The London School of Economics and Political Science

Essays in Applied Macroeconomics

Jonathan Pinder

A thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy. London, July 2017

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author.

Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of approximately 38,000 words.

Jonathan Pinder wrote this thesis in a personal capacity and any views herein solely belong to him or his co-authors.

Statement of conjoint work

I certify that chapter 2 of this thesis was co-authored with Xiaoming Cai and Wouter den Haan. I contributed 33% of the work in this chapter.

Statement of use of third party for editorial help

I confirm that chapter 3 of this thesis was copy edited for conventions of language, spelling and grammar by Roger Pinder.

Abstract

I present a thesis in three chapters in the broad field of Applied Macroeconomics.

The first chapter is an empirical investigation into the “granular hypothesis” - the hypothesis that shocks to extremely large firms can have aggregate economic consequences. Identifying this channel is nontrivial as it may be the case that large firms respond more to aggregate shocks than most firms. I present a new way to identify true firm-level shocks by looking at stock price movements around the times that firms release financial information. I argue such movements reflect firm-specific, rather than aggregate information. Using a measure of firm shocks recovered using this information suggests that the importance of such shocks for aggregate economic outcomes has been overestimated by previous work in the literature.

A good univariate representation of US GDP is a random walk with drift. The second chapter shows that nonetheless US recessions have been associated with predictable short-term recoveries with relatively small changes in long-term GDP forecasts. To detect these predictable changes, it is important to use a multivariate time series model. We discuss reasons why univariate representations can miss key characteristics of the underlying variable such as predictability, especially during recessions.

The third chapter develops a general equilibrium model to investigate the macroeconomic consequences of liquidity regulation, a form of regulation which was strengthened substantially after the 2008 financial crisis. The model is used to identify two separate channels through which liquidity regulation can affect the cost of capital: the “crowding out” and “financial repression” channels. In the absence of these, I establish a neutrality result in which liquidity regulation does not affect the wider economy. The principal policy implication of this chapter is that regulators should not count safe assets which they require banks to hold for liquidity purposes against bank capital requirements.

Acknowledgements

This thesis would not have been possible without the help and support of a great many people.

First, I would like to thank my supervisor, Wouter den Haan, for his advice and guidance throughout my time at LSE. It is no exaggeration to say that this work would not have been possible without his input. I will forever be indebted to him for the enormous amount of time he has spent with me picking over models, discussing (and - thankfully - often dismissing) my latest idea and making me think critically about my own work.

I am also grateful to the entire macro faculty at the London School of Economics for comments in seminars and discussions about research. Ongoing conversations with Ethan Ilzetzki and David Baqaee were particularly helpful. The privilege of sitting in macro seminars at LSE - both for internal and external speakers - was a fantastic opportunity to learn from some of the sharpest minds in the field.

My peers at LSE have always been fantastic - and were able to understand the challenges as only those in the PhD process can. I would particularly like to thank Angus Foulis, Alex Clymo, Tom Carr, Andrea Lanteri, Sergio de Ferra and Markus Riegler for helpful discussions from the very start, and later Filip Rozsypal, Thomas Dreschel, Andreas Ek, Sevim Kosem, Ryan Liang, and Anna Rogantini Picco. I fear that I benefited a lot from the advice of the first group without being to directly reciprocate, but I hope that I was able to pass on some of their distilled wisdom to some in the second group (though to be clear I am grateful for discussions with and support from the second group as well). The benevolent overlapping generations model still applies between cohorts of LSE PhD students.

I would also like to thank the numerous people who encouraged me to study economics at a higher level, particularly Michael Pollitt and Max Beber, who encouraged me and pushed me as I started learning about this subject at a university level. I would also like to thank Ahmet Akarli and Rory MacFarquhar for pointing me in the direction of further study, and, in Ahmet's case, being understanding during the final year of writing up.

Outside of LSE, I am lucky to have been surrounded by a number of extremely supportive friends. There are perhaps too many to mention here, but at the risk of singling people out I would like to thank Caezar Al-Jassar, Kuly Heer, Majeed Neky, Fraser Davison and Karthik Ramanathan. Jess Dell and my brother Tim Pinder both deserve credit for tolerating me throughout the process.

Finally, I would like to thank my parents, Roger and Joyce Pinder, who have encouraged and supported me throughout my education. I dedicate this thesis to them.

Contents

List of Figures

List of Tables

Introduction

This thesis consists of three chapters which fall within the broad field of applied macroeconomics. The first and third chapters focus on investigating the importance of individual firms or sectors on aggregate fluctuations. One implication of the fact that many different firms or sectors can matter for the aggregate economy is that empirical models of the latter need to be sufficiently complex to capture rich dynamics such heterogeneity would imply. The second chapter explores this point.

The first chapter focuses on the importance of the very largest firms, which, as pointed out by Gabaix (2011) in principal can explain aggregate fluctuations due to their size. In a "big picture" sense such shocks could provide a microfounded explanation for Total Factor Productivity (TFP) shocks, which have remained a staple of modern macroeconomic models even though it is hard to identify examples of such shocks (as argued, for instance, by Summers (1986)). While the theoretical possibility that such shocks are quantitatively important comes from the firm size distribution, the empirical challenge has always been distinguishing true firm-level shocks from aggregate shocks. It is clearly not sufficient to note that the very largest firms are more cyclical than the average firm, as this is consistent *either* with the hypothesis that large firms drive the business cycle *or* with the hypothesis that large firms are more sensitive to the business cycle than the average firm. For example, in the US economy (the economy I study in Chapter One) the largest firms throughout a large part of the post war period were Car Manufacturers, namely General Motors and Ford. There are good reasons to think that Car Manufacturers are more sensitive to aggregate economic conditions than other companies – cars are a consumer durable, and all such goods are typically very cyclical as households postpone purchasing them during recessions. To overcome this key empirical challenge I use variation in firm stock prices in daily intervals around when firms release financial information (typically quarterly earnings results) which I argue (both *a priori* and using suggestive evidence regarding the behaviour of stock prices of different subsets of firms) contains firm-level information but not aggregate information. I then use this variation to create a measure of firm shocks which should be "clean" of aggregate information and examine how aggregate output and productivity respond to this measure of aggregate shocks relative to the existing measures in the literature.

While the first chapter focuses on firms which are distinguished by their size, the third chapter focuses on a sector – financial intermediation – which is distinguished because of its unique function. This chapter investigates the consequences of liquidity regulation on the cost of capital for productive firms and hence aggregate output. The model I develop identifies two channels through which liquidity regulation can raise the cost of capital, which I term the "crowding out" and "financial repression" channels. I demonstrate that in the absence of these channels, a *neutrality result* exists where financial intermediaries are indifferent about holding any amount of liquid assets and so liquidity regulation has no impact. I also calibrate my model to US data and quantitatively assess the impact that liquidity regulation can have on aggregate output, both in the steady state and during a transition period after a policy change which increases liquidity requirements. The principal policy recommendation from this work is that if financial intermediaries are required to hold liquid assets by regulatory bodies, these should not count towards capital requirements, which they do under a simple "leverage ratio" which have been introduced to supplement risk-weight based capital requirements since the 2008 financial crisis.

One implication of these chapters is that simple time series models are unlikely to be rich enough to capture the full dynamics of the macroeconomy, as many different shocks in principal can affect the economy. Campbell and Mankiw (1987) noted that a random walk with drift is a good representation of US GDP, which implies that any fall in GDP would be expected to be permanent relative to a pre-recession growth path. In this chapter we demonstrate that many US recessions are associated with predictable recoveries, but only if the econometrician uses a multivariate model. The reason for this is that GDP is a sum of stationary and nonstationary components. As such, the correct univariate specification of GDP would indicate that it is nonstationary, which would indicate that a shock to this model would imply that GDP would fall permanently. However any shock to a stationary component would imply predictable future changes in GDP, which would not be captured by a univariate representation. We investigate whether this point matters in US and UK data, and demonstrate that it does. We find that in recessions a simple multivariate model typically outperforms its univariate counterpart, though there is no discernible difference in normal times.

Chapter 1

Assessing the Granular Hypothesis with High Frequency Financial Information

1.1 Introduction

What are the shocks which drive business cycle fluctuations? In many business cycle models, shocks to Total Factor Productivity (TFP) are important. But such shocks do not have a ready interpretation, nor counterpart in the data. This missing element has meant that some, such as Summers (1986) have questioned the notion that such shocks are truly the source of business cycle fluctuations. More recently, Chari et al. (2007) treat a series of exogenous TFP recovered from a prototype business cycle framework as a reduced form "wedge" rather than a series of structural innovations.

If part of these wedges truly are structural, one possible interpretation is that they are the sum of productivity shocks to either firms or sectors. It is easier to conceptualise what a productivity shock might look like at the level of a firm than it is for the economy as a whole. One obvious candidate is variation in the firm-level adoption of new technology or production processes. Another related explanation would recognise that running a firm is a complex task, and so mistakes by firm management are possible (and can lead to both positive and negative productivity changes in firm productivity).[1] However, as there are many firms, the possibility that firm-level shocks might quantitatively explain aggregate fluctuations was discounted until recently. But in an influential paper, Gabaix (2011) argued that in principle shocks to the largest firms could explain a sizeable fraction of aggregate fluctuations in output and productivity. This is because the firm-size distribution in practice is sufficiently fat tailed that the effect of firm shocks decays much more slowly with the number of firms N than one would normally expect if we were able to apply a standard central limit theorem.[2]

The contribution of this paper is to introduce a novel way of identifying firm-level shocks. Distinguishing aggregate and firm-level shocks is non-trivial as the effect of the former can vary over firms, while shocks to large firms might affect other firms contemporaneously through either input-output linkages or market prices. I attempt to deal with this problem by using high frequency information on firm stock prices around the time that a firms releases their quarterly earnings reports. I argue that the movement of firm stock prices at this time is (i) correlated with the productivity shock which relates to the firm in question, and (ii) uncorrelated with any aggregate shock process. Given these two assumptions, I estimate a collection of VARs which contain both firm and aggregate information. I use the external proxy methodology proposed by Stock and Watson (2012) and Mertens and Ravn (2013) to obtain estimates of firm shocks. I then attempt to quantify the aggregate consequences of these shocks in practice using the same method as Gabaix (2011). I find that his original results considerably overstate the importance of such shocks relative to my measure.

This method contrasts with the two principal existing approaches in the literature. The first, performed by Gabaix (2011) simply used the (equally-weighted) firm productivity growth rate as a measure of the aggregate shock, and found that firm shocks appear to explain a large share of aggregate fluctuations. However, this ignores the possibility that larger firms might be more (or less) sensitive to aggregate shocks than the average firm. More recently, Stella (2015) has applied a factor model approach first developed at the

[1] The idea that firm managers can make mistakes seems at least as plausible as the idea that monetary policy makers' mistakes are the source of monetary policy shocks, which is one story used to justify the existence of the latter.

[2] Key in this is that granular shocks truly apply at the firm level, rather than (say) at the plant level. This is important as the former cannot be diversified by increasing the size of the firm. In this, appealing to managerial decisions or mistakes seems a natural source of firm-level productivity variation.

sectoral level by Foerster et al. (2011) to firms, an approach which can accomodate the different sensitivity of firms to aggregate shocks. This work is an important contribution to the empirical debate, and the results suggest that a very small fraction of aggregate fluctuations come from true granular shocks. However, this approach relies on both a number of strong theoretical assumptions and accurate data on input-output links between firms: without these, granular shocks can be misattributed as aggregate shocks. Unfortunately, data on network links between firms is highly incomplete at the firm level.

Aside from being related to the literature on granular business cycles, this paper is related to attempts to recover estimates of other fundamental economic shocks. There is a long literature attempting to do this for many other types of shock, including Romer and Romer (1989), Romer and Romer (2004) in the case of monetary policy shocks; Blanchard et al. (2002), Ramey (2011) and Mertens and Ravn (2013) for fiscal shocks and Kilian (2009) for shocks from the oil market, among many others. To my knowledge this is the first paper to attempt to apply the methodology common in these areas to examining the importance of firm-level shocks.

This chapter proceeds as follows. Section 1.2 discusses the literature. Section 1.3 sets out a standard business cycle model which I use to examine previous attempts to estimate firm shocks and to motivate my own emprical work. Section 1.4 discusses the data, while sections 1.5 and 1.6 discuss the firm and aggregate results. Section 1.7 concludes.

1.2 Literature Review

There is a long literature on whether idiosyncratic shocks to either firms or sectors can be large enough to explain a nontrivial quantity of aggregate fluctuations. Early important theoretical contributions include Long and Plosser (1983) and Jovanovic (1987). However, in many models, getting sectoral shocks to be quantitatively important relied on imposing only a small number of sectors.[3] The reason for this is standard: assuming that idiosyncratic shocks are distributed according to a "typical" (non-fat tailed) distribution,

[3] Gabaix (2011) argues that Long and Plosser (1983) use such a small number of sectors such that these shocks are, in effect "mini-aggregate" shocks

with N firms (or sectors) the importance of idiosyncratic shocks to overall volatility will diminish quickly as N grows due to a central limit theorem.

Gabaix (2011) argued that shocks to firms could be important for aggregate fluctuations *even if* the number of firms, N, is large. The essence of his argument is that firm size follows a power law distribution, and so shocks to the very largest firms might have a quantitatively important effect on economic aggregates. More precisely, the volatility of a sum of N independent and identially distributed Normal shocks declines at rate $N^{-\frac{1}{2}}$, while the volatility of the sum of N i.i.d. shocks distributed according to Zipf's law declines at rate $ln(N)$. He also presents some empirical evidence, based on demeaning the growth rates of large firms, that shocks to large firms are important. Carvalho and Grassi (2015) extend this logic by altering the heterogenous firm framework of Hopenhayn (1992) to allow for very large firms, and for shocks to large firms to dampen firm-level shocks. Calibrating their measure of firm shocks using firm sales data, and matching the firm size distribution in the US they are able to quantitatively explain around a quarter of aggregate fluctuations just with firm-level shocks.

Gabaix (2011)'s striking empirical results rely on a number of simplifications. First, his setup ignores the possibility that firms may vary in their response to aggregate growth. If large firms are more cyclical than the average firm, his "granular residual" would tend to overstate the importance of firm level shocks, as some of the reaction of large firms to aggregate fluctuations would be mis-recorded as firm level shocks. Conversely, the opposite would be true if large firms were less cyclical than the average firm. In addition, his empirical measure uses the growth rate in the sales to employee ratio as the measure of firm productivity, rather than value added, which would be a more typical way to measure firm TFP.

These results have been assessed in a number of other empirical studies. di Giovanni et al. (2014) use a panel of large French firms with detailed data on inter-firm linkages and find that firm-level shocks matter as much as aggregate shocks for aggregate fluctuations in France, with approximately one quarter of the effects of firm level shocks coming from direct effects, and the remaining three-quarters from input-output linkages. However, due to the relatively short time dimension of their panel, they are forced to assume (as with Gabaix (2011)) that all firms respond symmetrically to aggregate fluctuations.

An alternative approach is to use a dynamic factor model with a sufficiently long time dimension which allows the loadings on the aggregate factor to vary across firms. An important contribution was made by Stella (2015) who adapts the method of Foerster et al. (2011) to do this on a panel of 500 large US firms using quarterly Compustat data. The factor model approach requires that the econometrician knows how shocks to firm i affect all other firms to allow the factor model to be able to distinguish firm and aggregate shocks. In the model developed by Foerster et al. (2011), such spillovers can occur through network links between firms and accounting for these is important in performing the decomposition of productivity. Stella (2015) uses the Compustat Historical Customer Segments database to construct firm linkages. As he notes, however, this database likely understates the true input-output linkages across firms as it only records "major customers" (those customers accounting for more than 10% of firm revenue), and the low estimated input shares of such links (the calibrated intermediate-input shares of firms in his sample has a mean of 0.02 and a median "close to" zero) also may indicate that some input-output links are missing.

Aside from data problems, the structural setup of Foerster et al. (2011) might also be overly restrictive when quantifying the importance of firm shocks. Atalay (2014) argues that assuming that intermediate inputs enter in a Cobb-Douglas production function (i.e. with a unitary elasticity of substitution) understates the importance of sectoral shocks. He argues that the inputs provided by different sectors have an elasticity of substitution with the output of other sectors far below 1 (typically between 0.2 to 0.4) which dramatically increases the fraction of aggregate output fluctuations which can be accounted for by sectoral shocks. It is not clear whether this result holds when considering firms, rather than sectors (indeed, firms with competitors in the same sector have close substitutes by definition, which might mean that the elasticity of substitution is greater than 1).

Practice in the existing granular literature is that the productivity growth of firms is typically measured using sales.[4] In addition to such measures, I construct a measure of firm-level TFP growth using Compustat data in line with the method proposed by Imrohoroglu and Tüzel (2014) as a robustness check.

[4] Gabaix (2011) uses the growth of the ratio of sales to employees and Stella (2015) uses the growth rate of sales alone.

1.3 Framework

This section first outlines a simple model which will be useful to structure thinking about my emprical work. I then examine existing empirical methods in the literature which attempt to recover a measure of firm shocks through the lens of this model, before explaining my empirical strategy.

1.3.1 Model

The model consists of a continuum of households, which supply labour and consume $N \times S$ varieties of consumption goods, each produced by a different firm.

Household

The household supplies labour to firms and receives any profit, which are spent on N varieties of consumption goods from each of S sectors:

$$\max_{\{c_{ist}, l_t\}_{t=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t (u(C_t) - v(L_t)),$$

subject to:

$$C_t = \left(\sum_{s=1}^{S} \sum_{i=1}^{N} \nu_{is}^{\frac{1}{\varepsilon}} c_{ist}^{\frac{\varepsilon-1}{\varepsilon}} \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{1.1}$$

$$\sum_{s=1}^{S} \sum_{i=1}^{N} p_{it} c_{it} = w_t L_t + \Pi_t, \tag{1.2}$$

where $\sum_{s=1}^{S} \sum_{i=1}^{N} \nu_{is} = 1$. The household's first order conditions are standard:

$$c_{ist} = v_{is} \left(\frac{p_{ist}}{P_t} \right)^{-\varepsilon} C_t, \text{and} \tag{1.3}$$

$$\frac{v'(L_t)}{u'(C_t)} = w_t, \tag{1.4}$$

where

$$P_t = \left(\sum_{s} \sum_{i} \nu_{is} p_{ist}^{1-\varepsilon} \right)^{\frac{1}{1-\varepsilon}}. \tag{1.5}$$

Firms

Firms choose labour input to maximise profits subject to their downward sloping demand curve and taking wages as given:

$$\max_{c_{ist}, n_{ist}, p_{ist}} \pi_{ist} = p_{ist} z_{ist} n_{ist}^{\alpha} - w_t n_{ist}, \tag{1.6}$$

subject to:

$$c_{ist} = z_{ist} n_{ist}^{\alpha}, \tag{1.7}$$

$$z_{ist} = \lambda_{si} f_t + \gamma_{si} f_{st} + u_{ist}, \tag{1.8}$$

and the demand curve implied by (1.3). Note that the firm's productivity follows a factor structure, where the factor loading λ_{is} measures the firm specific response to aggregate shocks f_t, γ_{is} measures the firm response to a process f_{st} which only affects firms in sector s, and u_{ist} is the idiosyncratic firm-specific component of productivity.[5] The firm's first order condition is

$$w_t + \frac{\partial w_t}{\partial n_{ist}} n_{ist} = z_{ist} \Big(\frac{\partial p_{ist}}{\partial c_{ist}} \frac{\partial c_{ist}}{\partial n_{ist}} n_{ist}^{\alpha} + p_{ist} \alpha n_{ist}{}^{\alpha-1} \Big), \tag{1.9}$$

where the marginal cost of production (on the left-hand side of this expression) is increasing in n_{ist} while the marginal revenue (right-hand side) is decreasing both because of the effect of increased supply on both prices (first term) and the marginal product of labour at a given price (second term). Again, this is standard except for the fact that I allow for "large" firms in that the choice of n_{ist} may affect aggregate wages.

Market clearing implies that $L_t = \sum_{s=1}^{S} \sum_{i=1}^{N} n_{ist}$, and the representative household receives the profit that all firms make: $\Pi_t = \sum_{s=1}^{S} \sum_{i=1}^{N} \pi_{ist}$.

[5] One possibility not discussed in the literature is a time varying factor loading:

$$\begin{aligned} z_{ist} &= \lambda_{ist} f_t + \gamma_{si} f_{st} + u_{ist} \\ &= \bar{\lambda}_{is} f_t + \gamma_{si} f_{st} + \{(\lambda_{ist} - \bar{\lambda}_{is}) f_t + u_{ist}\}, \end{aligned}$$

where the "firm shock" (now contained in braces) has a slope component (depending on the aggregate factor) as well as a level component. It is reasonable to interpret a shock to λ_{ist} as a firm (rather than an aggregate) shock, as it affects the productivity of firm i in sector s without directly influencing the productivity of any other firm. In practice accomodating such a specification would be computationally prohibitive in a factor model approach. In the external proxy approch detailed in section 1.3.4 the proxy would simply be correlated with the term in braces and uncorrelated with f_t.

Notice that firms are technologically symmetric, and variations in firm size come from variations in taste (i.e. ν_{is}). Firms which produce goods with a high ν_{is} face higher relative prices and hence employ more labour. Idiosyncratic shocks to these large firms u_{ist} will have a larger effect on total output than the equivalent shocks for smaller firms, and following the argument of Hulten (1978) discussed in Section 1.3.4, the effect of each shock on aggregates is linear in the sales of the firm.

The task of the econometrician is to extract estimates of u_{ist} in this economy. First, I assume that the econometrician can observe the full economy, and show that the factor structure of productivity z_{ist} potentially confounds the original "demeaning" approach of Gabaix (2011) in recovering u_{ist}. Second, I assume that the econometrician observes $p_{ist}z_{ist}$ but is unable to distinguish relative price changes from shocks to z_{ist}, which I argue is true in practice. In this case, I show that estimating a factor model as in Stella (2015) on firm level data will also give inconsistent estimates of true firm shocks. I then outline the emprical approach I use in this paper to deal with this problem.

1.3.2 Demeaning approach

The growth of productivity of firm j in sector q is given by:

$$\Delta z_{jqt} = \lambda_{jq}\Delta f_t + \gamma_{jq}\Delta f_{qt} + \Delta u_{jqt}. \tag{1.10}$$

Gabaix (2011) attempts to isolate the aggregate factor by subtracting the (equally-weighted) mean productivity growth rate of (i) all firms, and (ii) all firms in sector q.[6] In this setup that would imply

$$\Delta z_{jqt} - \Delta z_{\bullet\bullet t} = (\lambda_{jq} - \lambda_{\bullet\bullet})\Delta f_t + (\gamma_{jq} - \gamma_{\bullet\bullet})\Delta f_{qt} + \Delta u_{jqt} - \Delta u_{\bullet\bullet t}, \tag{1.11}$$

or, when deducting sectoral averages:

$$\Delta z_{jqt} - \Delta z_{\bullet qt} = (\lambda_{jq} - \lambda_{\bullet q})\Delta f_t + (\gamma_{jq} - \gamma_{\bullet q})\Delta f_{qt} + \Delta u_{jqt} - \Delta u_{\bullet qt}, \tag{1.12}$$

[6] To be precise, Gabaix (2011) uses the sales to employees ratio as a measure of productivity, which I abstract from here.

where $x_{\bullet\bullet t} = \frac{1}{S}\Sigma_{s=1}^{S}\frac{1}{N}\Sigma_{i=1}^{N}x_{ist}$ and $x_{\bullet st} = \frac{1}{N}\Sigma_{i=1}^{N}x_{ist}$. Gabaix takes each of these expressions to be the "firm shock", and then aggregates these to form what he terms the "granular residual":

$$\Gamma_{ALL,t} = \sum_{s}\sum_{i} w_{ist-1}\Delta(z_{ist} - z_{\bullet\bullet t}), \tag{1.13}$$

$$\Gamma_{SIC,t} = \sum_{s}\sum_{i} w_{ist-1}\Delta(z_{ist} - z_{\bullet st}), \tag{1.14}$$

where the weights w_{it-1} are lagged sales scaled by lagged nominal GDP (first suggested by Domar (1961)). Gabaix regresses GDP and TFP growth of the different measures of $\phi(L)\Gamma_t$, where $\phi(L)$ is a lag polynomial, and finds that (i) Γ_t is positively related to growth in these economic aggregates and (ii) between 25-33% of the variation of these aggregates can be explained by this measure of firm shocks (judging by the R^2).

However notice that either measure of Γ_t will be contaminated by a measure of the aggregate shock if for some i, $\lambda_i \neq \lambda_\bullet$. In particular, even if there were no firm shocks, Gabaix's granular residual Γ_t would be positively correlated with total output if larger firms were more sensitive to the aggregate factor than smaller firms (and vice-versa).

1.3.3 Factor model approach

Given the problems with the demeaning approach, Stella (2015) imposes a factor structure to attempt to distinguish aggregate from firm shocks, using the model proposed by Foerster et al. (2011). The use of this approach has a number of advantages, notably in allowing the factor loadings to vary over firms. Stella (2015) reports an impact of firm-level shocks which are substantially smaller than a Gabaix (2011)-like measure applied to his dataset, which casts some doubt on the original findings.

However, this approach is not without problems too. The factor-model based approach is an adaption of Foerster et al. (2011) model of US manufacturing sectors, and is solved by reference to the social planner's problem. While this might be appropriate for disaggregating output into sectors, when applied to firms this is problematic as in a decentralised setting large firms would use their market power to choose different prices/output than would be chosen by a social planner.

Another problem is data limitations: particularly the paucity of data documenting network linkages between firms. Input-output linkages create positive correlation between the outputs of connected firms, and if this information is missing then idiosyncratic shocks could be misclassified as aggregate shocks. Stella (2015) obtains network linkages data from the Compustat customer segments database; but this only records "major customers" (which account for $> 10\%$ of revenue) and hence is likely to understate the true degree of linkages. That said, a robustness exercise where he uses BEA sector-level estimates to calibrate the input-output linkages does seem to generate similar results.

Finally, model misspecification is another potential problem for this structural approach. Atalay (2014) argues that the elasticity of substitution between sectors is much lower than that implied by the Cobb-Douglas production function used in Foerster et al. (2011), which would tend to understate the effect of sectoral shocks. When applying this model to firm-level data, the misspecification could work the other way - it might be the case that firms within the same sector are close substitutes for one another.

In the model presented in this section, a similar problem arises if the consumer's elasticity of substitution between different goods is non-unitary and if firm-specific deflators are not available. Then:

$$\frac{p_{ist} z_{ist}}{P_t} = \frac{p_{ist}}{P_t}(\lambda_{si} f_t + \gamma_{si} f_{st} + u_{ist}) \tag{1.15}$$

and recall (1.3):

$$\frac{p_{ist}}{P_t} = v_{is} \left(\frac{c_{ist}}{C_t} \right)^{-\frac{1}{\varepsilon}}.$$

Suppose that $\varepsilon > 1$, and consider a positive shock to a large firm j in sector q, i.e. $u_{jqt} > 0$. Then by (1.9), $\Delta n_{jqt} > 0$. In general equilibrium, assuming diminishing marginal utility of leisure, the wage will increase which will increase the total labour supply and will "crowd out" the labour used by other firms, lowering their output. In turn, this will lead to a fall in the price of output of the firm hit by the shock relative to that of other firms, i.e. $\Delta \frac{p_{jqt}}{p_{-jqt}} < 0$. This means that while z_{-jqt} will not be affected by u_{jqt}, the *measured* productivity ($p_{-jqt} z_{-jqt}$) of other firms will increase as their relative prices increase. Such crowding out is a feature of more complex models of heterogenous large firms, such as Carvalho and Grassi (2015). If $\varepsilon < 1$ then the contamination will be converse (i.e. the same shock would lead to "crowding in" of other firm's output and lower measured productivity for these firms). To work with firm-level data in the framework of Foerster et al. (2011) we require $\varepsilon = 1$, which is a problem if this is not true in the data.

1.3.4 An external proxy approach

In this section I outline an alternative methodology to recover estimates of granualar shocks, which should ovecome some of the shortcomings of existing methods in the literature. First, I discuss the external proxy approach in principle and how it applies within the framework outlined above. Second, I discuss the particular external proxy used in the paper and why it is plausible that it is both related to firm-level productivity and unrelated to aggregate or sectoral information. Finally, I briefly discuss the strategy for relating firm level shocks recovered with this strategy to aggregate information.

A collection of VARs

Consider a VAR containing the measured productivity growth of firm j and the equally weighted average growth rate of the productivity of all firms:[7]

$$\phi(L) \begin{bmatrix} \Delta(p_{jqt} z_{jqt}) \\ \Delta(p_{\bullet qt} z_{\bullet qt}) \\ \Delta(p_{\bullet\bullet t} z_{\bullet\bullet t}) \end{bmatrix} = \begin{bmatrix} e_{jqt} \\ e_{j\bullet t} \\ e_{\bullet\bullet t} \end{bmatrix}, \tag{1.16}$$

where e_{jqt}, $e_{\bullet qt}$ and $e_{\bullet\bullet t}$ are the reduced form residuals corresponding to the firm, sectoral and aggregate equation respectively. As is well understood in the VAR literature, in general e_{jqt} cannot be used as a measure of structural firm shocks as it is made up of a linear combination of firm and aggregate shocks:

$$\begin{bmatrix} e_{jqt} \\ e_{j\bullet t} \\ e_{\bullet\bullet t} \end{bmatrix} = B \begin{bmatrix} u_{jqt} \\ u_{j\bullet t} \\ u_{\bullet\bullet t} \end{bmatrix}, \tag{1.17}$$

[7] As with the factor model approach, to map the model from section 1.3.1 into a linear VAR framework would require linearising it around a steady state.

where the structural shock in the first equation is the idiosyncratic shock pertaining to the large firm in question (u_{jqt}) while those in the other two equations correspond to sectoral and aggregate shocks. To extract estimates of the u_{jqt} from the VAR we need to know B, and the fact that I am focussing particularly on large firms makes many standard identification assumptions (e.g. a Cholesky decomposition) invalid.

