations. *Erkenntnis* **9**: 163-173.

Tversky, A. and D. Kahneman (1981). The Framing of Decisions and the Psychology of Choice. *Science* **211**: 453-458.

Tversky, A. and D. Kahneman (1986). Rational Choice and the Framing of Decisions. *Journal of Business* **59**: 251-278.

Tversky, A. and E. Shafir (1992). Choice under Conflict: the Dynamics of Deferred Decision. *Psychological Science* **3**: 358-361.

Tversky, A., P. Slovic, et al. (1990). The Causes of Preference Reversal,. *American Economic Review* **80**: 204-217.

Tversky, A. and R. Thaler, H. (1990). Anomalies: Preference Reversals. *Journal of Economic Perspectives* **4**: 201-11.

Van Daal, J. and A. H. Q. M. Merkies (1984). *Aggregation in Economic Research: from Individual to Macro Relations.* Dordrecht, D. Reidel.

Vaughn, K. I. (1989). Invisible Hand. *The New Palgrave.* J. Eatwell, M. Milgate and P. Newman. London, Macmillan. **2:** 997-999.

Vercelli, A. (1991). *Methodological Foundations of Macroeconomics: Keynes and Lucas.* Cambridge, The University Press.

Verma, T. S. and J. Pearl (1990). Equivalence and Synthesis of Causal Models. *Proceedings of the 6th Conference on Uncertainty in AI*: 220-227.

Villegas (1977). On the Representation of Ignorance. *Journal of the American Statistical Association* **72**: 651-654.

Von Wright, G. H. (1971). On the Logic and Epistemology of the Causal Relation. *Causation and Conditionals (1987).* E. Sosa and M. Tooley. Oxford, Oxford University Press: 105-124.

Vriend, N. J. (1996). Rational Behaviour and Economic Theory. *Journal of Economic Behaviour and Organization* **29**: 263-285.

Wahba, G. and S. Wold (1975). A Completely Automatic French Curve: Fitting Spline Functions by Cross Validation. *Communications in Statistics* **4**: 125-142.

Wall, K. (1993). A Model of Decision Making under Bounded Rationality. *Journal of Economic Behaviour and Organisation* **21**: 331-52.

Wallace, N. (1980). The Overlapping Generations Model of Fiat Money. *Models of Monetary Economics.* J. H. Kareken and N. Wallace. Minneapolis, Federal Reserve Bank of Minneapolis, MN).

Warner, B. and M. Manavendra (1996). Understanding Neural Networks as Statistical Tools. *The American Statistician* **50**: 284-293.

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology* **44**: 92-107.

Weakliem, D. (1999). A Critique of the Bayesian Information Criterion. *Sociological Methods and Research* **27**: 359-397.

Weatherford, M. S. (1983). Economic Voting and the "Symbolic Politics" Argument. *American Political Science Review* **77**: 158-74.

Weisberg, S. (1985). *Applied Linear Regression.* New York, John Wiley and Sons.
Welsch, R. E. (1986). Comment. *Statistical Science* **1**: 403-5.

Wermuth, N. (1980). Linear Recursive Equations, Covariance Selection, and Path Analysis. *Journal of the American Statistical Association* **75**: 963-972.

Wermuth, N., D. R. Cox, et al. (1994). Explanations for Multivariate Structures Derived from Univariate Recursive Regressions, University of Mainz.

White, H. (1984). Tests of Specification in Econometrics: Comment. *Econometric Reviews* 3(1984): 261-67.

White, H., Ed. (1992). *Artificial Neural Networks: Approximation and Learning Theory.* Cambridge and Oxford, Blackwell.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics.* Chichester, England, John Wiley and Sons.

Williams, L. J., H. Bozdogan, et al. (1996). Inference Problems With Equivalent Models. *Advanced Structural Equation Modelling.* A. G. Marcoulides and R. E. Schumacker. Mahwah, New Jersey, Lawrence Erlbaum Associates, Publisher: 279-314.

Williams, P. M. (1980). Bayesian Conditionalisation and the Principle of Minimum Information. *British Journal for the Philosophy of Science* **31**: 131-144.

Williamson, J. (1999). Countable Additivity and Subjective Probability. *British Journal for the Philosophy of Science* **50**: 401-416.

Williamson, P. (1997). Learning and Bounded Rationality. *Journal of Economic Surveys* **11**: 221-230.

Winkler, R. L. (1980). Prior Information, Predictive distribution, and Bayesian model-building. *Bayesian Analysis in Econometrics and Statistics.* A. Zellner. Amsterdam, North: Holland.

Winkler, R. L. (1994). Model Uncertainty: Probabilities for Models? *Model Uncertainty: Its Characterization and Quantification.* A. Mosleh, N. Siu, C. Smidts and C. Lui. Washington, D.C., U. S. Nuclear Regulatory Commission: 107-116.

Woodward, J. (1988). Understanding Regression. *PSA: The Philosophy of Science Association* **1**: 255-269.

Woodward, J. (1999). Causal Interpretation in Systems of Equations. *Synthese* **121**(199-257).

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation.* New York, Oxford University Press.

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research* **20**: 557-585.

Wright, S. (1934). The Method of Path Coefficient. *Annals of Mathematical Statistics* **5**: 161-215.

Yatchew, A. J. (1998). Nonparametric Regression Techniques in Economics. *Journal of Economic Literature* **36**: 669-721.

Yule, G. U. (1903). Notes on the Theory of Association of Attributes in Statistics. *Biometrica* **2:**: 121-134.

Yule, G. U. (1926). Why Do We Sometimes Get Nonsense Correlations Between Time Series? A Study of Sampling and the Nature of Time Series (with discussion). *Journal of the Royal Statistical Society* **89**: 1-64.

Zarembka, P. (1974). *Frontiers in Econometrics*, Academic Press.

Zellner, A. (1969). On the Aggregation Problem: A New Approach to a Troublesome Problem. *Estimation and Risk Programming: Essays in honour of Gehard Tintner*. Berlin, Springer: 365-378.

Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York, John Wiley.

Zellner, A. (1975). Bayesian Analysis of Regression Error Terms. *Journal of the American Statistical Association* **70**: 138-144.

Zellner, A. (1987). Bayesian Inference. *The New Palgrave: A Dictionary of Economics*. J. Eatwell, M. Milgate and P. Newman. London, Macmillan: 208-218.

Zucchini, W. (2000). An Introduction to Model Selection. *Journal of Mathematical Psychology* **44**: 41-61.