To attempt to deal with these problems, I appeal to the *external proxy* method proposed by Stock and Watson (2012) and Mertens and Ravn (2013). In particular, assume that for each firm I have an m_{ist} such that:

$$\mathbb{E}\left[m_{ist}u_{ist}\right] = \Phi \neq 0, \tag{1.18}$$

and

$$\mathbb{E}\left[m_{ist}u_{jst}\right] = 0 \quad \forall j \neq i, \tag{1.19}$$

$$\mathbb{E}\left[m_{ist}u_{jqt}\right] = 0 \quad \forall j, q \neq s, \tag{1.20}$$

$$\mathbb{E}\left[m_{ist}f_{qt}\right] = 0 \quad \forall q, \tag{1.21}$$

$$\mathbb{E}\left[m_{ist}f_{t}\right] = 0. \tag{1.22}$$

This states that the proxy is correlated with the shocks to the own firm but uncorrelated with shocks to other firms or aggregate shocks. Conditions (1.18) and (1.19)-(1.22) are analagous to the relevance and validity conditions from instrumental variable estimation, though on VAR residuals rather than directly observable variables. As I show in Appendix 1.B, a series m_{it} which satisfies these conditions is sufficient to back out a unique estimate the series of u_{it}.

To be precise, the methodology (the first part of which is analagous to two-stage least squares) proceeds in three steps. First, regress the (potentially confounded) reduced-form residual corresponding to the firm equation on the instrument, pooling by sector.[8] The fitted value from that regression can then be formed:

$$\hat{e}_{ist} = \hat{\beta}_{\bullet s} m_{ist}, \tag{1.23}$$

[8] In principle, the first-stage regression could proceed firm-by-firm. In practice, because of the noisiness of the relationship between m_{ist} and e_{ist} it will be useful to run this regression as a panel by two-digit sector.

where the fitted value is correlated with the firm shock but by (1.19)-(1.22) it is uncorrelated with other shocks. Note that although the parameter estimate $\hat{\beta}_{\bullet s}$ only varies by sector, the fitted value varies at the company level as it depends on the realised values of the high-frequency instrument.

Second, as detailed by Gertler and Karadi (2015), we can then run the "second stage" regressions:

$$e_{\bullet st} = \gamma_0 + \gamma_1 \hat{e}_{ist} + \nu_{1,st}, \tag{1.24}$$

$$e_{\bullet\bullet t} = \mu_0 + \mu_1 \hat{e}_{ist} + \nu_{2,st}, \tag{1.25}$$

Under the assumption that the validity conditions (1.19)-(1.22) hold, then these regressions give an estimate of the responsiveness of sector s's productivity and the overall economy to a shock to the firm's productivity, respectively. Notice that although the first-stage regression is pooled, the fact that this regression is estimated at the firm level allows productivity shocks affecting larger firms to have stronger impacts on sectoral or economy-wide average productivity than smaller firms.

Third, we can use the estimated parameters $\hat{\gamma}$ and $\hat{\mu}$ as additional restrictions which decomposing the variance-covariance matrix of reduced-form residuals on a firm-by-firm basis:

$$\Sigma_{is} = B_{is} B'_{is} \tag{1.26}$$

These estimated parameters are enough to pin down a unique estimate of the first column of B_{is} when combined with the implicit restrictions in Σ_{is}. As detailed in Appendix 1.B, we can then recover a unique estimate of the structural shock associated with firm productivity.

High-frequency stock price movements as an external proxy

The proxy variable I propose in this paper to isolate the effect of firm shocks is the stock price return of the company in question on the day they release their quarterly results. This is defined as:

$$m_{ist} \equiv \widehat{R}_{ist} \equiv R_{ist} - \widehat{\beta}_{0y} - \widehat{\beta}_{1y} R_{IND,t}, \tag{1.27}$$

where R_{ist} is the total stock return (including dividends) of company i in sector s on day t, $R_{IND,t}$ is the total index return on day t, and $(\widehat{\beta}_{0y}, \widehat{\beta}_{1y})$ are the estimated constant and slope coefficients of the ordinary least squares regression of R_{it} on a constant and $R_{IND,t}$ for year y. I drop excess returns corresponding to days on which the $R_{IND,t}$ is more than two standard deviations away from its annual average, as this would suggest that a lot of aggregate information might have been released on that day, which would make the estimate of the firm specific shock less reliable.

There are good *a priori* reasons to believe that excess stock returns over relatively short windows around firm specific announcements might be both related to firm productivity and orthogonal to direct shocks to economic aggregates. First, it has been well documented in the event study finance literature that positive earnings surprises are associated with higher stock returns on the day of release (MacKinlay (1997)). Taking a short window around these results reduces the chance that aggregate news influences both firm stock prices and productivity. I elaborate on the validity and relevance of the proxy in section 1.5.

Relating firm-level shocks to economic aggregates

Once a method has been chosen to recover a series of firm or "granular" shocks, the task is to quantify the effect of such shocks on economic aggregates. To do this, I adopt the same approach as Gabaix (2011) and Stella (2015). These papers appeal to a theorem proposed by Hulten (1978), which states that the marginal contribution of a Hicks-neutral technology shock to one firm on aggregate TFP is scaled by that firm's sales. Aggregating across firms:

$$\frac{dTFP}{TFP} = \sum_i \sum_s w_{ist-1} u_{ist}, \tag{1.28}$$

where the weights for a productivity shock recovered from a production function based on gross output are the same as those proposed by Domar (1961): $w_{ist} = \frac{S_{ist}}{Y_t}$. This result is somewhat surprising in that it implies that if we have a measure of the firm shocks, we don't need to account for input-output linkages across firms to quantify the effect of u_{ist} on aggregate TFP.[9]

As an empirical counterpart to (1.28), consider the following:

$$\Gamma_{\tau,t} = \sum_{i=1}^{N} \sum_{s=1}^{S} w_{ist-1} \hat{u}_{\tau,ist}, \tag{1.29}$$

where $\tau \in \{ALL, SIC, HFI\}$ indexes the type of granular shocks used, where $\tau \in \{ALL, SIC\}$ are the 'demeaned' productivity residuals introduced by Gabaix (2011) as expressed in equations (1.11) and (1.12), while $\tau = HFI$ are the productivity residuals extracted using the external proxy approach, as detailed in section 1.3.4.

I consider two measures of productivity. First, following Gabaix (2011) I use the sales to employee ratio. This measure is transparent and has the advantage that it makes my results more easily comparible with his original paper. For robustness, I also consider a TFP measure following Imrohoroglu and Tüzel (2014). This is based on the value added of the firm, rather than its gross output. In this case, the appropriate weights to use are instead value added over GDP, rather than sales, as noted by Guellec and de la Potterie (2001): $w_{ist} = \frac{V_{ist}}{Y_t}$, where $V_{ist} = S_{ist} - M_{ist}$ (value added $\equiv$ sales less intermediate inputs).

1.4 Data

I use four different types of data in this paper: firm accounts data (to construct measures of firm productivity), stock price data (to construct the high-frequency instrument), analyst earnings forecasts (as a cross-check on the stock price data), and macroeconomic data (to relate the firm shocks to economic aggregates). Details about the first two of these are described briefly below, with a more detailed exposition about how the data were cleaned and otherwise adjusted relegated to the Appendix.

The National Accounts data are taken from the FRED database maintained by the

[9]The intuition behind this result is the following. Consider a 1% increase in a firm's TFP. If firms don't respond to that change by altering their capital and labour choices (or, in a richer model, their intermediate input choices) then the impact of that shock on produced values is indexed by the sales of the firm. However, because firms before the shock were optimising, the envelope theorem implies that this is also the increase in total TFP when we allow firms to alter their choices).

Federal Reserve Bank of St Louis and is originally compiled by the Bureau of Economic Analysis, while the measure of TFP growth that I use is taken from the San Francisco Federal Reserve Bank.[10],[11]

1.4.1 Firm productivity data

Firm level accounts data are taken from the Compustat Quarterly database. I follow Gabaix (2011) and take firms which were in the top 1000 by sales in each quarter from 1963 to 2015 which were not one of (i) oil firms; (ii) non-oil energy firms; or (iii) financial firms. This is the same as Gabaix's choice of sample except that I also exclude conglomerates with a large financial component, notably General Electric and Berkshire Hathaway.[12] This yields a baseline unbalanced panel of 3337 firms.

My baseline measure of firm productivity is the sales to employee ratio, following Gabaix (2011). This allows my work to be more closely compared with his results. Unlike Gabaix (2011), this paper has to work with quarterly data to take full advantage of information in quarterly earnings releases. Details of how I clean and seasonally adjust the data are relegated to Appendix 1.A.1.

For robustness, I construct a measure of firm TFP. I follow Imrohoroglu and Tüzel (2014) in adapting the method of Olley and Pakes (1996) for company accounts recorded from COMPUSTAT. Firm TFP is defined as:

$$TFP_{it} = log(y_{it}) - \beta_0 - \beta_k log(K_{it}) - \beta_l log(L_{it}),$$

Where y_{it}, K_{it} and L_{it} are value added; labour input (defined as the number of workers) and capital input in real terms respectively. I follow Imrohoroglu and Tüzel (2014) closely in the construction of this variable: value added is taken to be the sum of operating income before depreciation and a proxy of labour costs. Capital and labour are only

[10]The GDP series used is code GDPC1

[11]The methodology is discussed in Fernald et al. (2012) and the data are available from `http://www.frbsf.org/economic-research/indicators-data/total-factor-productivity-tfp/`

[12]Even ignoring the financial component of such firms, the fact that they operate in a number of unrelated business areas makes estimating these firms' TFP based on a single production function problematic.

available annually: as such, I interpolate these series with a spline.

1.4.2 Stock price data

Stock market data is obtained from the Compustat/CRSP linked data. Excess returns are estimated on an annual rolling basis, as the residual of the daily stock price return for company i and the return on the S&P 500 on that day.

The principal measure of "firm news" that I use is this excess stock return on the day on which a firm releases its quarterly results. This day varies by firm, but it is typically within a month of the end of a firm's financial quarter. These days are identified by combining information from the Compustat Quarterly database with Institutional Broker Estimate System (I/B/E/S). The latter database has the advantage that it contains the time of release as well as the day, so I am able to attribute results information released after market close (16:00 EST) to stock changes on the following day (unfortunately the I/B/E/S database has missing datapoints, so it is necessary to combine it with the information from Compustat).

Market-relevant information about firms may also become public (and incorporated into market prices) on days other than those which coincide with quarterly earnings releases. To try and capture some of this information, I compile filings of Form 8-k from the Securities and Exchange Commission's EDGAR database. The SEC requires listed firms to file a Form 8-k if there is a "material corporate event" which shareholders should know about. I record the dates (and times, where available) of Form 8-k filed under Sections 7, 8 and 9, which correspond to news plausibly related to firm productivity (I drop filings on days which coincide with quarterly results days, to avoid double counting). I exclude Form 8-k filed under other sections, as these relate to information such as changes in senior management or the firm's accountant (for instance) which might be market-moving but are unlikely to be related to firm productivity.

Figure 1.1 shows the average standard deviation of firm excess stock returns ($\widehat{R}_{it}$) in a twenty-business-day window either side of the release of a firm event: either a quarterly earnings report or the filing of a form 8-k with the SEC. The volatility of excess stock returns increases substantially on the day that quarterly results are released. On the month leading up to a results day, daily volatility in excess returns (averaged across all firms) is lower than 2%: on days on which quarterly results are released, this jumps to over 3.5% before falling again. Two days after the results day, the excess return volatility falls again to its pre-event day average. A similar pattern, if less dramatic, is observable using excess return volatility around 8-k days. In this case, the "spike" on the day of the news is smaller, but the effect appears to be more persistent, though this is consistent with the fact that 8-k events are not known in advance and are irregular, which may mean it takes longer for market participants to fully assess their effects.

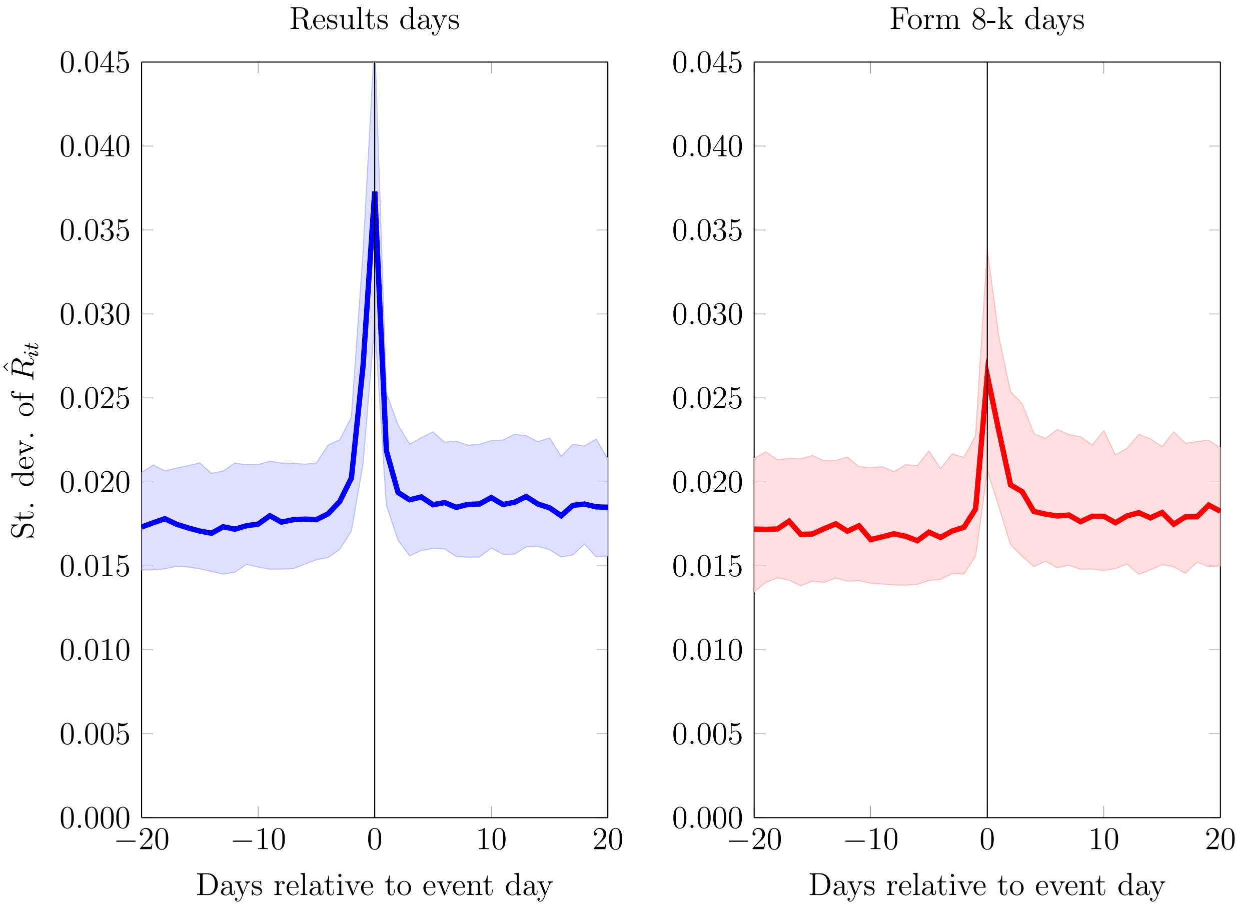

Figure 1.1: Excess return volatility increases on results days
Notes: Figures show the median standard deviation of excess returns of stocks around days on which quarterly results are released (left-hand panel) and days on which a company files a form 8-k (section 7-9) with the Securities and Exchange Commission. Shaded areas correspond to the 32/68 percentiles of the distribution.

1.4.3 Analyst Recommendations

To evaluate the high-frequency instrument, I use information based on analyst forecasts from the Thomson-Reuters I/B/E/S detail database. In particular, I construct two measures, relating to "surprises" to earnings in the current quarter (assessed by the difference between the final analyst consensus forecast and the realised value) and surprises in one quarter out (assessed by revisions to the analyst consensus before and after results days).

The analyst-based external instrument has one principal advantage over the stock-price based measure described above: the ability to distinguish surprises which affect the quarter to which the results pertain (the previous financial quarter) and surprises to expectations in the next quarter (that is, the quarter during which the results day occurs). Against this advantage is the fact that analyst forecasts are typically made well in advance of results days, which increases the risk that some of the "surprise" reflects aggregate, rather than firm-level shocks.

To form my first measure, I take an average of the earnings per share forecasts of analysts within the fourteen days ahead of the results day. I then subtract the realised earnings per share value from the I/B/E/S database and scale the result by the stock price on the day before the results were released. The second measure is similar: I take the average earning per share forecasts for the next quarter from a two week window prior to the results day, and subtract the average earnings per share forecast of all forecasts made in the five days after the results day (forecasts recorded on the day of the results are ignored), and scale by the same stock price used for the first measure.

To assess the relationship between the size of the earnings surprise and the excess stock return I conduct a similar exercise to MacKinlay (1997). I divide earnings surprises (scaled by stock prices) into three groups: those which are more than one standard deviation above the mean surprise ("good news"), more than one standard deviation below the mean surprise ("bad news"), and all others ("no news").[13] The cumulative

[13] I drop very large earnings surprises (absolute value of 0.5 or greater) to stop them distorting the results. The standard deviation is calculated by taking the standard deviation of forecast errors by quarter, and averaging over quarters. The mean surprise is slightly positive. This is consistent with some evidence in the finance literature that earnings or expecations are manipulated by companies so that they on average beat the average analyst's estimate, see Terry (2015). This is another reason to be wary of earnings surprise relative to the analyst consensus as a measure of the firm shock.

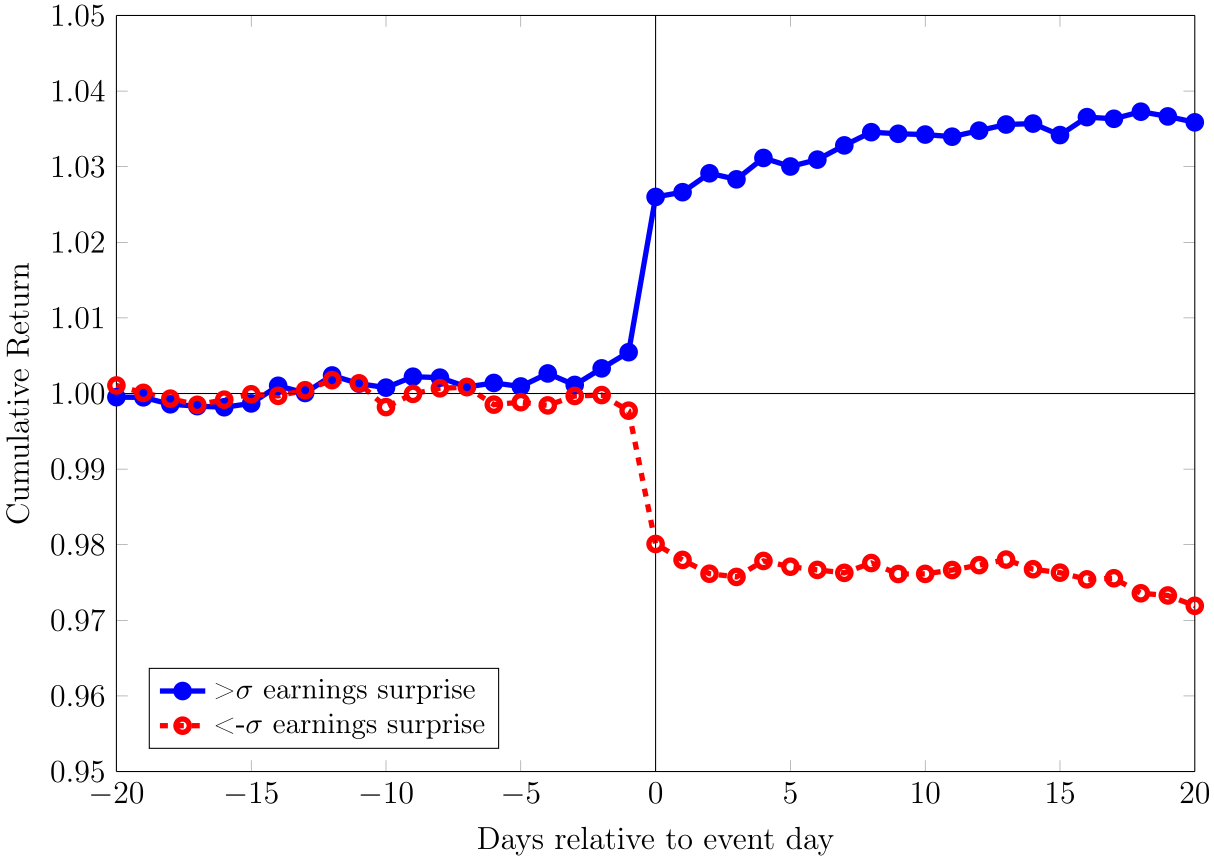

Figure 1.2: Mean cumul. stock returns of firms by results relative to consensus
Notes: Figure shows the cumulative stock returns of two groups of companies: (i) those which report earnings at least one standard deviation higher than the median analyst expectation (as captured by the Thomson Reuters I/B/E/S database) and (ii) those which report earnings one standard deviation lower than this consensus expectation.

stock returns for the three groups of firms in a forty-one day window around earnings releases are shown in Figure 1.2. On average, firms which receive "good news" outperform "no news firms" by 2% over the window, with most of the gain coming on the day on which results are released. "Bad news" firms underperform by a similar amount, again, with most of the loss coming on the day on which results are released.

However, these averages hide quite a lot of noise. While cumulative returns for "good news" firms outperform "bad news" firms on average, there is substantial overlap. This noisiness weakens the high frequency instrument, and as such pooling firms by sector is necessary to find relationships of sufficient strength such that weak instrument concerns are mitigated.

1.5 Firm-level results

This section documents the relationship between firm-level productivity measures and stock price surprises. I first discuss how closely my external proxies are related to the residuals of interest (i.e. the reduced-form residuals from the VAR), which is the counterpart to the "relevance" condition for instrumental variables. I then turn to a discussion of the validity of my proxy - i.e. that it is uncorrelated with other structural shocks of interest (in particular, sectoral and aggregate productivity shocks). While this assumption is not directly testable, I present suggestive evidence using both the differential timing of firm results days within a quarter and using the panel nature of my dataset that stock price movements on firm results days are not systematically related to aggregate information.

1.5.1 Relevance of the proxy

As in standard instrumental variable regression, the results of the procedure adopted here are only meaningful if the external proxy is sufficiently highly correlated with the reduced form residual of interest. As detailed in equation (1.23), I pool the first-stage regressions by (two-digit) sector. While it is possible to run these regressions at the firm level, the resulting relationships are noisy (as measured by the first-stage F statistic). This is not surprising: a component of any firm productivity innovation might be anticipated by investors but not by the VAR used here (investors form their expectation about firm productivity using a much wider range of information than simply lagged productivity of the firm or sector). This problem is compounded for firms with shorter time series (due to exit or entry). To deal with this, pooling at the sector level increases the sample size making it easier to identify the true relationship, at the cost of assuming that the relationship between stock returns on results days and the firm-specific productivity shock to be common across firms in the same sector. *A priori* it does not seem unreasonable that a 1% increase in firm productivity due to a firm level event is associated with the same increase in the firm's stock price independent (e.g.) of the size of the firm or other covariates.

After discarding small sectors (those with fewer than 300 firm-quarter observations), the first-stage regression results indicate that 11 of 53 two-digit sectors covering 47.9% of firms in my sample have an F-statistic of 10 or greater stipulated as the "rule of thumb" by Staiger and Stock (1997) using the sales: employees measure of productivity favoured by Gabaix (2011). The relationship is somewhat stronger when we use the measure of firm TFP based on the methodology of Imrohoroglu and Tüzel (2014): here 26 of 49 sectors covering some 81.4% of firms meet this criterion. As not all of the sectors meet the F$\geq$10 rule of thumb, I construct two different measures of the granular residual.The first measure includes all firms, while the second excludes firms in sectors which have a first-stage regression with an F statistic less than 10. I present results of both below: where they differ I prefer the latter.

1.5.2 Validity of the proxy

The other key assumption needed for identification is that the external proxy is uncorrelated with structural shocks other than the structural shock of interest. Here, taking stock returns over a relatively short window (one day) minimises the chance that news pertaining to economic aggregates is released on that day which affects asset prices. While it is true that shorter windows are often used in investigations of monetary policy shocks, I do not follow this practice for two reasons. First, most (but not all) firm results are released outside of market hours. Taking the market close to market close change in stock prices standardises the measure I am taking over all firms. Second, firms typically follow up the release of their earnings numbers with an analyst call to elaborate on the results - for example, to explain whether a surprise in the profit numbers was truly the consequence of a change in productivity or due to some accounting change. By taking a short intraday window around results I would risk throwing away stock price changes resulting from this information.

Another concern about the proxy could be that earnings reports themselves contain aggregate information. One might expect that this is unlikely *a priori* as company results are released after a large amount of higher frequency, more aggregated information has been released relating to the quarter in question (for example, industrial production, retail sales and labour market statistics are all released on a monthly basis, and around half of companies release their results after the first estimate of GDP is released by the

Bureau of Economic Analysis). As such, it is sensible to think about aggregate growth as being known, but the distribution (or source) of that growth being revealed in full when company earning season has completed.

However, we can attempt to address this concern - indirectly - with the data. For most firms, financial quarters coincide with calendar quarters, but the release of their earnings information are staggered throughout the following quarter. The date of release is normally chosen a year or more in advance and the timing is relatively persistent across quarters. If firm earnings reports do systematically release information about aggregate or sectoral productivity, then early-reporting firms would reveal more information to the market than late-reporting firms (as the former would release both idiosyncratic and aggregate information, while the latter would only release idiosyncratic information). If so, then we should see two features of the stock price data. First, early-reporting firms should see a larger increase in the daily volatility of R_{it} on its results day than late-reporting firms, and second, early-reporting firms should see a smaller fall in the rolling correlation between their stock price return on results days and the stock index (the S&P 500) than late reporting firms.

To get a sense if this is happening, I do the following. First, I drop firms for which financial quarters do not coincide with calendar quarters. I then examine the remaining firms by three-digit SIC sector. For each sector and quarter, I include a firm-quarter in an "early" sub-sample if it is one of the first two firms in that sector and quarter to report, and I include a firm-quarter in a "late" sub-sample if it is one of the last two firms in a quarter to report (I ignore sector-quarters with fewer than four firms reporting). Figure 1.3 shows the rolling excess return volatility for these two groups. The first panel of the chart makes clear that there is no discernable difference in the percentage increase in return volatility of early- or late- reporting companies, which is consistent with the idea that aggregate information is not revealed by a subset of companies. The bands shown correspond to the 32/68 percentiles respectively - in each case the point estimates for either sub-sample of companies lie within the percentile bands shown of the other, which indicates that any difference is not statistically significant. The second panel shows the rolling correlation between the total return of the two pairs of companies and the S&P 500: again. The fall in this rolling correlations on the results day indicates that firm-specific information is being released, and the fact that the correlations are not statistically distinguishable suggests that firms are not releasing aggregate information systematically when they release their quarterly results.

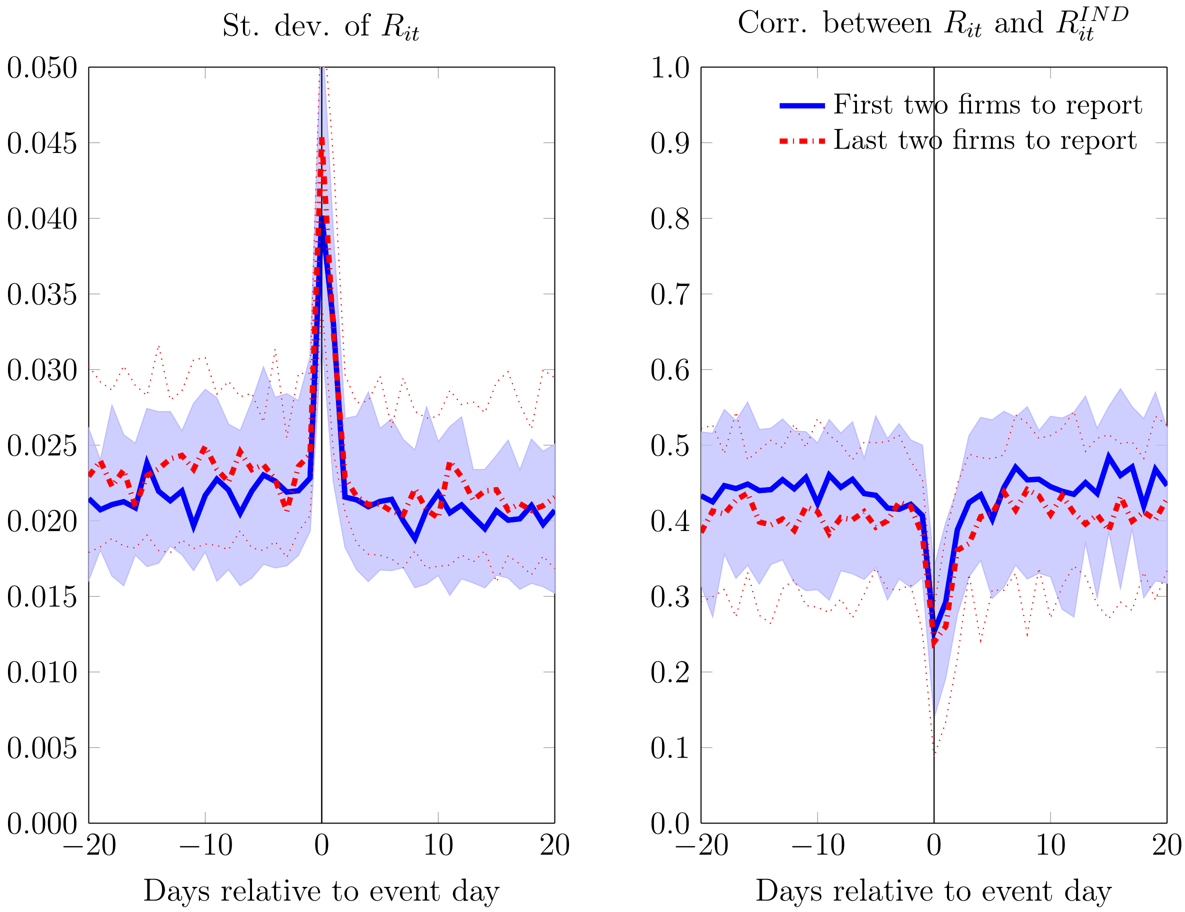

Figure 1.3: Behaviour of stock returns of early vs. late reporters by sector around events
Notes: Contrast of stock returns of companies around days on which quarterly results are released. Companies divided into two sub-samples: those which are among the first two to report their results in a quarter and those which are among the last two to report in a quarter (within the same three-digit sector). Left panel shows the standard deviation of total returns and the right panel shows rolling correlation with S&P 500 of each sub-sample. Shaded area/dotted interval are the 32/68 percentile of the distribution of each sub-sample.

1.6 Aggregate results

In this section I contrast the properties of the granular residual constructed with the "demeaning" method proposed by Gabaix (2011) with the "high-frequency" approach proposed in this paper.

First, I perform a similar exercise to Gabaix (2011) in attempting to quantify the importance of the granular residual on economic aggregates, namely GDP and TFP growth. I find that the granular residual calculated with the "demeaning" method substantially overstates the importance of the granular shock relative to the estimates associated with the "high-frequency" measure.

To investigate this potential discrepancy, I investigate the possibility that "demeaning" does not remove aggregate shocks perfectly as larger firms respond more strongly to aggregate shocks. I show that certain macroeconomic aggregates appear to Granger cause the "demeaned" granular residuals, which is consistent with the notion that these measures are not completely clean of aggregate shocks. I also verify that the same is not true for the method of constructing the granular residual based on high-frequency financial information.

1.6.1 Comparison with existing estimates

As a first exercise, I recreate the method that Gabaix (2011) uses to quantify the importance of the demeaned granular residual. To stay as close as possible to Gabaix's results, I take the quarterly measures of productivity (i.e. either the "demeaned" measures or the measure recovered by the high-frequency instrument), use each to construct an index of productivity shocks and then aggregate this series by year. Then, on an annual basis, I regress GDP or TFP growth on a constant and the contemporaneous and two lagged values of the demeaned granular residuals. This gives a sample from 1968-2014.[14] The results are shown in Tables 1.1 and 1.2 for GDP and TFP growth, respectively.

Each table has four columns, each relating to a different measure of firm shock. The first two columns use a measure of the granular residual related to the measures of firm shocks proposed by Gabaix (2011): the first refers to the growth rate of log(sales: employees) less the average growth rate of that measure for all firms in the sample (corresponding to equation (1.11)), while the second is same measure but instead demeaning at the three-digit sectoral level (corresponding to equation (1.12)). The R^2 in each case suggests that a substantial fraction of GDP and TFP are explained by these measures of the granular residual: 17% (15%) in the case of demeaning by all large firms for GDP (and TFP), and

[14]The qualitative conclusions of this section do not change markedly if I use the top 1000 firms

	(1)	(2)	(3)	(4)
Const.	0.023***	0.019***	0.027***	0.027***
	(0.003)	(0.003)	(0.003)	(0.003)
Γ_t	1.141	4.082**	-1.315	-1.872
	(0.860)	(1.533)	(0.649)	(1.073)
Γ_{t-1}	1.975**	3.780***	0.341	1.259
	(0.843)	(1.378)	(0.618)	(1.024)
Γ_{t-2}	0.577	1.847	0.646	0.985
	(0.849)	(1.476)	(0.642)	(1.028)
T	46	46	46	46
R^2	0.170	0.307	0.143	0.127
R^2 adj.	0.089	0.239	0.060	0.042
Γ def.	Γ_{ALL}	Γ_{SIC}	Γ_{HFI}	$\Gamma_{HFI\|F>10}$

Table 1.1: Regression of GDP growth on different measures of Granular Residual derived from firm sales:employees ratio

	(1)	(2)	(3)	(4)
Const.	0.006**	0.005	0.009***	0.009***
	(0.003)	(0.003)	(0.002)	(0.002)
Γ_t	1.685**	3.408**	-0.699	-1.275
	(0.690)	(1.315)	(0.524)	(0.863)
Γ_{t-1}	0.636	1.288	0.461	1.001
	(0.676)	(1.182)	(0.499)	(0.824)
Γ_{t-2}	0.336	0.598	0.636	0.694
	(0.681)	(1.266)	(0.518)	(0.827)
T	46	46	46	46
R^2	0.152	0.191	0.115	0.103
R^2 adj.	0.070	0.112	0.029	0.016
Γ def.	Γ_{ALL}	Γ_{SIC}	Γ_{HFI}	$\Gamma_{HFI\|F>10}$

Table 1.2: Regression of aggregate TFP growth on different measures of Granular Residual derived from firm sales:employees ratio

31% (and 19%) in case of demeaning by the average in the same three digit sector alone. These results are somewhat lower than those presented in Gabaix (2011), but this can be largely explained by a shorter sample (Gabaix (2011) presents results using a sample from 1950-2008, the longer time series made possible by use of annual data).[15]

The third and fourth column in each table construct the granular residual using measures of firm shocks using the infomation from firm stock price movements as outlined in Section 1.3.4. In this case, the fraction of GDP and TFP explained by this measure of firm shocks is noticeably smaller - at 14% and 13% for GDP, and 12% and 10% for TFP, respectively). These are noticeably smaller than the measures calculated using the "demeaning" methodology. More concerning still is that on impact, the point estimate of the effect of granular shocks is wrongly signed, i.e. a positive granular shock is associated with lower contemporaneous GDP and aggregate TFP (though these point estimates are not significantly different from zero).

As a robustness check of these results, I also consider a measure of firm shocks constructed from a value-added measure of TFP as outlined in Section 1.4.1. These results are shown in Tables 1.3 and 1.4, respectively. Here the R^2 values are considerably lower for both the "demeaned" and "high-frequency" granular residuals - around 0.12-0.15 when the regressand is GDP, and less than 0.1 when the regressand is (aggregate) TFP. Unlike with the sales-based measure there is no noticable fall in the fraction of the variance statistically explained with the "high-frequency" measure relative to the "demeaned" measure.

A visual inspection of a four-quarter moving average of the three sales-based measures of the granular residual can help shed light on these results, as shown in Figures 1.4, 1.5 and 1.6. Both the "demeaned" granular residuals and the "high-frequency" measure proposed in this paper turn negative during the 2001 recession and 2008 recessions, consistent with the "granular hypothesis". The 2008 recession is particularly interesting - most narrative accounts of the crisis associate it with problems in the financial sector but the large negative granular residual (a large part of which is explained by the poor performance of car manufacturers) suggests that there was a "granular" component to this recession as well.[16] However, the two measures of granular residuals differ for earlier recessions. There

[15]My sample is over the period 1968-2015. The Compustat quarterly database starts in 1961 and is initially thinly populated.

[16]Recall that the granular residual is constructed after dropping financial firms from the sample.

	(1)	(2)	(3)	(4)
Const.	0.026***	0.026***	0.027***	0.027***
	(0.003)	(0.003)	(0.003)	(0.003)
Γ_t	-0.031	-0.506	-0.317	-1.218
	(1.329)	(2.008)	(0.209)	(0.648)
Γ_{t-1}	1.617	2.608	0.084	0.701
	(1.201)	(1.768)	(0.228)	(0.695)
Γ_{t-2}	2.442*	2.664	0.317	0.886
	(1.305)	(1.910)	(0.208)	(0.645)
T	46	46	46	46
R^2	0.129	0.119	0.121	0.145
R^2 adj.	0.044	0.033	0.036	0.061
Γ def.	Γ_{ALL}	Γ_{SIC}	Γ_{HFI}	$\Gamma_{HFI\|F>10}$

Table 1.3: Regression of GDP growth on different measures of Granular Residual derived from firm TFP

	(1)	(2)	(3)	(4)
Const.	0.007***	0.008***	0.009***	0.009***
	(0.002)	(0.003)	(0.002)	(0.002)
Γ_t	1.151	0.171	-0.222	-0.750
	(1.075)	(1.655)	(0.169)	(0.533)
Γ_{t-1}	1.069	1.716	0.063	0.455
	(0.971)	(1.458)	(0.184)	(0.572)
Γ_{t-2}	1.708	1.012	0.223	0.483
	(1.055)	(1.575)	(0.168)	(0.531)
T	46	46	46	46
R^2	0.097	0.050	0.095	0.081
R^2 adj.	0.008	-0.042	0.007	-0.009
Γ def.	Γ_{ALL}	Γ_{SIC}	Γ_{HFI}	$\Gamma_{HFI\|F>10}$

Table 1.4: Regression of aggregate TFP growth on different measures of Granular Residual derived from firm TFP

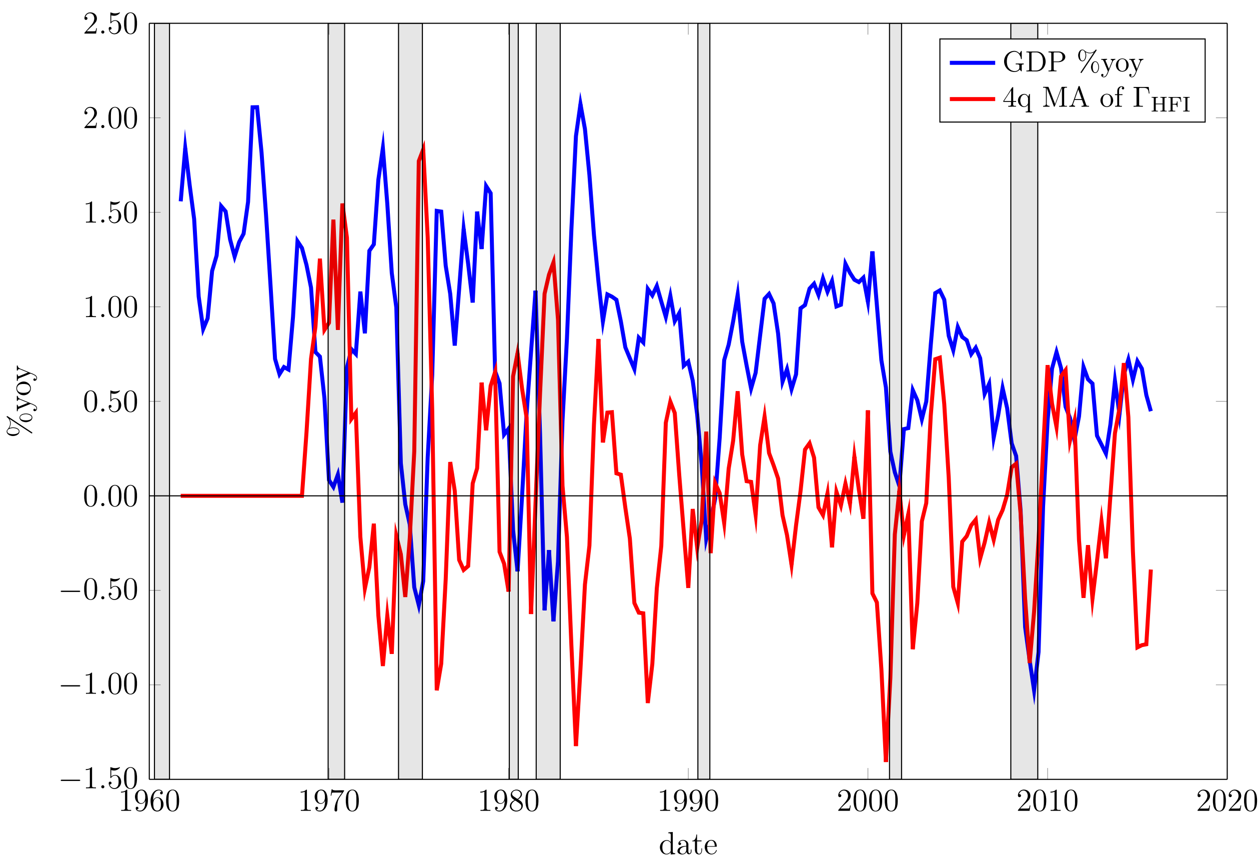

Figure 1.4: Year-on-year GDP growth vs. 4q moving average of "high-frequency" Granular residual (sales: employee measure)

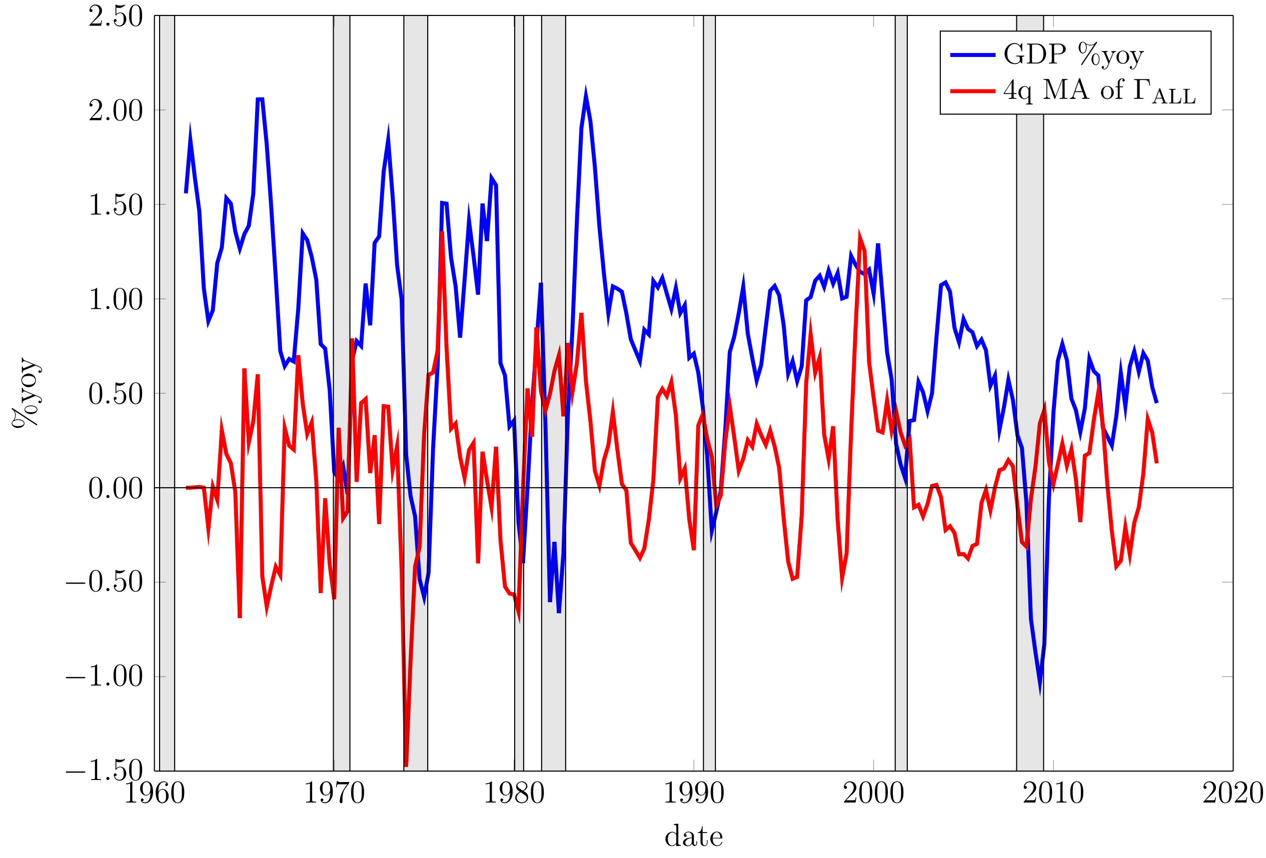

Figure 1.5: Year-on-year GDP growth vs. 4q moving average of "all" demeaned Granular residual (sales: employee measure)

were two NBER recessions at the start of the 1980s, and in this period the "demeaned" granular residuals are procyclical (and consistent with the granular hypothesis) while the "high-frequency" granular residual is countercyclical. This negative correlation is consistent with the negative point estimate of the (contemporaneous) "high-frequency" granular residual in columns 3-4 in Tables 1.1 and 1.2.

In sum, the formulation of the granular residual proposed by Gabaix (2011) suggests that granular shocks are somewhat more important in explaining aggregate flucutations than the measure introduced in this paper, using high-frequency stock price movements as an external proxy.

1.6.2 Granger Causality tests

In section 1.3.2, I argue that a potential weakness of the "demeaning method" of Gabaix (2011) is the fact that subtracting the mean productivity growth (of either the whole

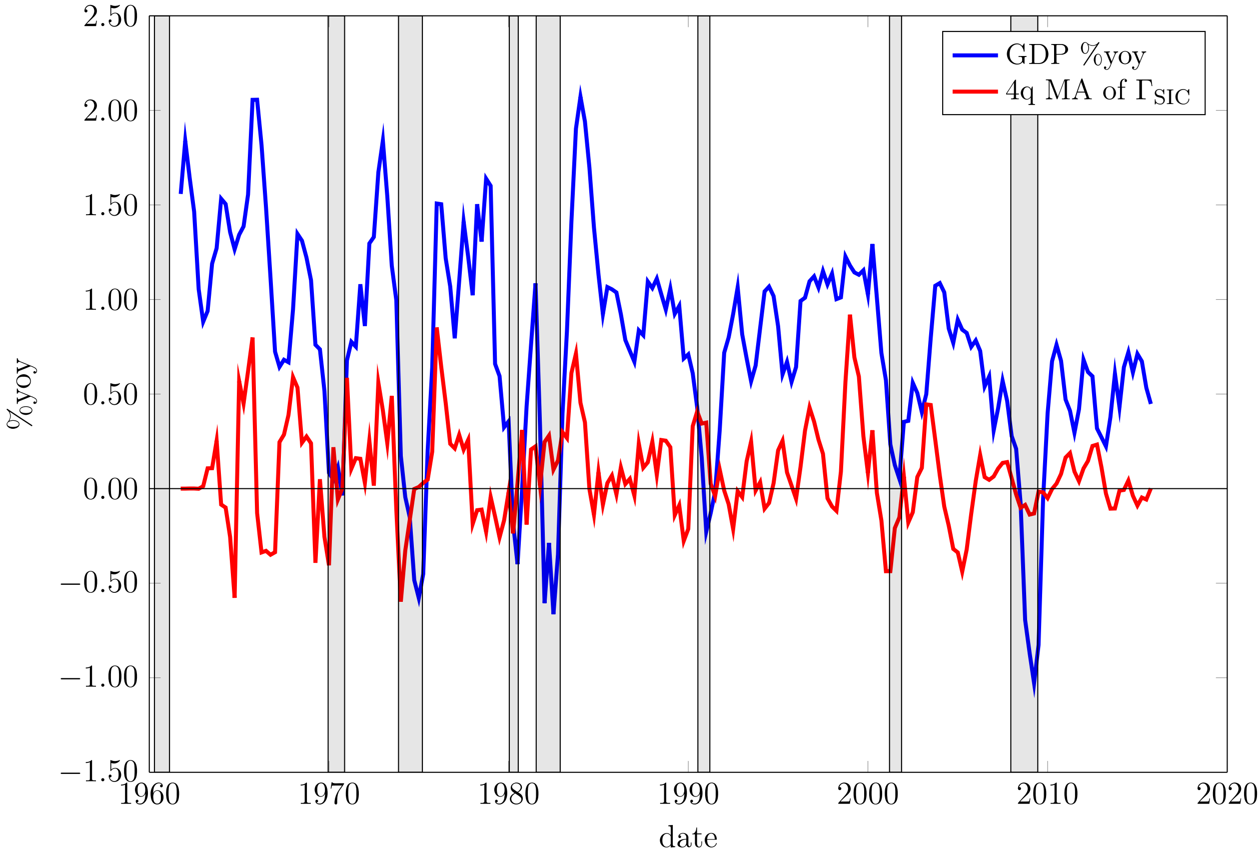

Figure 1.6: Year-on-year GDP growth vs. 4q moving average of "sector" demeaned Granular residual (sales:employee measure)

economy or a specific sector) will only remove aggregate or sectoral processes if firms respond to such processes symmetrically. If the productivity process is composed of a firm, sectoral and aggregate process as in (1.10), namely:

$$\Delta z_{ist} = \lambda_{is}\Delta f_t + \gamma_{is}\Delta f_{st} + \Delta u_{ist},$$

then subtracting the mean growth rate of all firms will only remove the aggregate factor if $\lambda_{is} = \bar{\lambda}\ \forall i, s$ and subtracting the mean growth rate of a sector s from firms in that sector will only remove the sectoral factor if $\gamma_{is} = \bar{\gamma_s}\ \forall i \in s$.

To test if this violation might matter in practice, I conduct the following Granger causality tests. First, I estimate regressions of the following form using quarterly data:

$$\Gamma_{\tau,t} = c + \sum_{j=1}^{P} \beta_j x_{t-j} + \sum_{j=1}^{P} \theta_j \Gamma_{\tau,t-j} + \varepsilon_t, \tag{1.30}$$

where the regressand is a measure of the granular residual of type $\tau \in \{ALL, SIC, HFI\}$, x_t is the variable of interest, where the first two measures are the economy-wide and sectoral demeaned granular residuals as introduced in Gabaix (2011) and the final type refers to the high-frequency instrument for firms in sectors which have a first-stage F-statistic of at least 10 (i.e. these measures correspond to columns 1, 2 and 4 in Tables 1.1 and 1.2).

I then conduct a standard Wald test of the hypothesis that the β coefficients are jointly equal to zero. If the measure of Γ used is truly an exogenous shock process then we would not expect that it should be granger-caused by macroeconomic variables. In contrast, if $\lambda_{is} \neq \bar{\lambda}$ (for instance) and if the aggregate process is persistent then we would expect that some lagged macroeconomic variables could Granger cause the demeaned granular residual.

The variables of interest I consider are GDP growth, the growth in the real price of oil and the tightness of Credit Standards as measured by the Federal Reserve Board's survey of Senior Loan Officers.[17] The rationale for including GDP and the real price

[17]Data for oil prices are taken from Kilian (2009) and deflated with the CPI. The Senior Loan Officer Survey Question is the "Net Percentage of Domestic Banks Tightening Standards for Commercial and Industrial Loans to Large and Middle-Market Firms". This measure refers to firms with sales of over \$50m annually, a condition which all of the firms in my sample satisfy.

of oil is that some of the largest firms in the US economy happen to be in sectors which might be both more cyclical than the average and more sensitive to energy prices and aggregate economic activity - notably large manufacturers such as General Motors, Ford and Hewlett-Packard. If such firms are disproportionately represented among the largest firms in the US, then we might expect aggregate shocks to (say) energy prices to disproportionately affect very large firms, which would generate a correlation between the demeaned granular residual and aggregate outcomes independent of the correlation resulting from true firm-level shocks. The rationale for including a measure of credit tightness is that this is an aggregate shock which larger firms may be more able to deal with than smaller firms. All of the firms in the sample are listed on equity markets, so they are not likely to be financially constrained in the same way that an unlisted firm might be. However, listed firms do borrow from banks (either long-term or for financing working capital) and we might expect smaller listed firms to be more reliant on bank borrowing than the very largest firms in the sample. If so, then an aggregate shock to credit standards would affect the demeaned granular residual.

The results of this exercise are shown in Table 1.5.[18] The first three specifications test the joint significance of lagged GDP growth and lagged growth in oil prices for predicting the various measures of the granular hypothesis. The first column shows that these variables appear to jointly cause the "economy-wide" demeaned granular residual $\Gamma_{ALL,t}$, with a p-value of less than 0.05.[19] However, the second column indicates that for these variables, this problem is largely dealt with if we use the sectorally-demeaned measure of the granular shock. Together, these results are consistent with the intuition that for oil prices and GDP growth the largest firms are disproportionately in sectors which are sensitive to energy and real activity, which cast doubt on whether the "economy-wide" demeaned granular residual is clean of non-firm specific disturbances. The third column verifies that the "high-frequency" granular residual introduced in this paper is not granger caused by these variables either.

In specifications (4)-(6), the regressand is credit standards as measured through the Senior Loan Officers survey. In specification (4), there is not enough evidence to conclude

[18] I use four lags in each specification

[19] One possible objection to this finding is that oil market traders could anticipate shocks to large firms, and oil prices adjust accordingly. However, the coefficients of the unrestricted model suggest that higher oil prices precede lower values of the granular residual, which is inconsistent with this interpretation and consistent with the hypothesis that oil prices are reflecting some aggregate factor.

	(1)	(2)	(3)	(4)	(5)	(6)
x	$dGDP, dp_{oil}$	$dGDP, dp_{oil}$	$dGDP, dp_{oil}$	Cred Stds	Cred Stds	Cred Stds
y	Γ_{ALL}	Γ_{SIC}	$\Gamma_{HFI\|F>10}$	Γ_{ALL}	Γ_{SIC}	$\Gamma_{HFI\|F>10}$
F	2.136	1.009	0.813	0.398	2.589	1.237
pval	0.036	0.431	0.592	0.809	0.042	0.301
T	163	163	163	99	99	99
Start	1974Q2	1974Q2	1974Q2	1990Q2	1990Q2	1990Q2
End	2015Q4	2015Q4	2015Q4	2015Q4	2015Q4	2015Q4

Table 1.5: Granger causality tests: Granular Residuals derived from sales: employees on different economic variables

at the 10% level of significance that the "economy-wide" granular residual is Granger caused by credit standards. However, credit standards do appear to granger cause the "sectoral" demeaned granular residual at the 5% level of significance. This suggests that an aggregate credit shock would affect small and large firms differently within the same sector, which indicates that this measure of firm-level shocks too might contain aggregate components. Finally, the sixth column verifies that we cannot reject the null hypothesis that the "high frequency" granular residual is not Granger caused by this measure of credit standards either.

To sum up, if any measure of the granular residual contains aggregate or sectoral shocks, then any correlation between this measure and economic aggregates might overstate the importance of firm shocks. One reason that the "demeaned" granular residual might not completely clean productivity growth rates of aggregate or sectoral shocks is that large firms respond differentially to smaller firms to such shocks, a problem which should not in principle apply to the "high-frequency" granular residual introduced in this paper. This section presents empirical evidence that certain macroeconomic variables which we might expect to reflect shocks which *a priori* affect large and small firms differently do appear to granger cause the demeaned granular residual (but not the "high-frequency" counterpart), which is consistent with this interpretation.

1.7 Conclusion

One possible microfoundation for total factor productivity (TFP) shocks would be shocks to individual large firms. Using a simple model, I argue that existing empirical approaches of extracting such shocks have important shortcomings - either they abstract from the possibility that the effect of aggregate shocks on individual firms can vary over firms or - in the case of the factor model literature - they are constrained by poor data on intra-firm linkages and strong structural assumptions needed to keep the model tractable. I propose a new method to recover estimates of such shocks, using the variation in stock returns around the days on which companies annouce results. I claim this variation is related to firm-specific information but unrelated to aggregate information. While I find that large firms do appear to drive aggregate fluctuations in GDP and TFP growth, quantitatively my estimates suggest a much smaller fraction of aggregate fluctuations are driven by these factors than the original quantifation exercise conducted by Gabaix (2011).

Appendices

1.A Data Appendix

1.A.1 Construction of firm-level TFP

The construction of firm level TFP on a quarterly basis is based on the methodology proposed by Imrohoroglu and Tüzel (2014). This section describes how I apply their method to my data: the principal difference with their study is that I attempt to construct a quarterly measure of TFP rather than an annual measure. I proceed in three steps. First, I outline the construction of the measure of value added at the firm level, including cleaning the data and seasonal adjustment. Second, I outline how I construct a measure of labour input on a quarterly basis, before explaining construction of the capital input.

Construction of Value Added

As Compustat accounts do not directly report intermediate inputs, it is only possible to construct a proxy of value added. Following Imrohoroglu and Tüzel (2014), the proxy of firm level value added I use is:

$$VA_{it} = OIBDPQ_{it} + LabourCost_{it} \tag{1.31}$$

Where OIBDPQ is operating income before depreciation from Compustat. By definition,

$$OIBDPQ_{it} = SALEQ_{it} - \underbrace{(COGSQ_{it} - XSGAQ_{it})}_{XOPRQ_{it}}$$

XOPRQ is operating expenses, the sum of COGSQ (the cost of goods sold) and XSGAQ (selling and general adminstrative expenses). Intuitively, the two terms in the expression for value added correspond to payments to capital and labour respectively.

The Compustat Quarterly database has a number of data anomalies which are not present in the annual database. To clean the OIBDPQ, SALEQ and XOPRQ series, I proceed as follows:

1. Any part of the time series which has the form [...NXN...], where X is data and N

is missing data, is transformed to [...NNN...]. This is because some annual values are erroneously stored in the Compustat Quarterly database at the financial year end.

2. Any part of a company time series of the form [...NXX...] is transformed to [...NNX...]. This is because a number of missing quarterly observations' data are combined into the datapoint for the following quarter.

3. For both SALEQ and XOPRQ I identify "unusual" one-off spikes in the following way: first, I calculate the quarterly growth rate, and discard the top and bottom 5% of my sample. I then calculate the standard devation of the remaining sample. I then identify quarters with "exceptional" sales/costs in the following way: if a quarter has either a postive growth rate followed by a negative growth rate, or a negative growth rate followed by a postive growth rate which are both more than three times this "trimmed" standard deviation, that quarter is treated as exceptional. I discard such quarters.[20]

I proxy labour cost as:

$$LabourCost_{it} = EMP_{it} * wage_t$$

where total employment EMP_{it} is defined below, and $wage_t$ is the average wage provided by the Bureau of Labour Statistics. This is the same proxy as used by Imrohoroglu and Tüzel (2014). The use of this proxy is motivated by the fact that the explicit entry for labour costs is sparsely populated in Compustat Annual (and does not exist in Compustat Quarterly). Imrohoroglu and Tüzel (2014) perform a robustness check and claim that the growth rate of this proxy matches well with the the direct entry of labour costs where both exist.

Note that as labour costs is a slow moving term, the bulk of variation in the quarterly growth of firm-value added comes from volatility in $OIBDPQ_{it}$. I also deflate the measure of value added by the GDP deflator.

[20]Typically such one off spikes are due to accounting changes which do not reflect the underlying productivity of the firm. For example, from 2012 Verizon adjusts its pension liabilities in the fourth quarter of every year to account for changes in interest rate and mortality assumptions. This charge is technically a labour cost so passes through operating expenses, but does not reflect the underlying productivity of Verizon.

Labour and Capital

The number of employees is recorded in the Compustat Annual database, but not the Compustat Quarterly database. As such, I assign the Compustat Annual number to the final quarter in the financial year for each company (as the annual number corresponds to the end-of-year number, not an annual average) and then interpolate the intervening quarters using a spline. The results reported in the paper are robust to using a linear interpolation.

While for some companies a quarterly measure of Property, Plant and Equipment (the basic measure of capital) does exist, the series is sparse and implausibly volatile (in the case of General Motors and Boeing, for example, the end of year numbers align with the Compustat Annual number, but the intervening quarters show a sharp drop in Property, Plant and Equipment recorded). As such, I use the annual data (series PPEGT in Compustat).

Following Imrohoroglu and Tüzel (2014), I need to convert this (nominal) value for the capital stock into real terms. I proceed as follows. First, I calculate the average age of the capital stock in years by taking the ratio of current depreciation (DP) over accumulated depreciation (DPACT). Then, I deflate the Property, Plant and Equipment at time t by the investment deflator (from the national accounts) at time $t-j$, where j is the average age of the capital stock. This is equivalent to assuming that all of the capital stock was created j periods before. After I have estimates for the real capital stock, I interpolate between the annual values to get a quarterly series in the same manner as the Labour series.

Calculating TFP

With value added, labour and (real) capital estimates on a quarterly basis, I can construct estimates of TFP by applying a two step procedure similar to that proposed by Olley and Pakes (1996) which is designed to deal with both bias from simultaneity between productivity and the labour input decision and survival bias. The exposition here is very close to that of Imrohoroglu and Tüzel (2014): I report it for the convenience of the reader. Consider the log-linear production function:

$$v_{it} = \alpha_0 + \alpha_k k_{it} + \alpha_l l_{it} + z_{it} + u_{it} \tag{1.32}$$

where v_{it} is log value added of firm i in period t, k_{it} is the log of the real capital stock, and l_{it} is the log of the number of employees, z_{it} is an unobserved hicks-neutral productivity shock and u_{it} is an error. Naively estimating (1.32) by least squares leads to two separate problems. First, the unobserved productvity shock z_{it} is likely to be correlated with l_{it} (a "time to build" assumption means that it is usual to treat k_{it} as predetermined). Second, if firms which recieve a low productivity draw are more likely to exit, this will also generate survivorship bias. To proceed, Olley and Pakes (1996) note that the current investment decision can be described by the some function

$$i_{it} = i(z_{it}, k_{it})$$

where i_{it} is investment by firm i in period t, and is monotonically increasing in z_{it}. This allows us to invert the investment function:

$$z_{it} = h(i_{it}, k_{it})$$

Define

$$\phi_{it} = \alpha_0 + \alpha_k k_{it} + h(i_{it}, k_{it}) \tag{1.33}$$

Combining (1.32) and (1.33):

$$v_{it} = \alpha_l l_{it} + \phi_{it} + u_{it} \tag{1.34}$$

By appoximating ϕ_{it} with a second order polynomial in i_{it} and k_{it}, it is possible to get consistent estimates of α_l which controls for the simultaneity problem. The second step of the Olley and Pakes (1996) regression deals with surviorship bias. Consider:

$$\mathbb{E}_t \left[v_{it+1} - \hat{\alpha}_l l_{it+1} \right] = \alpha_0 + \alpha_k k_{it} + \mathbb{E}_t \left[z_{it+1} | z_{it}, survival \right] \tag{1.35}$$

The last term: the expectation of productivity at *t+1* given both current productivity and survival, is a function of z_{it} and $\hat{P}_{survival,t}$, the probability of survival from t to *t+1*. The latter is the fitted value from the estimation of a probit model regressing survival on a second order polynomial in capital and investment. It is then possible to estimate the following expression by nonlinear least squares:

$$v_{it+1} - \hat{\alpha}_l l_{it+1} = \alpha_0 + \alpha_k k_{it} + \rho z_{it} + \tau \hat{P}_{survival,t} + u_{it+1} \tag{1.36}$$

where $z_{it} = \widehat{\phi}_{it} - \alpha_0 - \alpha_k k_{it}$ is assumed to follow an AR(1) process. This estimation step gives a consistent estimator for α_0 and α_k. With the parameters of the production function in hand, it is possible to extract estimates of z_{it}.

I follow this procedure using data on v_{it}, l_{it} and k_{it} from the Compustat Annual database at the three digit industry level.

Once I have estimates for $\alpha_0, \alpha_k, \alpha_l$, I can recover an estimate of TFP by using (1.32). I then seasonally adjust this measure using a moving average filter. Following Gabaix (2011) I winsorize the quarterly productivity growth rates. This reduces the impact of extreme growth rates which might be brought about (for example) by corporate actions such as mergers or accounting changes rather than true growth in firm productivity. I winsorize this data at a 30% quarterly growth rate.

1.A.2 Stock price information

Calculating Excess Stock Returns

Following the methodology of MacKinlay (1997), I calculate excess stock returns for company i as:

$$\widehat{R}_{it} = R_{it} - \widehat{\beta}_{0y} - \widehat{\beta}_{1y} R_{IND,t} \tag{1.37}$$

where R_{it} is the total stock return (including dividends) of company i on day t, $R_{IND,t}$ is the total index return on day t, and $(\widehat{\beta}_{0y}, \widehat{\beta}_{1y})$ are the estimated constant and slope coefficients of the ordinary least squares regression of R_{it} on a constant and $R_{IND,t}$ for year y. I drop excess returns corresponding to days on which the $R_{IND,t}$ is more than three standard deviations away from its annual average.

Stock price data is obtained from COMPUSTAT/CRSP linked data. The benchmark results use the total returns of the S&P 500 as the appropriate index, though the results are robust to using the CRSP equally-weighted index as well.

Results days

Results days are obtained by combining the RDQ series from the COMPUSTAT Quarterly database with data from Thomson Reuters Institutional Broker Estimate System (I/B/E/S). The latter is used as it records the time of release as well as the date. If the time falls after market close (16:00 EST) then I take the following day's excess return. If the I/B/E/S data is missing, I assume that the results are released before the market close. I also check my results for robustness by using two-day returns, but this makes little difference to the reported results.

Days on which 8-k forms are released are identified by the time of filing in the Security and Exchange Commission's EDGAR database. As with results days, I treat filing times (when available) after 16:00 EST as pertaining to the following day's market move.

1.B Recovering structural shocks using an external proxy

Consider the VAR system:

$$\phi(L)y_t = e_t$$

$$e_t = B\varepsilon_t$$

The standard identification problem arises as $\Sigma = BB'$ contains n^2 unknowns but only $\frac{(n+1)n}{2}$ equations. Mertens and Ravn (2013) show that if we have an external proxy, m_t, which satisfies conditions (1.18) and (1.19)-(1.22) then it is possible to identify the column of B corresponding to variable p and calculate impulse responses to shocks to this variable. In general, this method does not deliver full identification of the system in question. However, it is possible for us to back out a unique time series of the structural shock correlated with the proxy. Suppose that this shock is ordered first, and note that the decomposition must satisfy both

i. $\hat{\Sigma} = BB'$

ii. the estimates of the first column of B using the external instrument

Let B_0, B_1 be matrices satisfying both (i) and (ii). Then there exists a Q such that:

$$B_1 = B_0 \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix} = B_0 \widetilde{Q} \tag{1.38}$$

where $Q'Q = I$. Then

$$\begin{aligned} \varepsilon_t &= B^{-1} e_t \\ &= \widetilde{Q}^{-1} B_0^{-1} e_t \end{aligned}$$

Note that

$$\widetilde{Q}^{-1} = \widetilde{Q}' = \begin{bmatrix} 1 & 0 \\ 0 & Q' \end{bmatrix}$$

$$\begin{bmatrix} \varepsilon_{p,t} \\ \varepsilon_{q,t} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & Q' \end{bmatrix} B_0^{-1} \begin{bmatrix} e_{p,t} \\ e_{q,t} \end{bmatrix}$$

Which means that any Q which is orthonormal (satisfies $Q'Q = I$) will give the same time series of $\{\varepsilon_p\}$ as any other orthonormal Q.

Chapter 2

Predictable Recoveries

2.1 Introduction

Accurate forecasts of future economic growth are very valuable, for example, because they are needed for policymakers to decide on the appropriate stance of monetary and fiscal policy. Good forecasts are also important for the private sector, for example, for investment decisions or purchases of durable consumption goods. For these reasons, it is important that such forecasts are done with utmost care; forecasts that are too pessimistic or too buoyant could induce the wrong decisions and be quite harmful. Understanding what lies ahead is especially important during recessions, which explains the strong interest to understand what the short-term and long-term consequences of economic downturns are for future output levels.

Campbell and Mankiw (1987) argued that:

> "The data suggest that an unexpected change in real GDP of 1 percent should change one's forecast by over 1 percent over a long horizon."

Thus, shocks to GNP are permanent. Moreover, it implies that reductions in real activity are associated – if anything – with predictable deteriorations, not predictable recoveries. More recently, this quote was repeated on Mankiw's blog.[1] Campbell and Mankiw (1987)

[1] See: `http://gregmankiw.blogspot.com/2009/03/team-obama-on-unit-root-hypothesis.html`.

base their conclusion on estimated univariate ARMA models, that is,[2]

$$\phi(L)\,\Delta y_t = a_0 + \theta(L)\,e_t, \tag{2.1}$$

where y_t is the log of real GDP and e_t is a serially uncorrelated shock. In this class of time-series models, there is only one type of shock, that is, the response of output to realizations of e_t is always the same, independent of why there is a shock to output.

The contribution of this paper is twofold. First, we document that the claim made in Campbell and Mankiw (1987) is not very accurate. Using a simple *multivariate* time series model, we show that US recessions were often (but not always) followed by *predictable* recoveries.[3] Consistent with the results in Campbell and Mankiw (1987), these recoveries were not predicted by univariate time-series models.

The second contribution of this paper is to put forward reasons why univariate time-series models for GDP may lead to inaccurate forecasts. Key in our arguments is that GDP is an *aggregate* of other random variables.

The first reason is that a univariate representation does not have the flexibility to incorporate shocks with different persistence levels. A striking illustration is given in Blanchard et al. (2013). They construct an example in which the *correct* univariate specification of a stochastic variable that is the sum of an integrated variable *with* predictable changes and a stationary variable, also *with* predictable changes, is a random walk. That is, using only information about the aggregate variable, the correct univariate representation indicates that all changes are permanent, even though both innovations of the underlying system imply predictable further changes. We derive a more general version of this result.

The key lesson is the following. Macroeconomic aggregates are likely to be the sum of stationary and non-stationary variables. A correct univariate representation of such

[2]They allow for the possibility that $\theta(L)$ has a root equal to 1, which would imply that y_t is stationary around a deterministic time trend.

[3]We also compare univariate and multivariate time-series models to predict UK recoveries. Whereas several US recessions were followed by remarkable recoveries, economic recoveries in the UK were much more gradual and the predictions of the two types of models are similar. However, the multivariate model does outperform the univariate model during the great recession. In particular, the multivariate model correctly predicts a further deterioration in the initial phase of the economic downturn and correctly predicts its long-lasting impact.

a variable must indicate that it is non-stationary, which means that the impact of the shock of the univariate representation necessarily has a permanent impact. We show that similar distortions occur when a random variable is the sum of two stationary variables with different persistence levels.

The second reason that univariate models may prove problematic is that the true ARMA representation of an aggregate variable may be more complex than the most complex ARMA process of each of its component series. This argument, pointed out by Granger and Morris (1976) and Granger (1980), means that with a finite data sample it might be difficult to identify the correct ARMA specification. This means that univariate time series models for aggregate variables may generate misleading forecasts. In this paper, we analyze how the under-parameterization of a univariate time series model can lead to biased forecasts.

We compare predictions of the univariate representation with those based on a VAR of GDP's expenditure components. It strengthens our argument that even such a simple multivariate time series model generates quite different forecasts during recessions. This finding is consistent with results from the forecasting literature that richer models can outperform univarate time series models.[4] Nevertheless, univariate time-series models have a long history and remain important. Nelson (1972) documents that large-scale macroeconometric models with many equations do not outperform forcasts made by simple ARIMA models. Similarly, Edge and Gurkaynak (2010) and Edge et al. (2010) show that forecasts made by DSGE models can be worse than a simple forecast of constant output growth.[5]

In section 2.2, we provide some theoretical background and discuss reasons why univariate representations may overestimate the long-run impact of economic downturns. In section 2.3, we illustrate some key time-series properties of US GDP. In section 2.4, we compare the precision of forecasts made by univariate and multivariate time-series

[4]Fair and Shiller (1990) also show that GDP forecasts based on the sum of forecasts of GDP's components help improve forecasts when compared with univariate forecasts. They use univariate representations of the components, which makes it possible to disaggregate at a higher level. Stock and Watson (2002) generate forecasts using a small number of indexes that are based on the principal components of a large set of economic variables. We refer the reader to Chauvet and Potter (2013) for a recent survey of the forecasting literature.

[5]By contrast, Smets and Wouters (2007) show that their DSGE model performs better in forecasting than a Bayesian VAR.

models. In section 2.5, we document what this meant for forecasts made during US post-war recessions. In section 2.6, we show that multivariate representations also have advantages for predicting UK GDP, but for quite different reasons than the ones outlined above. The last section concludes.

2.2 Econometrics of univariate time-series models

In section 2.2.1, we illustrate why *univariate* time-series representations can give misleading predictions *even* if they are correctly specified. In particular, it is possible that the variable of interest, y_t, is a random walk and (i) it is not necessarily true that all changes in this variable have a permanent effect and (ii) the model's predictions made during recessions systematically overpredict the persistence of the downturn. In section 2.2.2, we give reasons why it may be difficult to get a correctly specified univariate representation for aggregate variables.

2.2.1 Univariate representation: Missing information and bias

Consider the following data generating process (*dgp*) for y_t:[6]

$$\begin{gathered}
y_t \equiv x_t + z_t, \\
(1-\rho L)\, x_t = e_{x,t}, \\
(1-\rho L)\,(1-\rho_z L)\, z_t = e_{z,t}, \\
\mathbb{E}_t\left[e_{x,t+1}\right] = \mathbb{E}_t\left[e_{z,t+1}\right] = \mathbb{E}_t\left[e_{x,t+1}e_{z,t+1}\right] = 0,\, \mathbb{E}_t\left[e_{x,t+1}^2\right] = \sigma_x^2, \mathbb{E}_t\left[e_{z,t+1}^2\right] = \sigma_z^2,
\end{gathered} \tag{2.2}$$

where $\mathbb{E}_t\left[\cdot\right]$ denotes the expectation conditional on current and lagged values of x_t and z_t. The persistence of the effects of $e_{x,t}$ on x_t is determined by the value of ρ and the persistence of the effects of $e_{z,t}$ on z_t is controlled by both ρ and ρ_z. We assume that

$$-1 < \rho < 1, \tag{2.3}$$

$$-1 < \rho_z \leq 1, \tag{2.4}$$

$$\frac{\rho_z}{\rho} > 1. \tag{2.5}$$

[6]This time-series specification is a generalization of the one studied in Blanchard et al. (2013).

We define $e_{y,t}$ such that the following holds:[7]

$$\left(1 - \rho_z L\right) y_t = e_{y,t}, \tag{2.6}$$

The unconditional autocovariance of $e_{y,t}$ and $e_{y,t-j}$, $\mathbb{E}\left[e_{y,t}e_{y,t-j}\right]$, is given by

$$\mathbb{E}\left[e_{y,t}e_{y,t-j}\right] = \frac{\rho^j}{1-\rho^2}\sigma_z^2 + \left(\left(\rho - \rho_z\right)\rho^{j-1} + \frac{\left(\rho-\rho_z\right)\rho^j}{1-\rho^2}\right)\sigma_x^2. \tag{2.7}$$

This implies that the autocovariances of $e_{y,t}$ are equal to zero if the following equation holds:[8]

$$\sigma_z^2 = \frac{\left(\rho_z - \rho\right)\left(1 - \rho_z\rho\right)}{\rho}\sigma_x^2. \tag{2.8}$$

If this equation is satisfied, then $e_{y,t}$ is serially uncorrelated, and the correct univariate time-series specification of y_t is an $AR\left(1\right)$ with coefficient ρ_z.

In this univariate representation for y_t, there is only one shock, $e_{y,t}$, and the persistence of the effects of this shock is solely determined by ρ_z. Thus, the value of ρ does not matter at all! This is remarkable given that ρ affects the persistence of both fundamental shocks, $e_{x,t}$ and $e_{z,t}$.

To understand why the univariate representation misses key aspects of the underlying system, consider the case considered in Blanchard et al. (2013) when $\rho_z = 1$. The univariate representation is then given by

$$y_t = y_{t-1} + e_{y,t}. \tag{2.9}$$

That is, Δy_t is white noise and y_t is a random walk. Although y_t is a random walk, almost all changes in y_t imply predictable further changes according to the underlying multivariate *dgp*.[9] In particular, if $\Delta y_t < 0$ because $e_{x,t} < 0$, then there is a *predictable*

[7] It is always true that

$$\left(1 - \rho_z L\right)\left(1 - \rho L\right) y_t = \left(1 - \rho_z L\right) e_{x,t} + e_{z,t}.$$

Thus, an equivalent definition of $e_{y,t}$ would be the following:

$$\left(1 - \rho L\right) e_{y,t} = \left(1 - \rho_z L\right) e_{x,t} + e_{z,t}.$$

These two equations are helpful in deriving the formulas in this section.

[8] $\sigma_z > 0$, since we assumed that $\rho_z/\rho > 1$.

[9] In the (very) special case that $(1-\rho)x_t$ happens to be equal to $\rho\Delta z_t$, then $\mathbb{E}[y_{t+k}] = y_t$ for $k \geq 1$.

recovery in y_t, since $x_t = \rho x_{t-1} + e_{x,t}$ and $0 < \rho < 1$. If $\Delta y_t < 0$ because $e_{z,t} < 0$, then there is a *predictable* further deterioration, since $\Delta z_t = \rho \Delta z_{t-1} + e_{z,t}$ and $\rho > 0$. If one only observes that $\Delta y_t < 0$, then one has to weigh the two possible cases and in this example the two opposing effects exactly offset each other, leading the forecaster to predict that the level of output will remain the same.

Although the implications are most striking when $\rho_z = 1$, which is the case considered in Blanchard et al. (2013), the analysis presented here makes clear that the univariate representation of y_t does not incorporate the role of ρ for *any* value of ρ_z such that $-1 < \rho_z \leq 1$.

The *dgp* considered in this section is special because the forecastability that is present in the different components cancels out and disappears in the univariate representation. It is true more generally, however, that important information is lost in the univariate representation of the sum of variables.

Is the predicted long-run impact correct on average? The previous discussion showed that the univariate representation given in equation (2.6) clearly misses some aspects of the underlying data generating process. Next, we turn to the question whether the univariate representation generates (long-term) predictions that are *on average* correct.

To simplify the discussion, we focus on a particular version of the *dgp* given in equation (2.2). We assume that $\rho_z = 1$ and equation (2.8) is satisfied, so that the univariate representation of y_t is a random walk. Moreover, we set $\sigma_x = \sigma_z = \sigma$, which implies that $\rho = 0.381966$ according to equation (2.8). Finally, we assume that $e_{x,t}$ and $e_{z,t}$ can take only two values, namely $-\sigma$ and $+\sigma$, both with equal probability. Note that the value of y_t remains unchanged if $e_{x,t}$ and $e_{z,t}$ have the opposite sign.

Although y_t has a random-walk representation, it systematically overpredicts the long-term consequences when output falls, i.e., during recessions, and it systematically underpredicts long-term consequences when output increases.

Before showing this, we first consider the case when output remains the same, which happens if $e_{x,t}$ and $e_{z,t}$ have the opposite sign. The (long-run) predictions based on the random-walk specification remain the same, since y_t remains the same. However, the true long-run predictions are affected as follows:

$$\begin{aligned} &\lim_{\tau\longrightarrow\infty} \mathbb{E}_t\left[y_{t+\tau}\right] - y_t = +\sigma/\left(1-\rho\right) \text{ if } e_{z,t} = +\sigma \text{ and } e_{x,t} = -\sigma \text{ and} \\ &\lim_{\tau\longrightarrow\infty} \mathbb{E}_t\left[y_{t+\tau}\right] - y_t = -\sigma/\left(1-\rho\right) \text{ if } e_{z,t} = -\sigma \text{ and } e_{x,t} = +\sigma. \end{aligned} \tag{2.10}$$

Thus, when y_t remains the same, then one fails to recognize that the long-run value of y_t has gone up half of the time and fails to recognize that this long-run value has gone down the other half of the time. However, the forecasts are not systematically wrong.

Now consider the case in which output drops, which happens when $e_{x,t} = e_{z,t} = -\sigma$. The drop in output is equal to $-\sigma_x - \sigma_z = -2\sigma$. The random-walk specification implies that the long-run impact is identical to the short-term impact, that is,

$$\lim_{\tau\longrightarrow\infty} \widehat{\mathbb{E}}_t\left[y^f_{t,t+\tau}\right] - y_t = -2\sigma, \tag{2.11}$$

where $\widehat{\mathbb{E}}_t\left[\cdot\right]$ is the expectation according to the (correct) univariate representation. The *true* long-run impact of the shock, however, is equal to

$$\lim_{\tau\longrightarrow\infty} \mathbb{E}_t\left[y_{t+\tau}\right] - y_t = -\sigma/(1-\rho) = -1.618\sigma. \tag{2.12}$$

That is, in a recession, the univariate representation systematically overpredicts the long-run negative impact of the economic downturn. Similarly, the univariate representation systematically overpredicts the long-run positive impact of an increase in y_t. So the predictions are not biased, but one clearly is too pessimistic during recessions and too optimistic during booms if one would make predictions based on the random-walk specification.

In this stylized example in which $e_{x,t}$ and $e_{z,t}$ can take only two values, one could drastically improve on the predictions of the univariate representation even if one could not observe x_t or z_t, but knows the true *dgp*. The reason is that a drop in y_t implies that $e_{x,t}$ and $e_{z,t}$ are both negative and an increase implies that both shocks are positive. The idea that the *magnitude* of the unexpected change in y_t has information about the importance of $e_{x,t}$ and $e_{z,t}$ is also true for more general specifications of $e_{x,t}$ and $e_{z,t}$, as long as one has information about the distribution of the two shocks. If one observes a very large drop in y_t, then it is typically the case that it is more likely that $e_{x,t}$ and $e_{z,t}$ are both negative than that $e_{x,t}$ is positive and $e_{z,t}$ is so negative it more than offsets the positive value of $e_{x,t}$ or vice versa. That is, the larger the economic downturn the larger the probability that a certain fraction of this downturn is driven by the transitory shock, that is, the larger the probability that a fraction of the drop in real activity will be reversed.

2.2.2 Aggregated variables and correctly specifying their *dgp*s

Aggregating ARMA processes. In this section, we highlight another problem with working with aggregated variables. We illustrate that the correct ARMA representation of an aggregate variable may very well be more complex than the most complex ARMA process for each of the component series. Formally, if x_t is an $ARMA(p_x, q_x)$ and z_t is an $ARMA(p_z, q_z)$, then $y_t \equiv x_t + z_t$ is an $ARMA(p, q)$ and p and q satisfy the following condition:[10]

$$p \leq p_x + p_z \text{ and } q \leq \max\{q_x + p_z, q_z + p_x\}. \tag{2.13}$$

These conditions give *upper* bounds for the $ARMA$ representation of the sum, y_t. Thus, the $ARMA$ representation of y_t is not necessarily of a higher order than those of x_t and z_t. In fact, in section 2.2.1 we gave an example in which an $AR(1)$ variable and an $AR(2)$ variable add up to an $AR(1)$ variable.[11] But that example relies on specific parameter restrictions. In practice, one should not rule out the possibility that the univariate representation of a sum of several random variables could be quite complex. In fact, Granger (1980) argues that an aggregate of many components—as is the case for typical macroeconomic variables—may exhibit long memory.[12]

One might think that the solution to this dilemma is to use more complex $ARMA$

[10] See Granger and Morris (1976).

[11] In theory it is, of course, even possible that the sum of random variables is not random.

[12] One aspect that seems to be ignored in the econometrics literature is that the *dgp*s of the individual components may be "aligned" to the same factors, which could mean that the time-series representations of the components are similar, making it less likely that the aggregate has a much more complex representation than its components. For example, if markets are complete, then market prices will align agents' marginal rates of substitution—and, thus, their consumption growth processes—even if agents face very different income processes.

processes for aggregate variables. The problem is that the model has to be estimated with a finite amount of data, consequently the values of p and q cannot be too high. But if the values of p and/or q are too low, then the *dgp* could be misspecified.[13]

Simple example. We will now give a simple example, in which the predictions of a univariate time-series model for an aggregated variable are quite bad if that time-series model is *not* more complex than the most complex time-series representation of the components.

Consider the following *dgp*:

$$\begin{aligned}
y_t &\equiv x_t + z_t, \\
x_t &= \rho_x x_{t-1} + e_{x,t}, \\
z_t &= e_{z,t}, \\
\mathbb{E}_t\left[e_{x,t+1}\right] &= \mathbb{E}_t\left[e_{z,t+1}\right] = 0, \\
\mathbb{E}_t\left[e_{x,t+1}^2\right] &= \sigma_x^2, \\
\mathbb{E}_t\left[e_{z,t+1}^2\right] &= \sigma_z^2,
\end{aligned} \tag{2.14}$$

with $-1 < \rho_x < 1$. Thus, y_t is the sum of two stationary random variables, an AR(1) and white noise. Equation (2.14) implies that

$$(1-\rho_x L)\, y_t = e_{x,t} + (1-\rho_x L)\, e_{z,t}. \tag{2.15}$$

The first-order autocorrelation of the term on the right-hand side is not equal to zero unless $\rho_x = 0$, but higher-order autocorrelation coefficients of this term are equal to zero. Consequently, y_t is an $ARMA\,(1,1)$. That is, there is a value for θ such that the following is the correct univariate time-series representation of y_t:

$$(1-\rho_x L)\, y_t = (1+\theta L)\, e_{y,t}, \tag{2.16}$$

where $e_{y,t}$ is serially uncorrelated. The value of θ is given by the following expres-

[13]The misspecification is likely to be worse than indicated in this section. Typically, log-linear processes are more suitable than linear processes. But if $y_t \equiv x_t + z_t$ and x_t and z_t are log-linear processes, then neither y_t nor $\ln(y_t)$ is a linear process and the convention of modelling $\ln(y_t)$ as a linear process is, thus, not correct. In fact, the effects of shocks on y_t would be time-varying. These issues are further discussed in Haan et al. (2011).

sion:[14]

$$\theta = \frac{\rho_x \left(-\mathbb{E}\left[e_{x,t}e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right]\right)}{\mathbb{E}\left[e_{y,t}^2\right]}. \tag{2.17}$$

The most complex component of y_t is x_t, which is an AR(1). So suppose that y_t is also modelled as an AR(1). That is,

$$y_t = \widetilde{\rho}_y y_{t-1} + \widetilde{e}_{y,t}. \tag{2.18}$$

If we abstract from sampling uncertainty, we can pin down the value of $\widetilde{\rho}_y$ using population moments:

$$\widetilde{\rho}_y = \frac{\mathbb{E}\left[y_t y_{t-1}\right]}{\mathbb{E}\left[y_t^2\right]} = \frac{\left(\rho_x + \theta\right)\left(1 + \rho_x\theta\right)}{\left(1-\rho_x^2\right) + \left(\rho_x + \theta\right)^2}. \tag{2.19}$$

We are interested in whether this AR(1) specification would tend to over- or underestimate the long term effects of shocks by comparing $|\widetilde{\rho}_y|$ with $|\rho_x|$. If $|\widetilde{\rho}_y| > |\rho_x|$, then the AR(1) specification would tend to overstate the true degree of persistence. It is straightforward to show that $|\widetilde{\rho}_y| > |\rho_x|$ if and only if $\theta\rho_x > 0$, that is, if ρ_x and θ have the same sign.[15] Equation (2.17) implies that this happens if

$$-\mathbb{E}\left[e_{x,t}e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right] > 0. \tag{2.21}$$

This condition is satisfied if the covariance of $e_{x,t}$ and $e_{z,t}$ is sufficiently negative. Similarly, $|\widetilde{\rho}_y| < |\rho_x|$ if and only if ρ_x and θ have the opposite sign, which happens if

$$-\mathbb{E}\left[e_{x,t}e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right] < 0. \tag{2.22}$$

[14]Since $e_{y,t}$ is white noise, it must be true that

$$\mathbb{E}\left[\left(1+\theta L\right)e_{y,t} \times \left(1+\theta L\right)e_{y,t-1}\right] = \theta\mathbb{E}\left[e_{y,t}^2\right].$$

It is also true that

$$\mathbb{E}\left[\left(1+\theta L\right)e_{y,t} \times \left(1+\theta L\right)e_{y,t-1}\right] = \rho_x\left(-\mathbb{E}\left[e_{x,t}e_{z,t}\right] - \mathbb{E}\left[e_{z,t}^2\right]\right),$$

since $\left(1+\theta L\right)e_{y,t} = e_{x,t} + \left(1-\rho_x L\right)e_{z,t}$ and both $e_{x,t}$ and $e_{z,t}$ are white noise. Combining both equations gives the expression for θ.

[15]Equation (2.19) implies that $|\widetilde{\rho}_y| > |\rho_x|$ if

$$\begin{aligned} &\tfrac{\left(1-\rho_x^2\right)}{\left(1-\rho_x^2\right)+\left(\rho_x+\theta\right)^2}\theta > 0 \quad \text{when } \rho_x > 0, \\ &\tfrac{\left(1-\rho_x^2\right)}{\left(1-\rho_x^2\right)+\left(\rho_x+\theta\right)^2}\theta < 0 \quad \text{when } \rho_x < 0. \end{aligned} \tag{2.20}$$

Consequently, $|\widetilde{\rho}_y| > |\rho_x|$ if and only if $\theta\rho_x > 0$, that is, if ρ_x and θ have the same sign.

This condition would be satisfied if the two shocks are positively correlated.

To shed some light on the possible consequences of using an $AR(1)$ as the law of motion for y_t, we consider the case when the two shocks have the following very simple relationship:

$$e_{z,t} = \alpha e_{x,t}. \tag{2.23}$$

Since $e_{x,t}$ and $e_{z,t}$ are perfectly correlated, there is only one type of shock and there is a univariate time-series specification of y_t that completely captures the dynamics of y_t. Now we investigate what the consequences of misspecifying the $ARMA(1,1)$ process as an $AR(1)$—as an $AR(1)$ is the most complex of the individual underlying time series processes.

Figure 2.1 plots $\widetilde{\rho}_y$, i.e., the value of the coefficient of the $AR(1)$ representation of y_t, as a function of the true dominant root in the *dgp* of y_t, i.e., ρ_x. The top panel considers the case when the two shocks are negatively correlated ($\alpha < 0$). In this case, $\widetilde{\rho}_y$ is greater than ρ_x and so the $AR(1)$ process overstates the true amount of persistence. Conversely, if the shocks are positively correlated $\widetilde{\rho}_y$ is less than ρ_x, as shown in the lower panel.

These two panels document that long-term persistence is increased substantially for lower values of ρ_x when α is negative and that long-term persistence is decreased substantially for higher values of ρ_x when α is positive.

Figure 2.2 displays IRFs for three sets of parameter values. Each panel plots the true response of y_t to a one-time shock in $e_{x,t}$ and the response according to the $AR(1)$ specification for y_t. These three panels clearly document that misspecifying the aggregate variable y_t as an $AR(1)$—the correct specification of the most complex of the underlying processes—can give inaccurate impulse responses at both short and long horizons. The $AR(1)$ representation of y_t overestimates the long-term consequences of the shock when $e_{x,t}$ and $e_{z,t}$ are negatively correlated and underestimates them when the two shocks are positively correlated. The bottom two panels document that these bad long-term predictions only become apparent at forecast horizons of over 30 periods. At forecast horizons shorter than 30 periods, the $AR(1)$ representation of y_t *over*estimates the consequences of the crisis by a large margin when the shocks are *positively* correlated and vice versa. For example, when the shocks are negatively correlated, then the $AR(1)$ representation

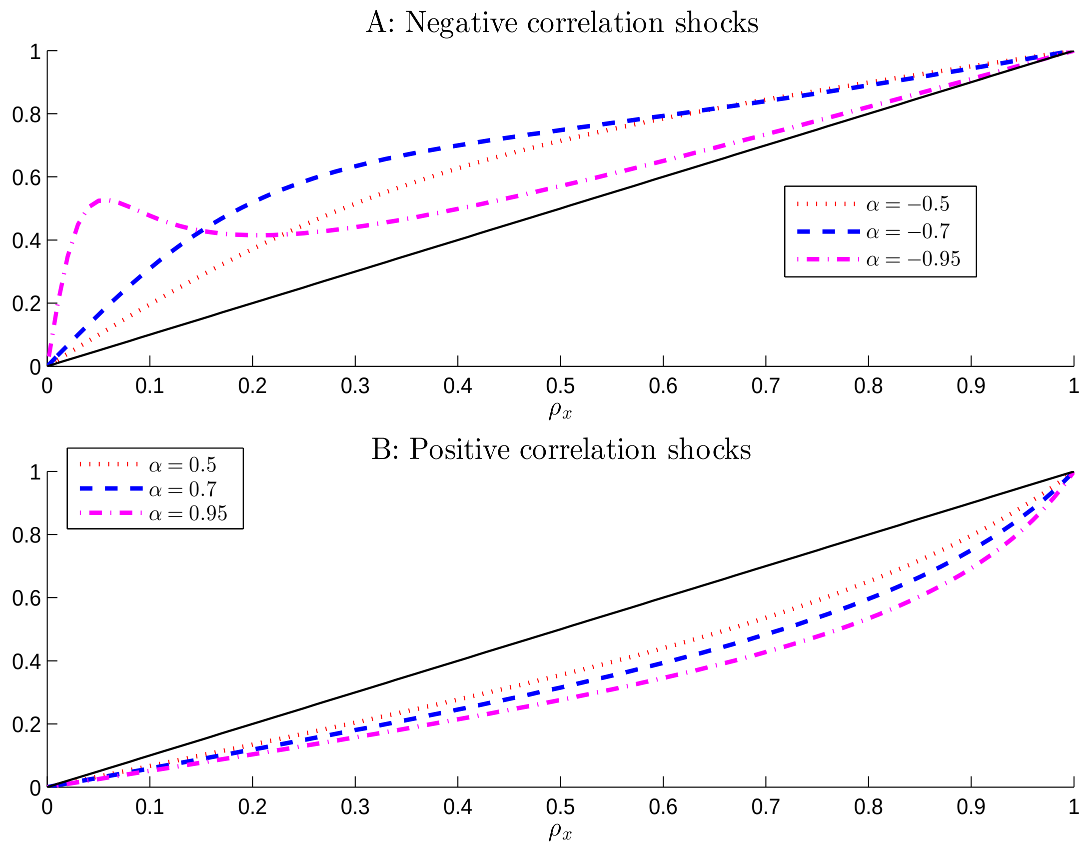

Figure 2.1: AR(1) coefficient of $y_t = x_t + z_t$ according to the incorrect univariate representation

Notes: The graph displays the root to the AR(1) representation of $y_t = x_t + z_t$ as a function of the AR root in the true time series representation of y_t when $e_{z,t} = \alpha e_{x,t}$. The solid line is the 45° line.

predicts that the initial reduction will be followed by an immediate but gradual recovery. By contrast, the true response is a further deterioration of almost the same magnitude followed by a somewhat faster recovery.

In this section, we focused on a case in which the most complex time-series specification of a component is an $AR(1)$, that is, a relatively simple process. Although the correct time-series specification of the aggregate is more complex, namely an $ARMA(1,1)$, it has only two parameters and one should be able to estimate this more complex time-series model with data sets of typical length. One can also improve on the $AR\,(1)$ specification by using higher-order AR processes, although these would—like the $AR(1)$—not be correct either, unless the number of lags is high enough to result in a sufficiently accurate approximation. However, the option to estimate a more complex representation may not always be feasible. If the two components are, for example, both an $AR(4)$, one would have to estimate an $ARMA(8,4)$, and if y_t is the sum of three $AR(4)$ processes, then one would have to estimate an $ARMA(12,8)$ to make sure that the univariate representation is not misspecified. In the next section, we document that a better strategy might be to estimate separate time-series models for the components and then explicitly aggregate the forecasts of the components to obtain forecasts for the aggregated variables.

2.3 Time series properties of US GDP

In this section, we discuss the relevance of the analyis in the last section by comparing an estimated univariate representation of US GDP with the representation that is implied by an estimated multivariate representation of its spending components.

2.3.1 Empirical specifications

The specification of the multivariate model is given by the following VAR:

$$\ln(s_t) = \sum_{j=1}^{p} B_j \ln(s_{t-j}) + e_{s,t}, \tag{2.24}$$

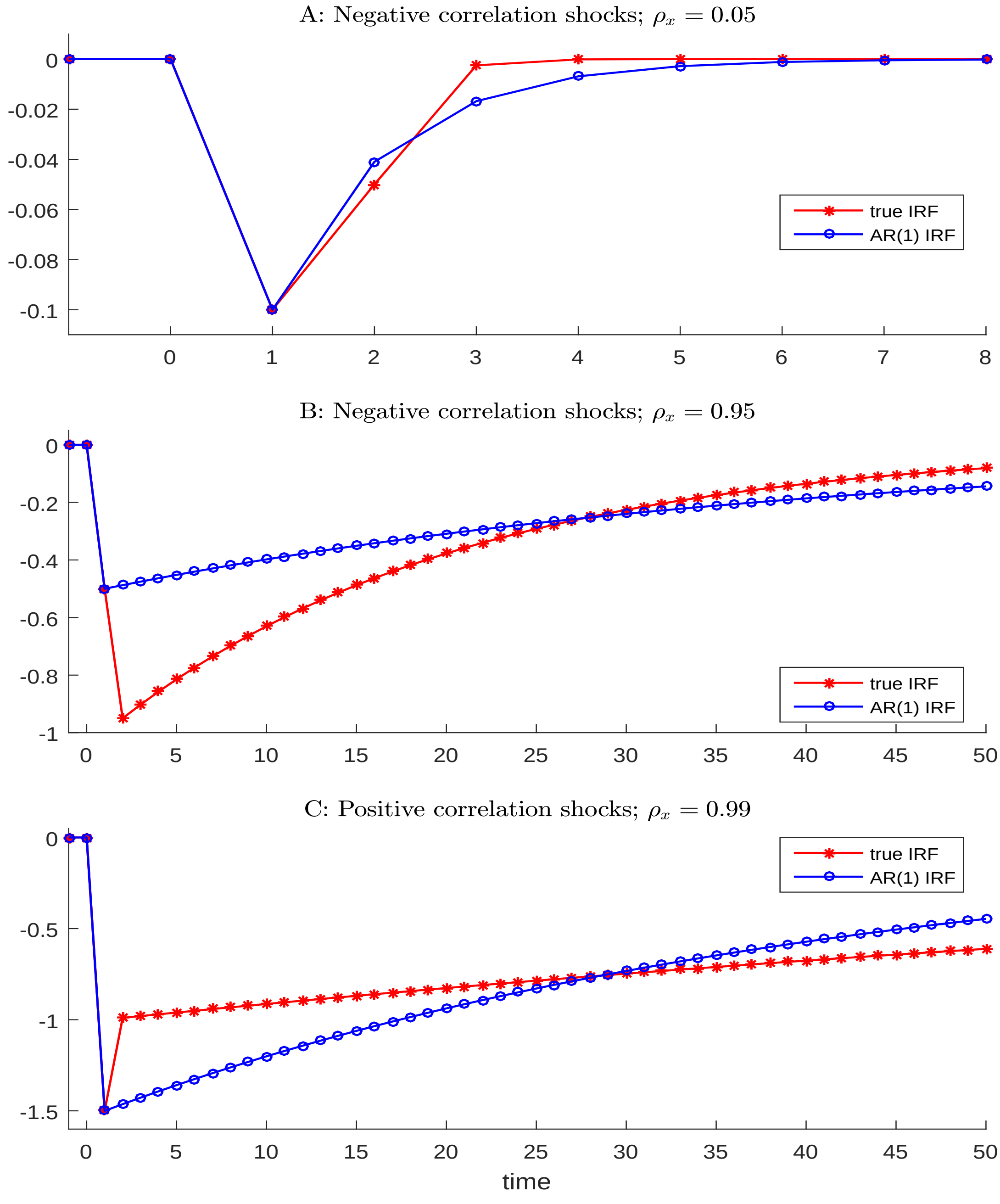

Figure 2.2: IRFs of $y_t = x_t + z_t$ according to the correct and incorrect univariate representation

Notes: The graph plots the true responses of $y_t = x_t + z_t$ to a one-time shock in $e_{x,t}$, and the response according to the AR(1) representation, which is the time series representation of the most complex of the y_t components. In panel A, $e_{z,t} = -0.9e_{x,t}$, in panel B $e_{z,t} = -0.5e_{x,t}$ and in panel C, $e_{z,t} = 0.9e_{x,t}$.

where s_t is a 5×1 vector containing the expenditure components, consumption, c_t; investment, i_t; government expenditures, g_t; exports, x_t; and imports. m_t. The forecast for $y_{t+\tau}$ follows directly from

$$y_{t+\tau} \equiv e^{\ln(c_{t+\tau})} + e^{\ln(i_{t+\tau})} + e^{\ln(g_{t+\tau})} + e^{\ln(x_{t+\tau})} - e^{\ln(m_{t+\tau})}. \tag{2.25}$$

The estimated univariate representation for aggregate output is given by:[16]

$$\ln(y_t) = \sum_{j=1}^{p} a_j \ln(y_{t-j}) + e_t. \tag{2.26}$$

The time series for y_t itself is also constructed using equation (2.25) so that we are comparing like with like exactly. The key feature of the univariate time-series model is that there is only one type of shock. If output turns out to be lower than expected, i.e., $e_t < 0$, then the predicted effect on future values of y_t will always have the same pattern with the magnitude proportional to the value of e_t.

Both time-series processes are estimated with ordinary least squares (OLS). Given that the variables could very well be integrated, it is important to add enough lags to ensure that the shocks are stationary and spurious regression results are avoided. If the time series are known to be integrated, then efficiency gains are possible by imposing this. Additional restrictions can be imposed if the series are cointegrated. If these restrictions are correct, but are not imposed, then the estimated parameter values will converge towards the true parameter values at rate T, that is, there is superconsistency. If the restrictions are not correct and are nevertheless imposed, then the system is misspecified and the estimated system will not converge towards the true system. Because of superconsistency, we prefer not to impose these types of restrictions on the system.

[16] We follow common practice and use four lags, unless stated otherwise. In appendix 2.B, we show that the results are similar when the number of lags is chosen by AIC, although the associated long-term forecasts are somewhat less precise. Results not reported here indicate that long-term forecasts are substantially less precise if the Bayesian Information Criterion (BIC) is used. All models in this paper also include a constant and a linear-quadratic deterministic trend. Appendix 2.B also shows that key results are very similar if no trend is included and when only a linear trend is included. Campbell and Mankiw (1987) also consider $ARMA$ representations, but the results are similar to those obtained with AR representations. The only exception is when third-order MA components are included, but the authors point out that the implied impulse response functions of this specification are estimated very imprecisely .

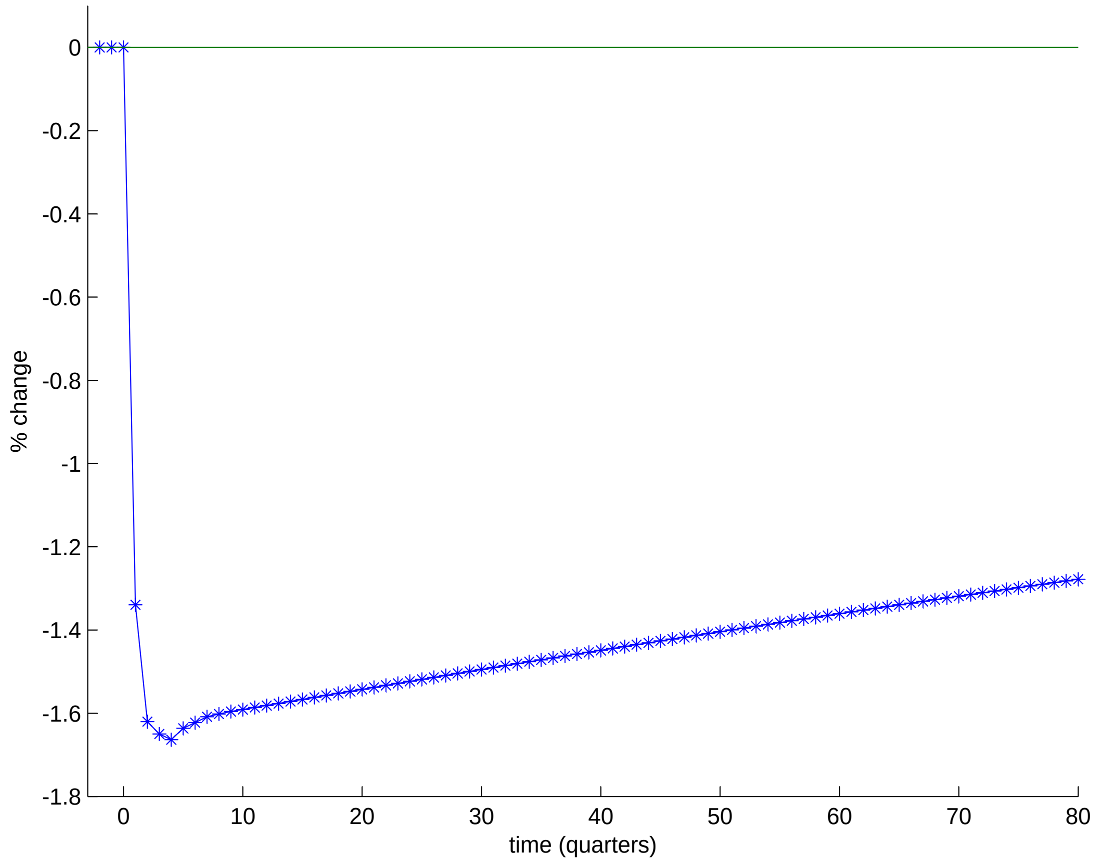

Figure 2.3: Effect of the shock in the univariate representatino of US GDP
Notes: The graph plots the response of output following a one-standard-deviation negative shock according to the univariate one-type-shock model.

2.3.2 Impulse response functions

The response of a negative one-standard-deviation shock to e_t on (the log of) US GDP, i.e., the impulse response function (IRF), is displayed in figure 2.3.[17] Even though the specification in equation (2.26) does not impose a unit root and contains a quadratic deterministic trend, the estimated specification documents that the response to the shock e_t is very persistent. It is exactly this type of result that underlies the argument of Campbell and Mankiw (1987) that one should expect economic downturns to have permanent effects.

If output is generated by the multivariate model, i.e., according to equations (2.24) and (2.25), then there are five reduced-form shocks that result in a drop in output.

[17] See Appendix 2.A for further details on data sources. Whereas the forecasting exercise discussed in the next section is based on real-time data, the results in this subsection are based on the full sample of quarterly US data from 1947Q1 to 2015Q1. The results are very similar if the sample ends in 2006Q4 and the financial crisis is, thus, excluded, except that the IRF of the “import” shock is then less persistent.

Consequently, there are five impulse response functions (IRFs), that is, five different ways in which output could respond. There are fierce debates in the economic literature on how to interpret shocks, but the interpretation of the shocks is not important for the point we want to make, that is, a model used to forecast GDP should allow for different forecasting patterns. For convenience, we will label the reduced-form shocks according to the dependent variable of the equation. For example, we will refer to $e_{c,t}$ as the consumption shock, but this is just a label and not meant to hint at a structural interpretation. The five IRFs are plotted in figure 2.4. The figure makes clear that according to the multivariate model there are shocks that have an extremely persistent impact on output. The figure also makes clear, however, that there are shocks that have a transitory impact on output.

2.3.3 Relevance of the theoretical arguments for modelling US GDP

The IRFs displayed in figure 2.4 indicate that several of the issues raised in section 2.2 could be relevant for forecasting US GDP using a univariate representation. The IRFs indicate that some events have long lasting consequences and others do not. For example, the "consumption shock" has a very persistent effect, but the "investment shock" and the "export shock" do not. This means that the analysis of section 2.2.1 is relevant. That is, since some components of US GDP are not stationary, the univariate representation will imply that *all* shocks to GDP will have a long-lasting effect.

With a finite sample, it is more difficult to determine whether the relatively parsimonious representation of GDP used here is the correct univariate representation. But the results of section 2.2.2 may give some guidance on potential problems. We find that the innovations of the components of GDP are positively correlated. As documented in figure 2.4, GDP consists of very persistent and not so persistent components. This resembles the example displayed in the bottom panel of figure 2.2. In this example, the univariate representation of the aggregate random variable overestimates the impact of shocks for a long period (up to 30 quarters), but underestimates the very long consequences.

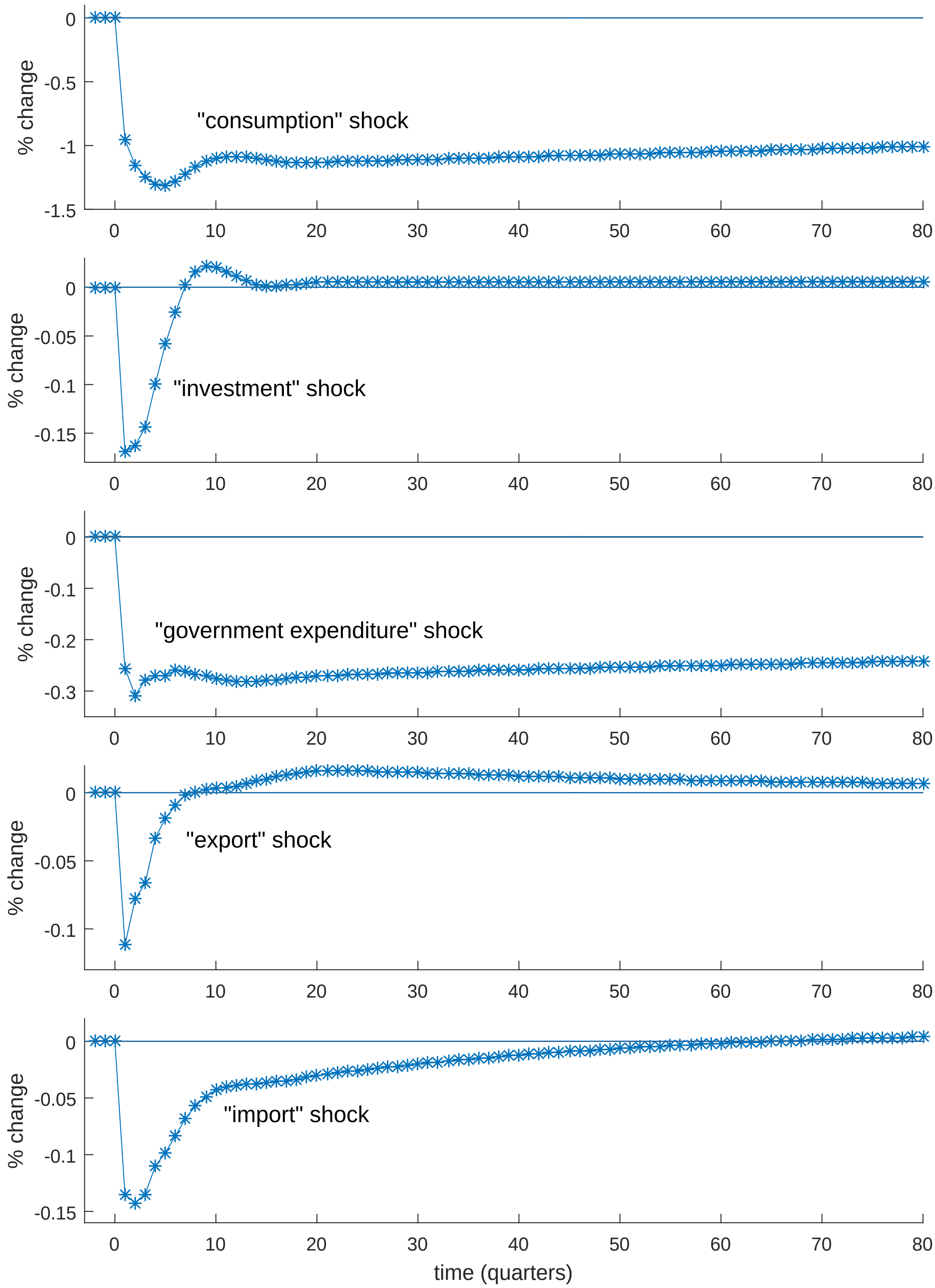

Figure 2.4: Effect of reduced-form VAR shocks on US GDP
Notes: The graphs plot the predicted repsonses of output following a one-standard-deviation shock in the indicated reduced-form VAR shock that leads to a reduction in GDP.

2.4 Forecasting US GDP with univariate and multivariate models

We use the univariate and the multivariate time-series models to forecast future GDP levels. Forecasts are out-of-sample forecasts, because forecasts made at t^* only use data up to date t^*.[18] We use the latest vintage of data for each forecast.

The left panel of figure 2.5 plots the average forecast error at different forecast horizons according to the univariate and the multivariate time-series models. The figure shows that the predictive power of the univariate model is just as good as that of the multivariate model in terms of average forecast errors. This does, of course, not imply that there are no multivariate models that outperform a univariate model. In fact, Stock and Watson (2002) document that a forecasting model that uses indexes based on the principal components of many economic variables outperforms autoregressive univariate for most (but not all) variables. Nevertheless, the result is somewhat surprising. After all, the IRFs of the expenditure components indicate that GDP has components characterized by different persistence levels and the theoretical analysis indicated that there should be advantages in constructing forecasts of the aggregate by combining the separate forecasts of the components.

But average forecast errors may obscure some interesting patterns. In particular, the multivariate model turns out to do substantially better in forecasting at longer forecast horizons during recessions. The right panel of figure 2.5 shows forecast errors averaged across the six US recessions starting with the 1973-75 recession. NBER dates are used to determine whether a quarter falls in a recssion. The figure shows that the multivariate model generates much better forecasts at higher forecasting horizons.

Since *average* forecasting errors of the two types of models are similar, there must be periods when the univariate time-series model generates better forecasts. Interestingly, that happens during "ordinary" times, when the economy is neither doing very well nor

[18]Strictly speaking, this is *pseudo* out-of-sample forecasting, since future data is available at each forecasting point. We estimate specifications with two lags if they have fewer than 135 observations and four lags otherwise. The exact cutoff point does not matter, but it is important to only use only two lags at the early dates of our forecasting exercise, because the specifcations with four lags generate strange forecasts, which is likely to be due to the low number of degrees of freedom. Note that four lags means estimating 23 coefficients per equation.

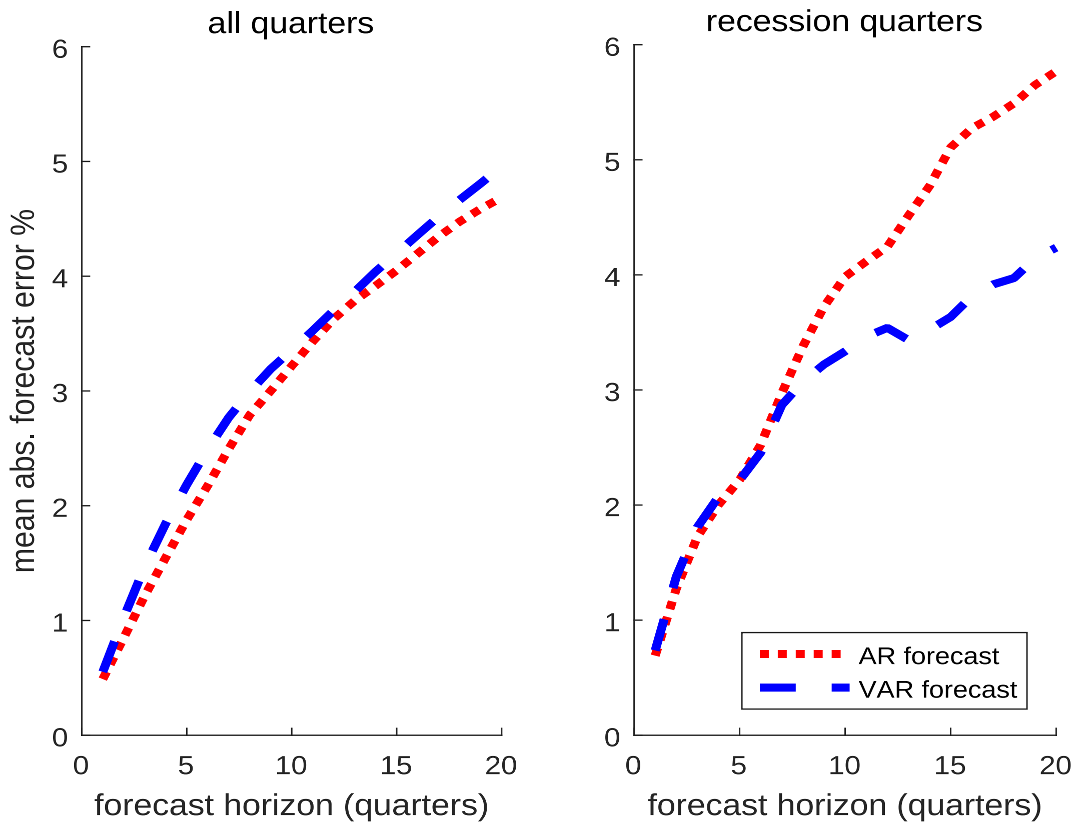

Figure 2.5: Average forecast errors, US GDP

Notes: These graphs plot the average forecat errors of the indicated time series model. NBER recession dates are used to determine whether a quarter is a 'recession quarter'.

very poorly, but continues to grow at a steady pace. The estimated multivariate models have fewer degrees of freedom and this seems to come at a cost during stable periods when simple forecasting rules suffice.

For the UK, the two time-series model generate forecast errors of similar magnitude even during economic downturns. The multivariate time-series model does generate more accurate forecasts, however, at the troughs of recessions. Below, we will discuss in more detail in which way UK recessions differ from US recessions.

2.5 Predictable US recoveries

In this section, we discuss in more detail the differences in forecasts of the univariate and the multivariate times-series model made at the trough of recessions.

Explaining the figures. Figures 2.6, 2.7, and 2.8 show the results for US recessions. The vertical lines in each figure indicate the forecasting point. The thick solid line plots the actual data. Each figure also plots the predicted growth path according to the two time-series models and a deterministic time trend.[19]

1973-75 US recession. The top panel of figure 2.6 displays the results for the 1973-75 recession.[20] Forecasts are made at the trough of the recession, 1975Q1. Forecasts from the univariate one-type-shock model indicate that output losses will be very persistent. Instead, there is a rapid recovery back to the long-term trend. Given that there are at times persistent changes in GDP, the univariate representation will always reflect this persistence to some extent.[21] By contrast, the forecast based on the multivariate model

[19] The time trend shown in the figures is a linear trend estimated on the full sample of GDP and is included as a point of reference. The linear-quadratic trends included in the univariate and multivariate models are estimated up until t^*.

[20] Because we focus on out-of-sample forecasts, we have only 109 quarterly observations for forecasts at the trough of this recession, which leaves few degrees of freedom when the VAR is estimated with the default specification, that is, four lags for each of the five variables and a quadratic deterministic trend. By using a VAR with only two lags for this recession, we avoid the strong sensitivity of forecasts when the forecasting date shifts slightly.

[21] However, since we use an $AR(4)$ to describe real output, our model does allow for a further predictable deterioration and/or for the possibility that (a large) part of the initial drop can be expected to be reversed.

captures the fast recovery of GDP after the trough of the recession. In addition to the predicted short-term increase in growth rates, the multivariate model also captures the subsequent return to normal growth rates. Not surprisingly, the path forecasted in 1973Q2 does not predict the recessions of the early eighties.

The exercise discussed here should not be considered as a horse race of two forecasting models. What the results show is that (i) some economic downturns are followed by faster than normal growth and seem to have little or no permanent effects and (ii) this type of pattern is unlikely to be predicted by univariate representations, whereas multivariate VARs do have the flexibility to capture this.

1980 US recession. The bottom panel of figure 2.6 displays results for the first recession of the early eighties. Forecasts are made at the trough, 1980Q3. Both models predict that the shortfall of GDP relative to its trend value observed in 1980Q3 will remain of roughly the same magnitude up till 1984. This means that both models miss the short-lived pickup in growth rates just after 1980Q3 and both miss the second recession in the early eighties. In 1984, the economy has recovered from the second recession, although GDP is still below its trend value, and GDP is in fact close to the levels predicted by both models using data up to 1980Q3.

The two 1980Q3 forecasts diverge in their predictions for the post-1984 period. The 1980Q3 forecast according to the univariate representation predicts that the gap between GDP and its (ex-post) trend value will not become smaller. By contrast, the 1980Q3 forecast based on the multivariate model indicates that the gap will become smaller, which is indeed what happened. In 1986, GDP was back to its trend value, which is in line with the 1980Q3 prediction according to the multivariate model.

The recovery predicted by the multivariate model in 1980Q3 is quite different from the recovery predicted in 1973Q2. Whereas, the multivariate model predicts a quick return at the trough of the seventies recession, it predicts a much more gradual return at the trough of the first early eighties recession.

1981-82 US recession. The top panel of figure 2.7 reports the results for the forecasting exercise when forecasts are made at the end of the second early-eighties reces-

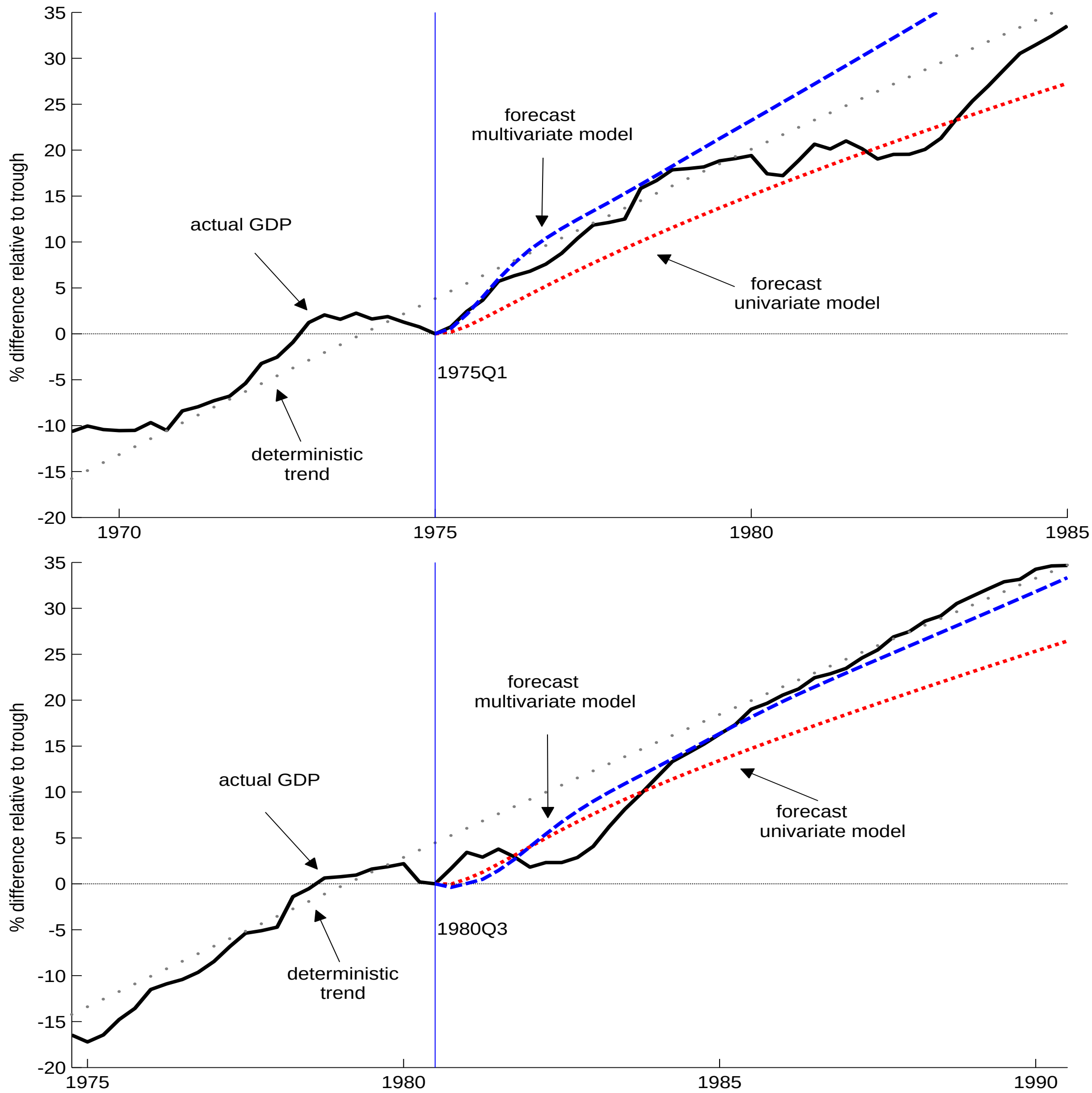

Figure 2.6: The 1973-5 and the 1980 US recessions
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

sion, 1982Q4. From this point onwards, the US economy recovers remarkably quickly. Whereas the economy is almost 9% below its (ex-post) trend level at the end of 1982, this gap is only 2.5% at the end of 1984 and only 1% at the end of 1985. The multivariate model captures this remarkable recovery very well. It does not capture, however, the fact that in subsequent years the gap gets even smaller. The univariate representation completely misses the recovery and predicts, again, that ground lost during the recession is permanent.

Both the behavior of GDP during this recession and the fact that the remarkable recovery can be predicted by a simple time-series model strongly suggest that it is not always the case that an unexpected change in real output of x percent should lead to a change of the long-term forecast of x percent.

Although our multivariate model is a simple VAR, with five variables and four lags, it allows for a rich set of dynamics. It is, therefore, not always easy to understand what features of the data lead to particular predictions. For this particular period, it is possible to point at the reason why the model predicts a sharp recovery. The period just before 1982Q4 is characterized by sharp drops in investment and exports. As documented in figure 2.4, these correspond to *temporary* reductions in GDP. Consequently, the multivariate model predicts that these negative influences will disappear quickly. During 1982, both consumption and government expenditures have started to grow already, which according to figure 2.4 correspond to permanent positive changes in GDP. This is consistent with the predicted persistence of the recovery.

1990-91 US recession. The bottom panel of figure 2.7 displays the results for the recession of the early 1990s. The results differ from those reported above for previous recessions in that now both models predict a permanent loss in GDP. Although the loss in actual GDP is indeed very persistent and GDP does not get back to its trend level until 1997, the actual loss is not permanent.

2001 US recession. The results for the early naughties recession are displayed in the top panel of figure 2.8. During this recession, there is not a sharp contraction in output. It is better characterized by a period of near zero growth rates. The recovery is also very

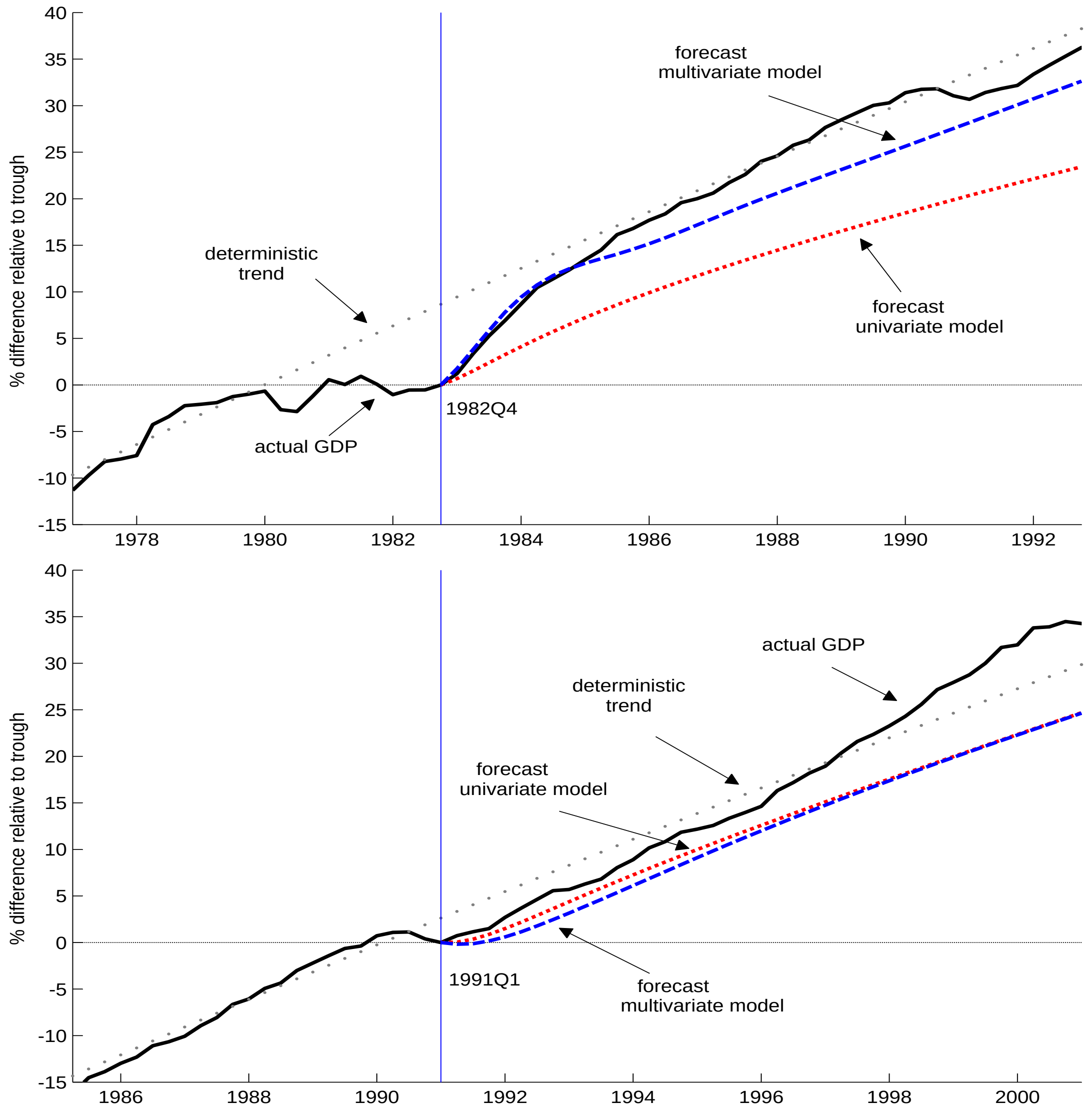

Figure 2.7: The 1981-2 and the 1990-1 US recessions

Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

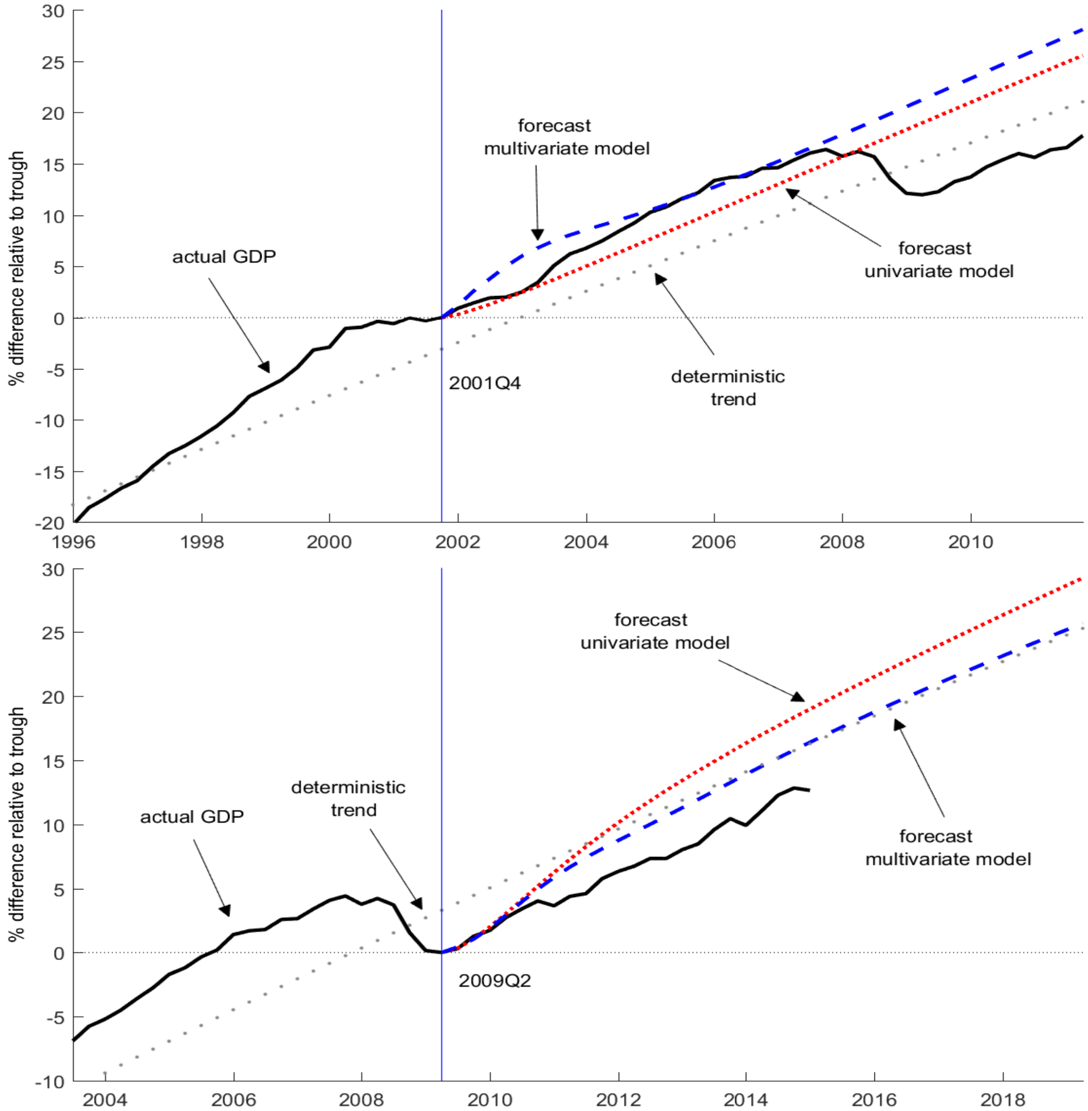

Figure 2.8: The 2001 and the great US recessions
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

gradual. The multivariate model is wrong in predicting a short-term pick up in growth rates, but is correct in its longer-term forecast that the loss in GDP is not permanent. The univariate representation predicts again that there will be no recovery, not in the short term, which in this case is indeed what happened, and also not in the long term, which is not what happened.

US financial crisis, 2008-2009 The bottom panel of figure 2.8 plots the results for the forecasts made in 2009Q2, when the sharp fall in GDP had come to a halt.[22] Similar to forecasts made in previous recessions, the multivariate model again predicts that part of the loss in output relative to trend will be recovered in a couple years. Different from forecasts made in previous recession is that the univariate now also predicts a recovery. In fact, at this point in time, the univariate model predicts stronger long-term growth than the multivariate model. Unfortunately, forecasts of both models were too optimistic.

Starting in 2012, the multivariate model starts to predict the future reasonably well. In particular, it correctly predicts that output loss relative to trend will not be reversed.[23] The univariate representation remains more optimistic than the multivariate model until the end of the sample, sometimes marginally more optimistic, but typically substantially more optimistic. Using data up to the end of our sample, the univariate model predicts that output in 2025 will be 1% below its extrapolated trend value whereas the multivariate model predicts that the gap will be 4.5%.[24]

Why are forecasts made with a univariate model too pessimistic? In section 2.2, we gave two reasons why univariate representations could be too pessimistic regarding the long-term impact of negative shocks. The common element in both reasons is that it is difficult for a univariate representation to generate the best possible forecast when the variable of interest is a sum of variables with different persistence.

The first reason focused on the case where the shocks affecting the aggregate where different shocks. Even the correct univariate representation has only one shock and would never be able to capture that there are actually multiple shocks that affect the aggregate for different lengths of time. The second reason focused on the case where the components are driven by the same shock, but the estimated univariate model is not complex enough.

[22] At the beginning of the financial crisis, both time-series models wrongly predict that a substantial part of the losses will be recaptured quickly. These results are not displayed in the graphs.

[23] These results are not displayed in the figures.

[24] The economy was substantially above its trend value before the crisis, which means that these long-term predictions imply larger losses relative to the hypothetical case when there would have been no financial crisis and subsequent average real output growth would have been equal to the trend growth rate.

Figure 2.4 showed that US GDP does consist of components with different degrees of persistence. Moreover, shocks to these components are clearly correlated. Nevertheless, we doubt that that the reason the univariate model generates different forecasts is that it is not complex enough. Our results are robust to alternative specifications and resemble those found in the literature for a variety of univariate representations. It seems more plausible to us that US GDP is affected by different types of events which affect the US economy for different durations. Univariate representations would not be able to capture this.

2.6 Predictable UK recoveries?

UK recessions before the financial crisis. Post-war UK recessions are not as interesting as US recessions. Instead of sharp contractions, like those observed for the US, UK recessions were typically prolonged periods of low growth rates. Similarly, recoveries were very gradual. Although the multivariate model has better long-term predictions than the univariate representation in all but one of the recessions that occurred before the financial crisis, the predictions of the two models are roughly similar. Moreover, forecasted paths are close to straight lines, which is not surprising given the shallow aspect of economic downturns in the UK. The exception to these observations is the financial crisis, which will be discussed next.

UK financial crisis, 2008-2010. Figures 2.9 and 2.10 plot the realizations of UK GDP together with forecasts made by the two models at four different forecasting points. First consider the two panels of figure 2.9, which plot the results when forecasts are made at the middle of the period with large negative growth rates, 2008Q4, and at the end of this period, 2009Q2.

In the middle of the period when GDP dropped sharply, the univariate representation predicts an immediate and sustained return to positive growth rates. It is even somewhat more optimistic than the prediction of a random walk model with drift in that it predicts that GDP will grow faster than its trend in the next couple years, that is, it predicts that part of the reduction of the pre-crisis positive gap between GDP and its trend value will

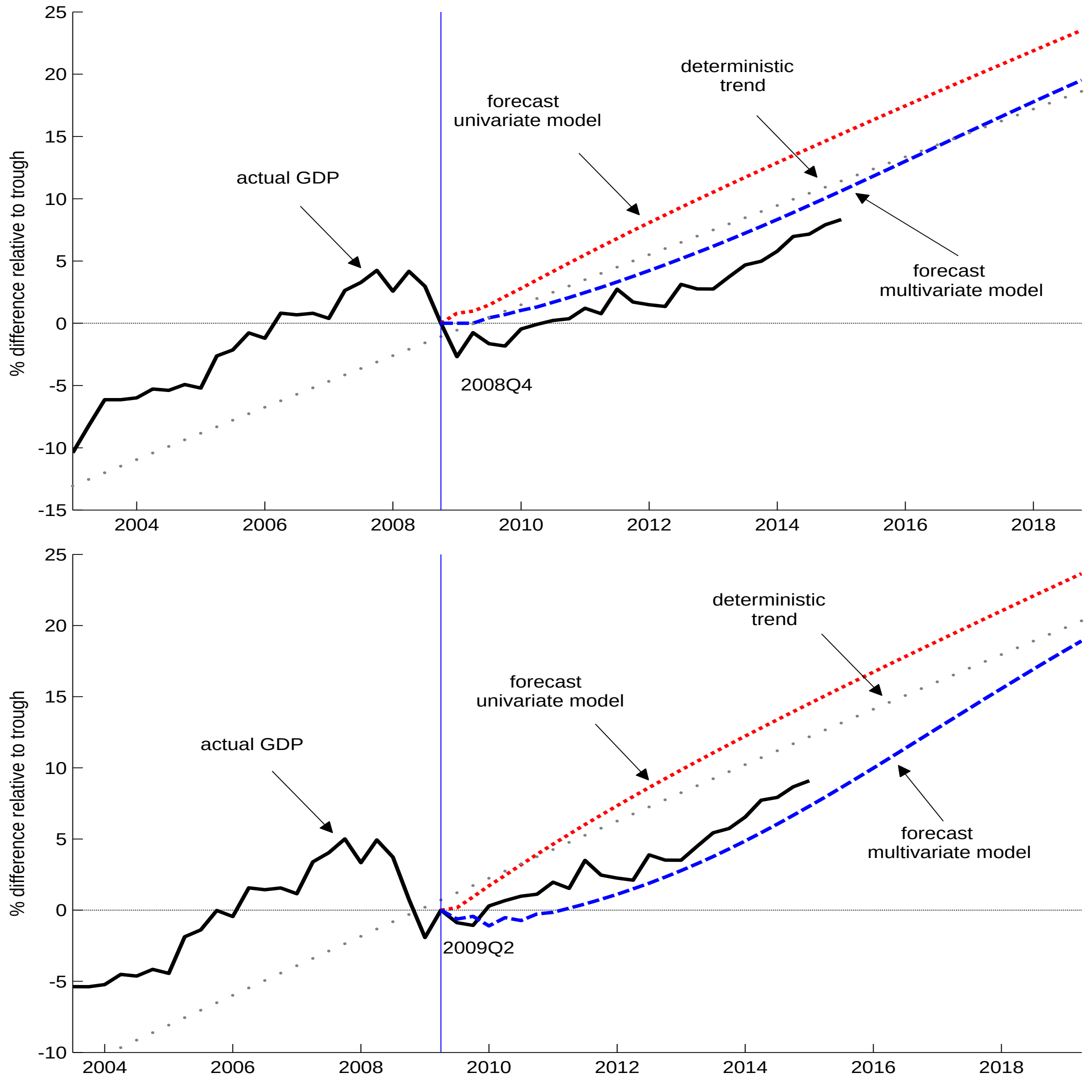

Figure 2.9: Start and trough of the great UK recession
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

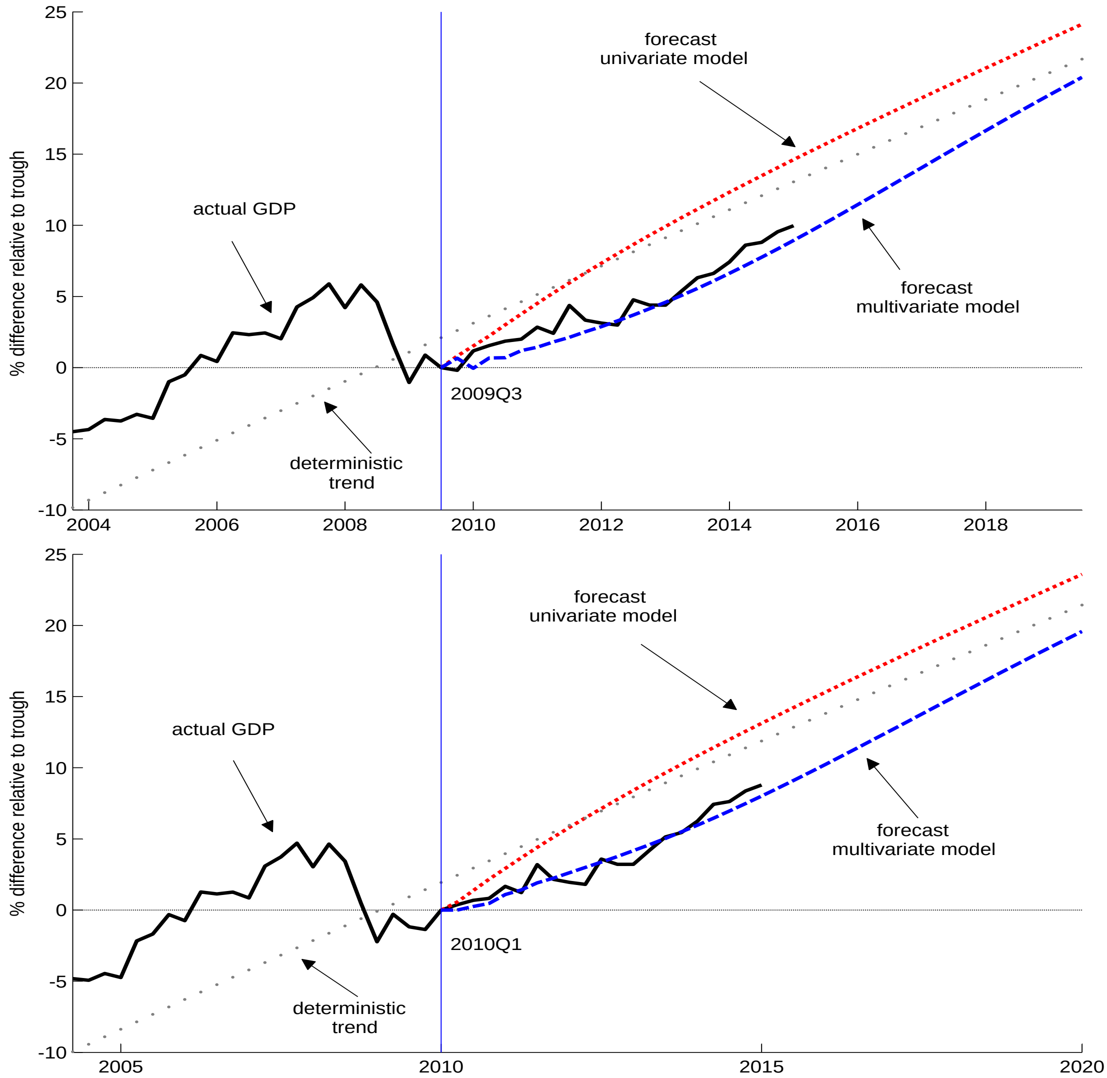

Figure 2.10: Start and trough of the great UK recession
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line.

be recovered. By contrast, the multivariate model predicts that GDP will grow at rates that are somewhat lower than the trend growth rate, which is closer to the observed outcomes, although also too optimistic. In 2009Q2, the univariate representation still predicts that GDP will end up substantially above its trend value. The multivariate model forecasts that growth rates would be around zero for several quarters followed by a very gradual recovery. These forecasts are slightly below the actual outcomes.

The two panels of figure 2.10 plot the results when forecasts are made in 2009Q3 and 2010Q1. Both of these quarters are in the period when the UK economy had just started its recovery. For both forecasting points, the univariate representation's predictions indicate that the economy will start growing at rates slightly higher than those observed in the past so that it still predicts that part of the losses will be recovered. By contrast, the multivariate model—using data up to 2009Q3—predicts that there first will be a period with low growth rates, which eventually is followed by a period of faster growth rates. This is indeed what happened, although the predictions are a little bit too pessimistic. Half a year later, in 2010Q1, the forecasts of the multivariate model have improved somewhat and do a good job in predicting the subsequent development of UK GDP.

We do not want to argue that the multivariate model is a remarkably good forecasting model. Neither model does very well in predicting subsequent output growth during this period, although it is worth noting that the multivariate model realizes quickly that output losses will be very persistent. The point that we want to make is that multivariate models have the flexibility to predict different types of forecasting patterns. By contrast, univariate representations are quite restrictive and may miss both predictable recoveries and—as is shown here—a predictable deterioration during a downturn. The main reason why the univariate representation is restrictive is that it has only one type of shock. Since the GDP data used to estimate the univariate representation contains a persistent component, changes in GDP will *always* lead to changes in the long-term forecasts of the univariate model. Although, univariate forecasts always have a permanent component, we allow for the possibility that short-term forecasts are different from long-term forecasts, since our empirical univariate representation has four lags. But all of our estimated univariate representations imply predictions that are quite close to those of a random walk with drift.

2.7 Concluding comments

Macroeconomic forecasts are made with simple univariate models, for example, Campbell and Mankiw (1987),[25] as well as with advanced multivariate models, for example, Stock and Watson (2002).

[25] More recently, Edge and Gurkaynak (2010) and Edge et al. (2010), show that the forecasting performance of estimated DSGE models can be worse than a simple forecast of a constant output growth.

In this paper, we reviewed reasons why univariate representations of a sum of random variables could miss key predictable aspects of this random variables. In fact, even if a random variable is a random walk, then that does not mean that there are no forecastable changes. In particular, if an aggregate consists of stationary and non-stationary variables, then the univariate representation will indicate that all shocks have permanent consequences even though that is, of course, not the case for shocks to the stationary components. Moreover, the correct specification of an aggregate of random variables could be quite complex. We argued that it might be better to estimate time-series models for the components and obtain forecasts for the aggregate by explicitly aggregating the forecasts of the components.

Despite the empirical observation that US GDP consists of very persistent and less persistent variables, the univariate and multivariate time-series model have similar forecasting performance in terms of average forecast errors. Such a finding may explain why forecasts based on univariate models are still taken seriously.

However, our simple multivariate time-series model clearly outperforms the univariate model, when it is used to forecast future GDP during recessions. Whereas the univariate model typically predicts that recessions have large and negative consequences, the multivariate model often correctly predicts that this is not the case. In some cases, for example, when the drop in GDP is mainly due to drops in components with less persistence such as investment and exports, it was possible to understand why the multivariate model performed better than the univariate model. In other cases it is not. Nevertheless, the sharply better performance of our simple multivariate model during recessions and the theoretical discussion indicate that one should be careful making forecasts with univariate time-series models.

One point that we do not address is the correct level of (dis)aggregation. Consumption is the sum of non-durable and durable consumption and both are sums of individual expenditures. So further disaggregation may lead to further improvements. It is not clear, however, whether one should disaggregate to the lowest possible level, since sampling variation typically increases when one considers disaggregated variables.

Appendices

2.A Data sources

US data. Data are downloaded from the web site of the Federal Reserve Bank of St. Louis. They are (i) Consumption: real personal consumption expenditures; (FRED code: PCECC96); (ii) Investment: real gross private domestic investment (GPDIC1); (iii) Government expenditures: real government consumption expenditures & gross investment (GCEC1); (iv) Exports: real exports of goods & services (EXPGSC1); and (v) Imports: real imports of goods & services (IMPGSC1). All time series are seasonally adjusted quarterly data measured in billions of chained 2009 dollars. The data were last updated May 29, 2015.

The GDP data used is the sum of the consumption, investment, government expenditures, and exports minus imports. Adding up these real time series generates a time series that is extremely close, but not exactly identical to the actual GDP data. Our approach ensures that the components used in the multivariate model add up *exactly* to the data used in the univariate model. This way, we avoid clutter by describing small differences in the GDP data used in the two types of time-series models.

UK data. Data are from the Office of National Statistics. They are (i) household final consumption expenditures (ONS code: ABJR) plus final consumption expenditure of non-profit institutions serving households (HAYO); (ii) total gross fixed capital formation (NPQT); (iii) general government: Final consumption expenditures (NMRY); (iv) balance of payments: Trade in goods and services: Total exports (IKBK); (v) Balance of payments: Imports: Trade in Goods and services (YBIM). All data are seasonally adjusted quarterly data and the base period is 2011. The GDP data used is the sum of these five components. Investment in inventories are excluded, since they contain some very volatile high frequency movements.

2.B Robustness

Figures 2.11 through 2.16 display the results for several robustness exercises. Figure 2.11 documents that our result that multivariate time-series models generate more accurate long-term forecasts than univariate models is also true when no deterministic trend term is included, when only a linear trend term is included, and when the number of lags are chosen by referenc to the Akaike Information Criterion (AIC). Figures 2.12 through 2.16 illustrate that even the actual forecasts are very similar when the number of lags are chosen with AIC. At the earlier forecasting dates, there is a bit of variation in the number of lags chosen by AIC, especially for the univariate specification. After this, the number of lags chosen for the univariate specification is three, which is one less than our benchmark number. For the multivariate specification, the number of lags remains two for a while and then jumps to five lags, one more than our benchmark number.

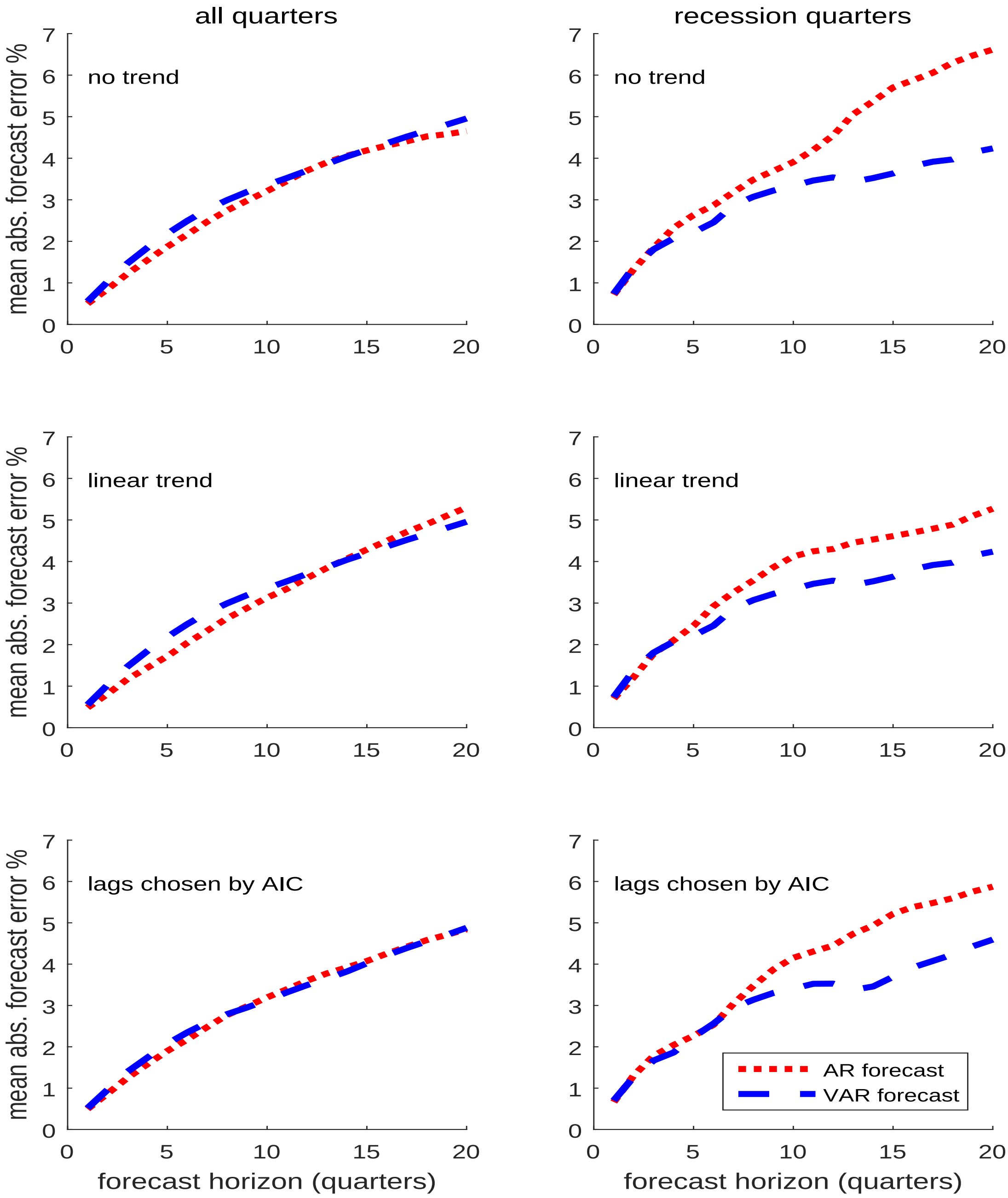

Figure 2.11: Average forecast errors, USA - robustness

Notes: These graphs plot the average forecast errors of the indicated time series model. NBER recession dates are used to determine whether a quarter is a 'recession quarter'

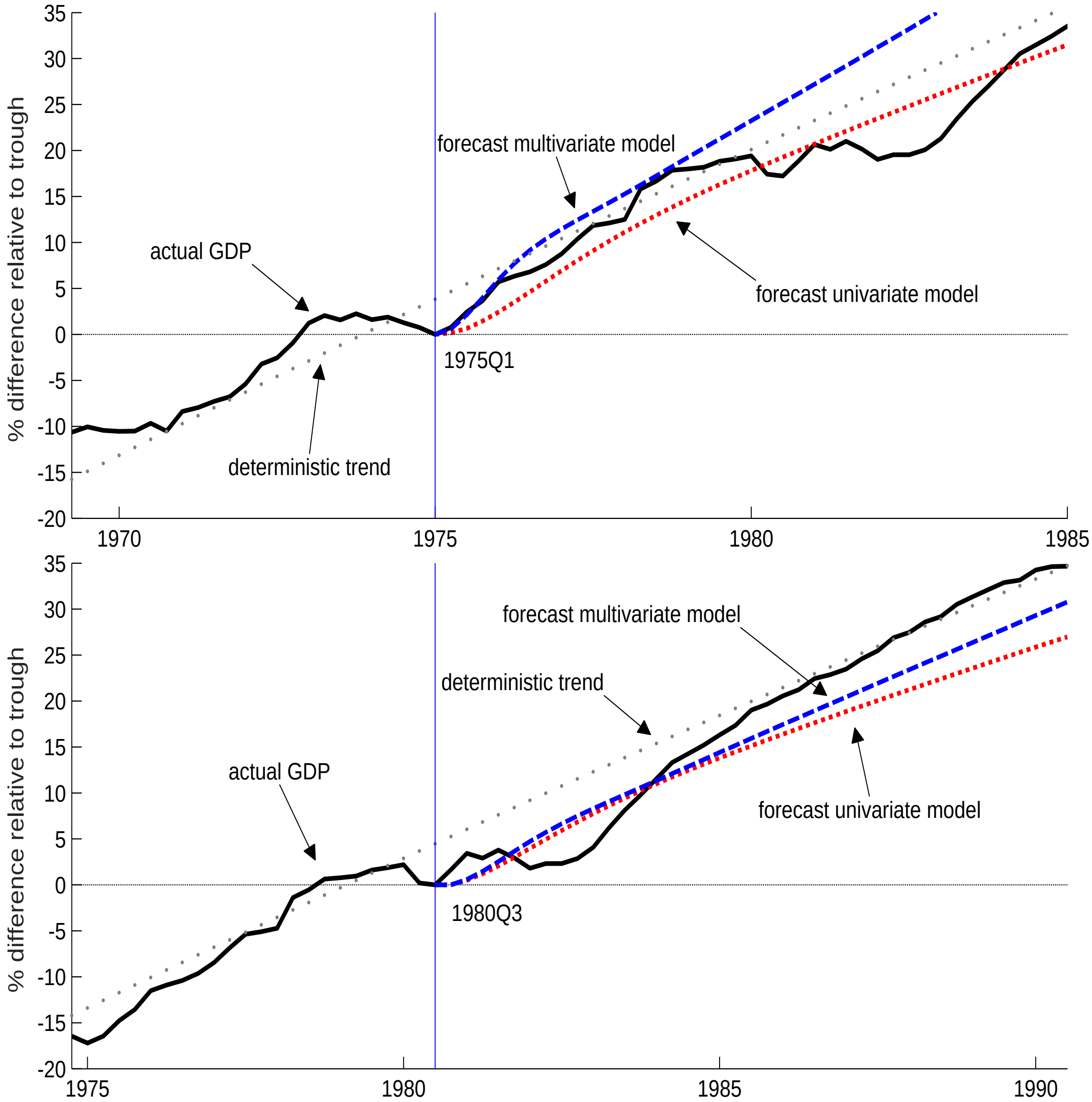

Figure 2.12: The 1973-5 and the 1980 US recessions-AIC
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line. Number of lags chosen using the Akaike Information Criterion.

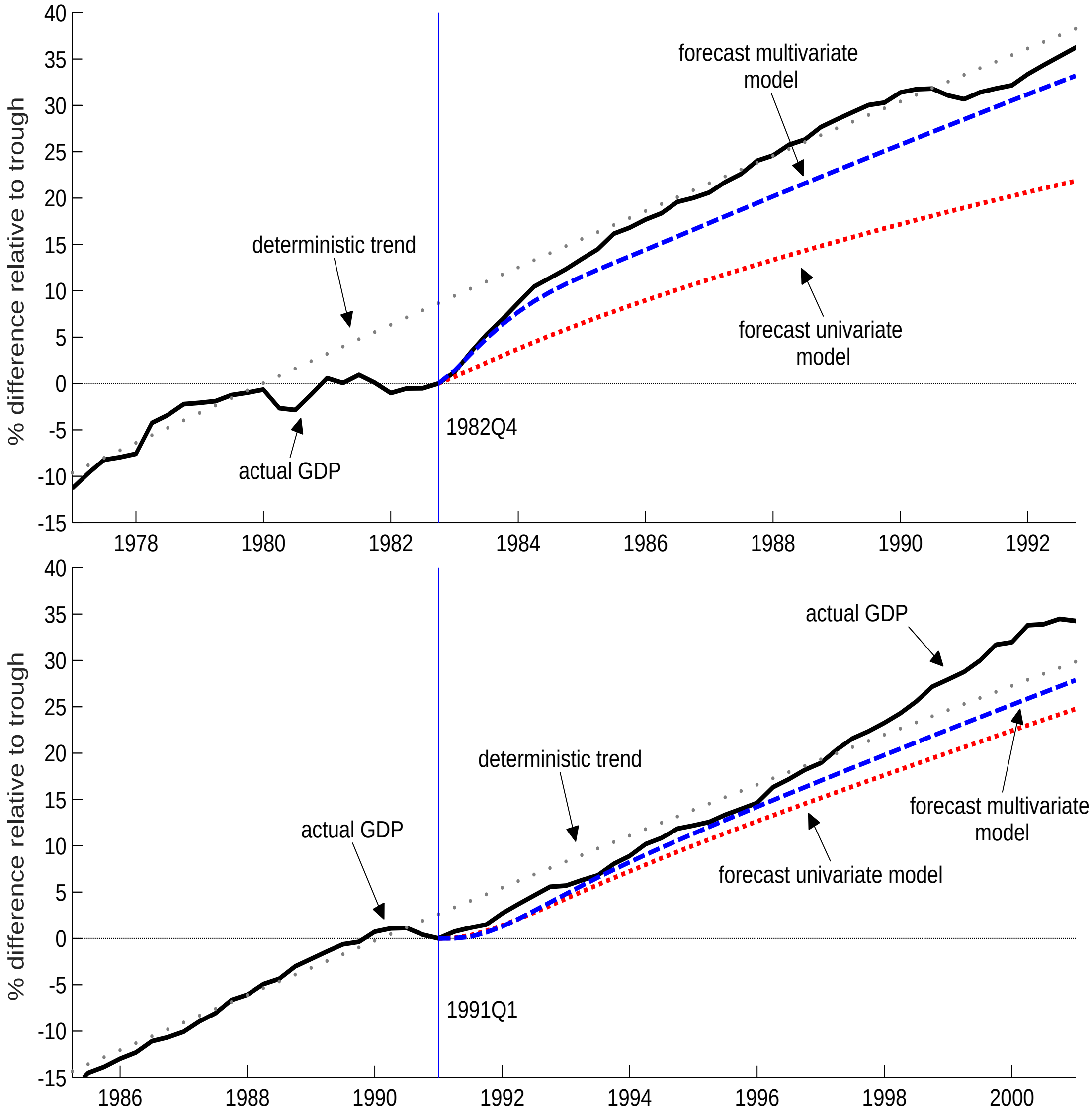

Figure 2.13: The 1981-2 and the 1990-1 US recessions-AIC
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line. Number of lags chosen using the Akaike Information Criterion.

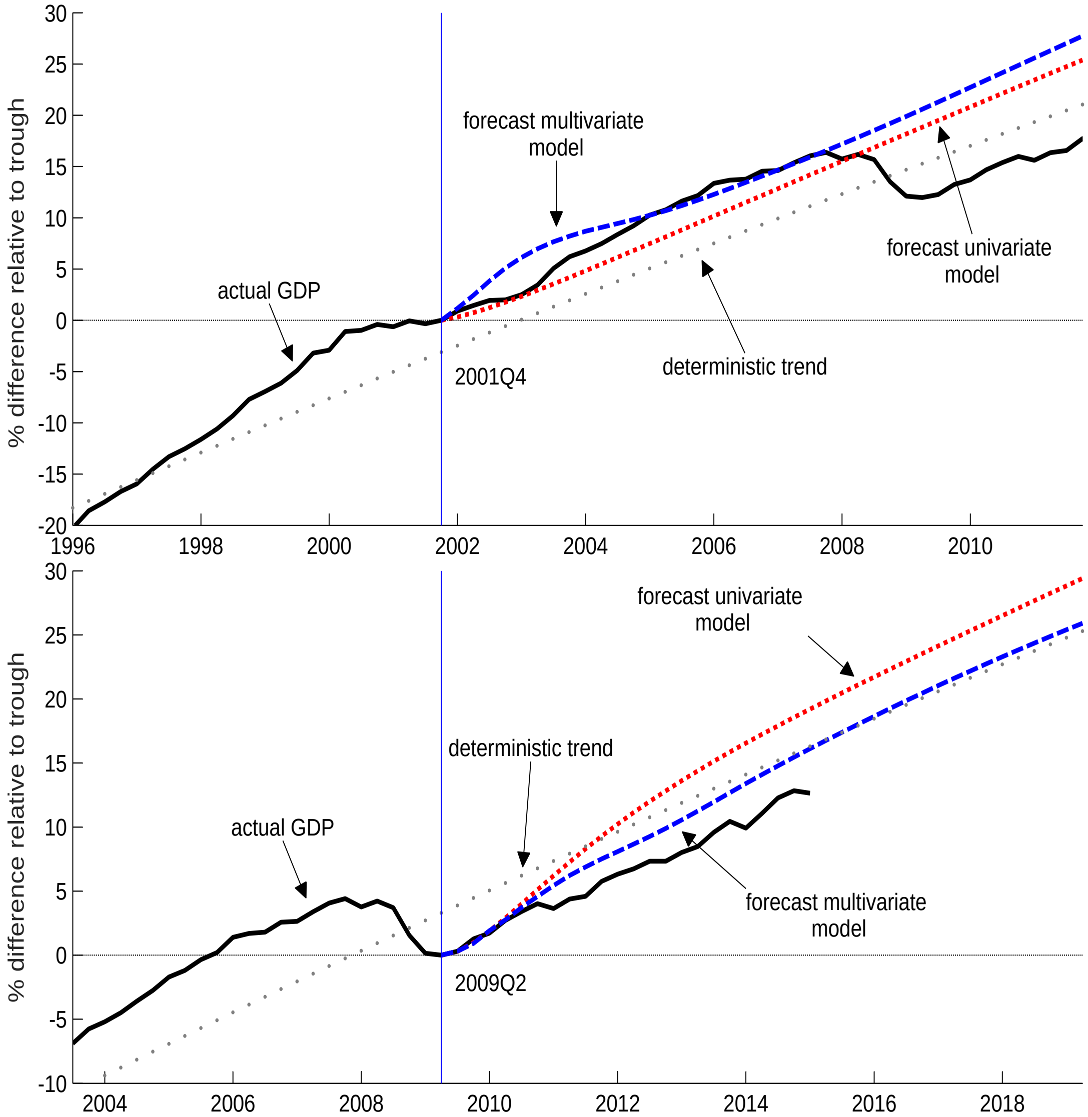

Figure 2.14: The 2001 and the great US recessions - AIC
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line. Number of lags chosen using the Akaike Information Criterion.

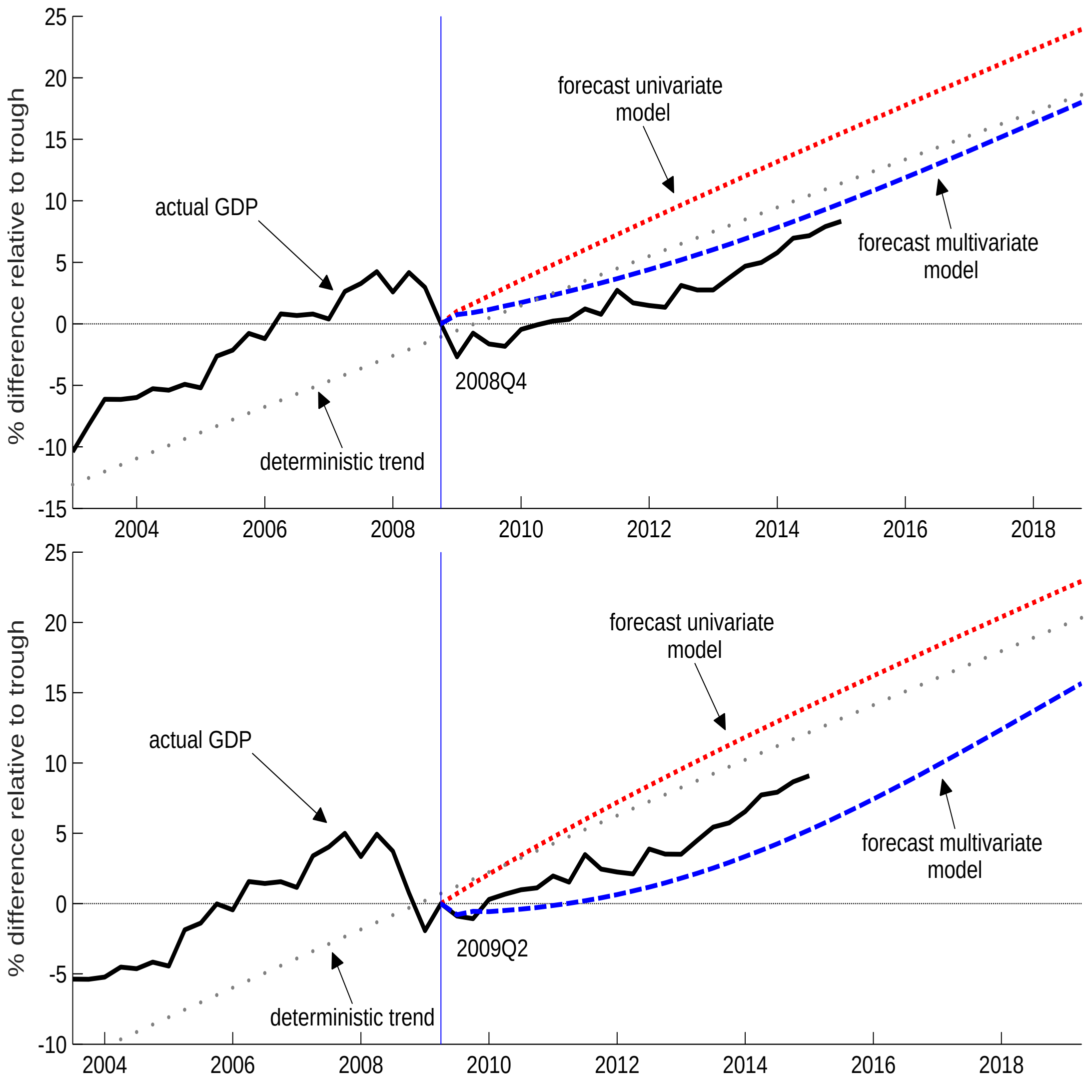

Figure 2.15: Start and trough of the great UK recession
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line. Number of lags chosen using the Akaike Information Criterion.

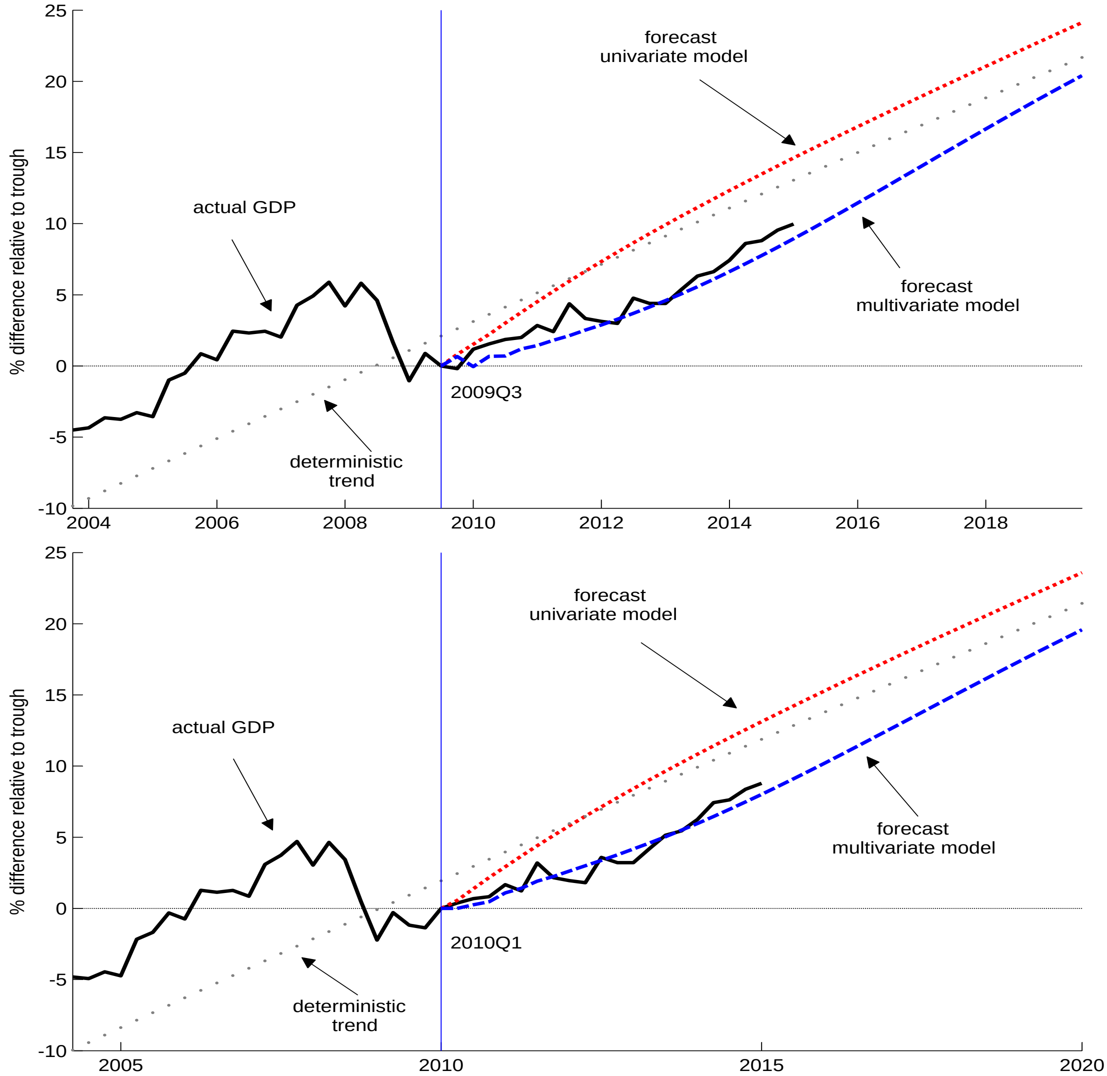

Figure 2.16: Start and trough of the great UK recession
Notes: This figure plots the two forecasted time paths for US GDP together with the realised values and a deterministic time trend. All four variables are relative to the value of GDP at the forecasting date, which is indicated by the vertical line. Number of lags chosen using the Akaike Information Criterion.

Chapter 3

Some Macroeconomic Consequences of Macroprudential Liquidity Regulation

3.1 Introduction

The financial crisis of 2008 has prompted a large overhaul of financial regulation, notably in the regulation of banks. The new Basel III regulations make two important qualitative changes. First, bank capital requirements, previously calculated with respect to risk-weighted assets, would be supplemented with a simple leverage ratio (which compares unweighted total assets with equity). Second, banks would have to hold a buffer of liquid assets (the "Liquidity Coverage Ratio" or LCR) which would cover at least the next 30 days of maturing market liabilities, as well as some proportion of retail deposits.

There are good reasons for thinking that both of these measures would either reduce the likelihood of a crisis in the financial sector, or reduce the severity of a crisis conditional on one occurring. However, such regulation may also have implications for the ability of banks to intermediate funds out of crisis periods. This is the question with which this paper is concerned. To this end, I set up a macroeconomic model to examine the effects (in general equilibrium) of changing liquidity regulation, which allows the effects of such a change, including those on prices, to be gauged.

I augment a standard neoclassical growth model with financial intermediaries similar to the type proposed by Gertler and Karadi (2011), and subject them to a "risk-weighted capital ratio" and a "liquidity ratio" which are meant to capture the essence of the most recent round of financial regulation. I then describe the mechanisms through which changing these constraints (though exogenous "policy") can affect the rest of the economy, which I term the "crowding out" channel and the "financial repression" channel. In the absence of either of these, I establish a neutrality result which shows that changes in liquidity regulation merely determine whether households hold bonds or bank deposits, and have no effect on the wider economy.

The intuition for the "crowding out" channel is as follows: suppose we start in a situation where banks have to adhere to a regulatory capital ratio which places a positive risk-weight on government bonds. Then suppose a regulatory change is introduced (for now, consider it to be unanticipated) such that banks are required to hold more of these bonds. As a consequence, they will have to reduce the amount of other assets that they hold for a given level of their own net worth. Over time, this change in the bank's portfolio mix will also typically reduce the rate at which banks accumulate net worth. This will further reduce the amount of productive assets that they can intermediate, and hence reduce the productive capital of the economy.

In contrast, the "financial repression" channel exists when banks earn a lower return - net of transaction costs - on their holdings of government bonds than they pay on their deposits. In this paper this is motivated by a positive resource cost of raising and servicing deposits.[1] As a consequence, forcing banks to hold more government bonds will require them to raise more deposits, increasing the amount of the resource cost that they have to pay and reducing the rate at which they accumulate net worth. This means that over time, banks will not be able to intermediate as much physical capital, reducing economy-wide output and consumption.

The neutrality result demonstrates that in the absence of either of these channels, the level of the liquidity ratio will simply determine whether households or banks hold gov-

[1]In the data, the interest rates on overnight retail bank deposits are typically below the yield on government bonds, which may offset such costs. In this paper I abstract from this difference as government bonds and bank deposits are both one period and riskless. It would be possible to extend the model to add "deposits in the utility function": in this case, the financial repression channel would still exist as long as the resource cost was greater than the difference in the interest rate spread between bonds and deposits.

ernment bonds, and have no effect on intermediation of productivie assets and hence other macroeconomic variables. This is true both in and out of steady state. This result is principally illustrative, as under these conditions banks are indifferent between holding any level of bonds financed out of deposits, and households treat bonds and deposits as perfect substitutes.

One clear policy implication from this paper is that if regulators require banks to hold liquid assets, these assets should not count towards capital or leverage ratios, to minimise the "crowding out" channel. Risk-weighted capital requirements have been criticised as allowing banks too much scope to "game" the regulatory system (for instance, see Admati and Hellwig (2013)). But if liquidity buffers consist of assets mandated by the regulator in any case, the scope for such gaming should be minimal, and would prevent the "crowding out" channel in the data.

This policy implication appears to be at odds with the new Basel III framework, and in particular the new "supplemental leverage ratio", as liquid assets held for regulatory reasons count towards this ratio. According to the Basel Committee's 2012 report on the implementation of the new framework, 56 of 209 global banks at the end of 2011 did not satisfy the new leverage ratio. So for a subset of banks, tighter liquidity requirements may lead to more "crowding out" of the type explained in this paper.[2]

The strength of the "financial repression" channel is harder to gauge. Insofar as banks compensate the costs of providing deposit services with a lower interest rate on deposits, a lower interest rate spread between government bonds and deposits (for example, due to the zero lower bound on nominal interest rates) may make this channel stronger. It would be better to assess the costs of deposit servicing separately, but this is a question for future research.

After discussing related literature, Section 3.2 describes the model. Section 3.3 discusses the conditions under which the "neutrality result" discussed above holds. Section 3.4 examines comparative statics of the model in steady state, while section 3.5 calibrates the model and presents some numerical results showing comparative statics and the transitional dynamics between regulatory regimes following a tightening of liquidity regulation.

[2]See: http://www.bis.org/publ/bcbs231.pdf. In the same report, 86% of LCR-eligible assets held by the 209 banks had a risk-weight of zero, so the "crowding out" channel would not appear to come from the risk-weighted capital requirements which were used under Basel II.

Section 3.6 concludes.

3.1.1 Related Literature

This paper uses financial intermediaries of the type developed and used by Mark Gertler, Peter Karadi and Nobu Kiyotaki in a number of papers (Gertler and Karadi (2011, 2013) Gertler and Kiyotaki (2010), Gertler et al. (2012) and Gertler and Kiyotaki (2013)). Of these, Gertler and Karadi (2013) is the closest to this paper from this strand of the literature, as it introduces a second asset (a long term government bond) to the Gertler and Karadi (2011) setup to study the effect of quantitative easing on the balance sheets of financial intermediaries. This paper is similar in that I also take Gertler-Karadi-Kiyotaki style financial intermediaries and add a second asset. However, this paper differs in three important ways. First, I substitute the "stealing" constraint facing financial intermediaries with a constraint designed to be interpreted as a risk-weighted capital ratio. This relates "risk-weighted assets" to bank net worth each period. Second, in this paper, banks also face a "liquidity ratio" or "reserve requirement". Thirdly, the second asset I include is a one-period government bond rather than a perpetuity. These changes allow for clearer results from comparative statics and also a clearer interpretation of the constraints.

The literature between financial regulation and macroeconomics has grown quickly since the financial crisis. There are a series of papers which look at the interaction between bank capital regulation, monetary policy and economic outcomes. These include Angeloni and Faia (2013), Farhi and Werning (2013) and Nuño and Thomas (2013). There is also a literature which examines liquidity and banking from a finite-horizon, partial equilibrium perspective. Notable recent contributions include Acharya et al. (2011), Acharya and Viswanathan (2011) and Allen et al. (2009).

There is a much smaller literature focusing on the effect of liquidity regulation on the financial sector. Particularly of note are papers by Goodhart et al. (2012) and Adrian and Boyarchenko (2013). Of these, the latter is most similar in spirit to this paper, in that it focuses on the same measure of liquidity (the Liquidity Coverage Ratio). Adrian and Boyarchenko (2013)) augment a continuous-time macro-finance model with financial intermediaries which are also subject to *ad hoc* capital and liquidity requirements. In their model, intermediaries can endogenously default. Their numerical results suggest that "tight" liquidity requirements and "loose" capital requirements typically maximise consumer welfare, as there is a stronger tradeoff between capital requirements and household consumption out of crisis periods than liquidity regulation. The paper by Goodhart et al. focuses on the welfare implications of the Net Stable Funding Ratio (NSFR), another addition to Basel III which attempts to manage liquidity of financial intermediaries over longer time horizons. They find that the NSFR can be a useful tool in improving consumer welfare.

Finally, there is a somewhat older literature which looks at the effectiveness of reserve requirements - which are similar to the recent liquidity regulations - as ways for the government to reduce their effective interest burden. Using an overlapping generation model of banks, Freeman (1987) shows that reserve requirements are typically worse from a welfare point of view than a tax on deposits, due in part to the "crowding out" of capital that reserve requirements imply.

3.2 Model Environment

3.2.1 Banking sector

The banking sector in this model is similar to that in Gertler and Karadi (2011) and subsequent papers. Banks have net worth n_t and raise deposits from households. With this they purchase capital (which they rent to firms) and one-period government bonds. In each period, bankers 'die' at rate $1-\sigma$, where the survival probability σ is low enough such that bankers are constrained by their net worth around the steady state.

Formally an individual bank's problem is:

$$\max_{\{b^b_{t+i}, d^b_{t+i}, k^b_{t+i}\}_{i=0}^{\infty}} \mathbb{E}_t \left[\sum_{i=1}^{\infty} \Lambda_{t,t+i}(1-\sigma)\sigma^{i-1} n_{t+i} \right]$$

subject to ($\forall t$):

$$n_t = R^k_t q_{t-1} k^b_{t-1} + R^f_{t-1} b^b_{t-1} - (R^d_{t-1} + \delta_c) d^b_{t-1}, \tag{3.1}$$

$$q_t k_t^b + b_t^b = n_t + d_t^b, \tag{3.2}$$

$$q_t k_t^b + \chi b_t^b \leq \phi n_t, \tag{3.3}$$

$$b_t^b \geq \pi d_t^b, \tag{3.4}$$

where b_t^b, d_t^b, k_t^b correspond to bond holdings, deposit liabilities and holidings of physical capital, respectively.[3] n_t is the bank's equity at t. R_t^k, R_t^f, R_t^d correspond to the gross rate of return on physical capital, bonds and deposits respectively, and q_t is the price of capital at t.

The bank's objective is to maximise the terminal value of n_t upon exit in terms of the marginal utility of households (as most of exiting bank net worth is redistributed lump-sum to households, and so net worth is discounted using the household's stochastic discount factor Λ, defined in the next section). Bank net worth evolves according to (3.1) and bank portfolio choices must be consistent with their period balance sheet constraint (3.2). (3.1) states that bank net worth at t is equal to the gross return from assets less the cost of repaying deposits. The gross return from bonds and deposits will come from the household's problem, and the gross return on capital is defined as:

$$R_t^k = \frac{r_t^k + (1-\delta)q_t}{q_{t-1}}, \tag{3.5}$$

where r_t^k is the rental rate on capital and δ is the depreciation rate. The only non-standard part of these constraints is the fact that I allow banks to face a period cost of raising deposits (δ_c). This reflects costs of raising and managing deposits, and I assume it is linear in the amount of deposits raised.

However, this paper deviates from the Gertler-Karadi-Kiyotaki model by assuming that banks face a regulatory capital ratio (equation (3.3)) rather than a stealing constraint. This constraint states that the risk-weighted sum of assets can be no greater than some multiple of net worth. Here, the parameter $\chi \in [0, 1]$ is analagous to the "risk-weight" applied to government bonds. $\chi = 1$ is analagous to a simple leverage ratio, while $\chi = 0$

[3] The superscript 'b' refers to the fact that the variable pertains to banks, so k_t^b is holdings of physical capital by banks at time t.

is equivalent to a zero risk-weight on government bonds. Implicitly, the risk weight on physical capital is normalised to one - this need not be the case in practice but then the "leverage" parameter ϕ would need to be adjusted accordingly.

The final constraint (3.4) is a liquidity ratio or reserve requirement: it states that banks have to hold liquid assets in proportion to their deposits. The proportion is determined by the parameter π. This is a simple way of attempting to construct a constraint in the spirit of Basel III's Liquidity Coverage Ratio (LCR). Strictly speaking, the Liquidity Coverage Ratio would set $\pi = 1$ for all market debt coming due in the next 30 days and zero for all other market debt.[4] However, in this model both deposits and government bonds have a one period maturity. A richer model would allow for variation of bank liability by (i) type and (ii) maturity. In this framework we abstract from this for simplicity.

The rationale for setting up the bank's problem with (3.3) and (3.4) rather than a stealing constraint consistent with Gertler and Karadi (2013) is twofold. First, the purpose of this paper is to attempt to clarify the effect of financial regulation (in particular, liquidity regulation) on the macroeconomy. (3.3) and (3.4) have a clear interpretation in terms of regulation. The second reason is that this particular setup will give clean results in terms of the comparative statics.

Solution to the bank's problem

Define the cost to the bank of raising a unit of deposits $\widetilde{R}_t^d = R_t^d + \delta_c$. The following proposition characterises the solution to the banks problem:

Proposition 1 *If $\pi \geq 0$ and $\chi \geq 0$, and if banks take prices $(R_{t+1}^k, R_t^f, \widetilde{R}_t^d)$ as given, banks maximise their present discounted value of net worth subject to (3.1)-(3.4) and if the following conditions hold in period t:*

$$(1-\pi)\mathbb{E}_t\left[R_{t+1}^k\right] + \pi(R_t^f) \geq \widetilde{R}_t^d, \tag{3.6}$$

$$\chi(\mathbb{E}_t\left[R_{t+1}^k\right] - \widetilde{R}_t^d) \geq (R_t^f - \widetilde{R}_t^d), \tag{3.7}$$

[4]Retail deposits are assumed to have a low "runoff" rate, with $\pi = 0.05$ or $\pi = 0.10$ depending on their type. See Basel Committee on Banking Supervision (2013) for details

then the regulatory capital ratio (3.3) and the liquidity ratio (3.4) bind.[5]

Proof: *See Appendix 3.A.1*

Condition (3.6) states that the weighted return on assets for a bank is at least as great as the cost of deposits. If this holds, then it then it will be optimal for the bank to hold as many (risk-weighted) assets as possible, and (3.3) will bind.[6]

Similarly, the second condition (3.7) implies that - holding constant risk-weighted assets ($q_t k_t^b + \chi b_t^b$) - banks will always choose to hold as few government bonds as possible, and hence (3.4) will bind. Intuitively, selling a bond and repaying a deposit will reduce risk-weighted assets by χ and returns by $(R_t^f - \widetilde{R}_t^d)$. To keep risk weighted assets constant, the bank can raise some deposits and buy χ units of physical capital, on which they will earn the (expected) return $(\mathbb{E}_t\left[R_{t+1}^k\right] - \widetilde{R}_t^d)$. Under condition (3.7), this swap will not decrease the banks returns (and will increase it if the inequality is strict).

In what follows we always assume that conditions (3.6) and (3.7) are satisfied. This is reasonable as regulations are set in the real world to influence bank behaviour, and in this model this is achieved through the regulatory constraints being binding.[7],[8]

Given this, the constraints (3.2)-(3.4) will bind and mechanically will solve for the optimal k_t^b, b_t^b, d_t^b. Note that the bank's policy functions will be linear in their n_t, which will allow aggregation. In particular, the policy functions are given by:

$$q_t k_t^b = \psi^k n_t = \frac{\phi(1-\pi) + \pi\chi}{1-\pi+\pi\chi} n_t, \tag{3.8}$$

$$b_t^b = \psi^b n_t = \frac{\pi(\phi-1)}{1-\pi+\pi\chi} n_t, \tag{3.9}$$

$$d_t^b = \psi^d n_t = \frac{\phi-1}{1-\pi+\pi\chi} n_t, \tag{3.10}$$

[5] Note that (3.6) and (3.7) are weak inequalities. If they hold exactly, it is also possible for the bank to set (b_t^b, d_t^b, k_t^b) such that (3.3) and/or (3.4) don't bind without reducing their objective function. In these boundary cases, I assume that the bank chooses (b_t^b, d_t^b, k_t^b) such that (3.3) and (3.4) bind.

[6] I show in the appendix that (3.6) and (3.7) together imply that the leverage ratio (3.3) binds even if the liquidity ratio (3.4) doesn't.

[7] If risk was added to the model and there was a penalty for breaching the requirements, it would be the case that the requirements would matter even if they did not normally bind.

[8] Additional evidence that current regulations bind can be seen from the large excess returns of Bank stocks following the unexpected victory of President Trump in November 2016, as documented by Wagner et al. (2017). Much of the financial reporting at the time focussed on the possible relaxation of post-financial crisis regulation which was seen to impede bank profitability.

Examination of these policy functions already hint at the effect of different policies - higher π appears to lead to more bonds being held by banks and less capital, for instance - but this ignores the general equilibrium effects of changing each parameter on bank net worth in equilibrium. To examine the effect of changing parameters we first have to set out the remainder of the model.

3.2.2 Household sector

Households supply labour inelastically, consume and save in one of three assets - bank deposits, government bonds, and physical capital (which they rent to firms). Importantly, households incur a management cost each period for holding physical capital. Formally, the household's problem is:

$$\max_{\{c_{t+i},b^h_{t+i},d^h_{t+i},k^h_{t+i}\}_{i=0}^{\infty}} \mathbb{E}_t \left[\sum_{i=0}^{\infty} \beta^i u(c_{t+i}) \right]$$

subject to a budget constraint and "no-shorting" contstraints, $\forall t$:

$$c_t+b^h_t+d^h_t+q_t k^h_t+f(k^h_t) = w_t+R^f_{t-1}b^h_{t-1}+R^d_{t-1}d^h_{t-1}+R^k_t q_{t-1}k^h_{t-1}-T_t+\pi^b_t+\pi^f_t+\pi^k_t, \quad (3.11)$$

$$d^b_t \geq 0, \quad (3.12)$$

$$k^b_t \geq 0, \quad (3.13)$$

Here, c_t is household consumption, b^h_t, d^h_t, k^h_t are household holdings of government bonds, bank deposits and physical capital respectively. Labour income is denoted by w_t, lump sum taxes by T_t, and $(\pi^b_t, \pi^f_t, \pi^k_t)$ are profits remitted to households by banks, final good firms and capital producing firms respectively. π^f_t will be zero in every period in this model, and π^k_t will be zero in steady state.

The household's problem yields the following first order conditions:

$$1 = \mathbb{E}_t \left[\Lambda_{t,t+1}\right] R^f_t, \quad (3.14)$$

$$1 = \mathbb{E}_t \left[\Lambda_{t,t+1}\right] R_t^d, \tag{3.15}$$

$$1 + f'(k_t^h) = \mathbb{E}_t \left[\Lambda_{t,t+1} R_{t+1}^k\right], \tag{3.16}$$

where $\Lambda_{t,t+1}$ is the stochastic discount factor between t and t+1:

$$\Lambda_{t,t+1} = \frac{\beta u'(c_{t+1})}{u'(c_t)}, \tag{3.17}$$

It is clear from (3.14) and (3.15) that households treat government bonds and bank deposits as perfect substitutes, and so a necessary condition for households to hold both is that $R_t^d = R_t^f$. The expected return on physical capital adjusted for aggregate risk and net of the marginal management costs $f'(k_t^h)$ must also equal the risk free rate if the household holds positive amounts of each asset in equilibrium.

The function $f(.)$ is strictly increasing and convex in the capital held by each household. It ensures that banks have a role in this economy - without it, banks would be as efficient as households in intermediating capital, and hence banks would have no role[9]. For numerical applications of the model in Section 4 onward, I will need to specify a functional form for $f(.)$:

$$f(k_t^h) = \frac{\gamma_h}{2}(k_t^h - \overline{k}^h)^2 \tag{3.18}$$

where $\overline{k}^h$ is a parameter governing how much of the capital stock households are allowed to hold costlessly.

3.2.3 Final good firms

Final good firms hire labour from households and rent capital from households and banks. They have a constant returns to scale production function which ensures that capital is paid its marginal product and labour is paid the remaining revenue. These firms make zero profit period by period. Formally their period problem is:

$$\max_{k_t^f, l_t^f} \pi_t^f = z_t (k_t^f)^\alpha (l_t^f)^{(1-\alpha)} - r_t^k k_t^f - w_t l_t^f$$

[9]Note that this function can be parameterised to nest the case in which households hold no capital at all.

The first order conditions for this problem, together with the fact that $l_t^f = 1$ in equilibrium imply that:

$$r_t^k = \alpha z_t (k_t^f)^{\alpha-1}, \text{ and} \tag{3.19}$$

$$w_t = (1-\alpha) z_t (k_t^f)^{\alpha}. \tag{3.20}$$

3.2.4 Capital producing firms

The capital producing firms in this economy are the same as that in Gertler and Karadi (2011). There are convex adjustment costs in net investment, which allows the price of capital to deviate from unity out of steady state. At the end of each period, capital producing firms purchase capital from households and banks, and then repair depreciated capital and build new capital, which they sell at price q_t.

Net investment and investment are defined as:

$$I_t^n = (K_t^b + K_t^h) - (K_{t-1}^b + K_{t-1}^h), \tag{3.21}$$

$$I_t = (K_t^b + K_t^h) - (1-\delta)(K_{t-1}^b + K_{t-1}^h), \tag{3.22}$$

where K_t^b and K_t^h are aggregate capital stocks held by banks and households, respectively, and steady state investment is denoted by I^{SS}. The fact that the flow adjustment cost depends on net investment ensures that the capital decision is independent of the market price of capital. The problem of capital producing firms is:

$$\max_{I_t^n} \mathbb{E}_t \left[\sum_{i=0}^{\infty} \Lambda_{t,t+i} \Big((q_{t+i} - 1) I_{t+i}^n - g(x_{t+i})(I_{t+i}^n + I^{SS}) \Big) \right],$$

where $g(.)$ is a convex adjustment cost in the flow of net investment, such that $g(1) = g'(1) = 0$, $g'(.) > 0$ and $g''(.) > 0$, and $x_{t+i} = \frac{I_{t+i}^n + I^{SS}}{I_{t-1+i}^n + I^{SS}}$. This problem gives the following first order condition:

$$q_t = 1 + g(x_t) + x_t g'(x_t) - \mathbb{E}_t \left[\Lambda_{t,t+1} x_{t+1}^2 g'(x_{t+1}) \right]. \tag{3.23}$$

These firms make zero profit in steady state, as $q_t^{SS} = 1$. Profits made out of steady state are rebated lump sum to households, and are defined by:

$$\pi_t^k = \Big((q_t - 1)I_t^n - g(x_t)(I_t^n + I^{SS})\Big).$$

3.2.5 Government

The government levies lump sum taxes T_t on households, and uses it to fund a constant amount of wasteful spending proportional to steady-state output (gY^{SS}) and service the interest on its outstanding stock of debt (B_t). The government's budget constraint is given by:

$$g + R_{t-1}^f B_{t-1} = B_t + T_t. \tag{3.24}$$

As taxes are lump sum the timing of taxes is not important to analysis of the model. However, to specify a path for T_t I assume that the government sets taxes with the following rule:

$$T_t = g + \psi_g(B_t - B^{SS}), \tag{3.25}$$

where $\psi_g > 0$ is a parameter and B^{SS} is the government debt level in the non-stochastic steady state.[10]

3.2.6 Market Clearing and Aggregation

As the policy functions for banks (3.8)-(3.10) are linear in net worth, we can sum over all banks such that:

$$K_t^b = \psi^k N_t,$$

$$B_t^b = \psi^b N_t,$$

$$D_t^b = \psi^d N_t,$$

[10]Note that throughout this paper I assume that parameters and shocks are such that $B_t^b < B_t$. I could relax this if I allow the household sector to short government bonds.

Where uppercase letters represent aggregate variables corresponding to the relevant lowercase symbol. In each period fraction $1-\sigma$ of banks are forced to exit. Their net worth on exiting is split between two sources: (most) is rebated lump-sum to households but some is transferred to entering banks of mass $1-\sigma$ (so that the mass of banks remains at unity over time). To be precise, profits rebated to households and transfers to entering banks respectively are:

$$\pi_t^b = (1-\sigma)[R_t^k q_{t-1} K_{t-1}^b + R_{t-1}^f B_{t-1}^b - (R_{t-1}^d + \delta_c) D_{t-1}] - N_t^{new}, \tag{3.26}$$

$$N_t^{new} = \omega \phi N_{t-1}. \tag{3.27}$$

Hence, the equation of motion for aggregate banking sector net worth is given by:

$$N_t = \sigma[R_t^k q_{t-1} K_{t-1}^b + R_{t-1}^f B_{t-1}^b - (R_{t-1}^d + \delta_c) D_{t-1}] + \omega \phi N_{t-1}, \tag{3.28}$$

where ω is a parameter denoting how much of the assets of exiting banks are remitted to entering banks. The market clearing conditions in the capital, government bond and deposit markets respectively are:

$$K_t^f = K_{t-1}^b + K_{t-1}^h, \tag{3.29}$$

$$B_t = B_t^b + B_t^h, \tag{3.30}$$

$$D_t^b = D_t^h. \tag{3.31}$$

Finally, aggregate output and the resource constraint are given by:

$$Y_t = z_t (K_t^f)^\alpha, \tag{3.32}$$

$$Y_t = C_t + I_t + g + f(k_t^h) + \delta_c D_t^b, \tag{3.33}$$

where $log(z_t)$ follows an AR(1) process with persistence parameter ρ_z.

3.3 A neutrality result

Proposition 2 *If* $\pi \in [0, 1)$ *and (3.6) and (3.7) are satisfied, and* $\chi = 0$ *and* $\delta_c = 0$ *then changes in the required liquidity ratio,* π*, alter* b_t^b, b_t^h*, and* $d_t^b = d_t^h$ *but have no effect on the rest of the economy.*

Proof: *See Appendix 3.A.2*

This neutrality result is interesting as under certain conditions it shows that changes in bank liquidity regulation need not have an effect on the wider economy, even in a model in which banks (i) are constrained by their net worth, and (ii) are better in intermediating capital than households. By studying changes which violate this neutrality result (in particular, allowing the "risk-weight" on bonds to be positive $\chi > 0$ or by assuming that banks find it costly to raise deposits $\delta_c > 0$) we can get a sense about the mechanisms through which liquidity regulation may affect the wider economy.[11]

To get some intuition behind this result consider the representative bank's constraints (3.1)-(3.4):

$$n_t = R_t^k q_{t-1} k_{t-1}^b + R_{t-1}^f b_{t-1}^b - (R_{t-1}^d + \delta_c) d_{t-1}^b,$$

$$q_t k_t^b + b_t^b = n_t + d_t^b,$$

$$q_t k_t^b + \chi b_t^b = \phi n_t,$$

$$b_t^b = \pi d_t^b,$$

where we assume that the conditions stipulated in Proposition 1 hold and so the two inequality constraints bind. Now consider an increase in π. For a fixed amount of deposits (d_t^b) the bank would have to buy bonds - which would mean selling some physical capital. However, if $\pi < 1$ then banks can also adjust to the higher liquidity requirement by raising deposits and using purchasing additional bonds, without selling any physical capital. If $\chi = 0$ then bonds have a zero risk weight, and so banks can buy bonds without increasing their risk-weighted assets - and as such, they don't have to reduce their holdings of physical capital. If in addition, $\delta_c = 0$ (banks don't face a resource cost in raising deposits), then increasing deposits and bonds one for one does not alter the

[11] It may be worth noting that under the neutrality result, banks are indifferent to the amount of government bonds they hold. As such, this case may not be realistic, but may still help understanding of the mechanisms involved.

rate at which banks accumulate net worth either. In this case, then, changing π does not alter anything apart from the size of the bank's balance sheet.

As the model is in general equilibrium, the neutrality result goes through only because households treat government bonds and bank deposits as perfect substitutes. As such, the banking system as a whole faces a perfectly elastic supply curve of bonds and bank deposits, and this is true even if the aggregate supply of bonds is held constant. If households did not treat bonds and deposits as perfect substitutes, then changes in π (which alter the amount of deposits and bonds that households hold in equilibrium) would have to be compensated by changes in their relative returns. In particular, higher π would require households to hold fewer bonds and more deposits in equilibrium. This would require R^d to increase relative to R^f if bonds and deposits were less than perfect substitutes. Hence extending this model to allow deposits to enter the utility function (for instance) would mean that this result would not in general be preserved.

The idea behind the proof is that in general, the equations of the model detailed in Section 3.2 need to be solved as a system. However, if $\chi = 0$ and if $\delta_c = 0$ we can show that this system reduces into two "blocks": each containing a subset of the equations listed above. Crucially, the system can now be solved recursively, with the first block solved without reference to the second. As the parameter π only appears in the second block, then it will only affect variables in this block. The second block contains three equations which pin down three variables (b_t^b, b_t^h, and $d^b = d_t^h$), which correspond to the representative household's holdings of bonds and deposits. If either $\chi > 0$ or if $\delta_c > 0$ then at least one of these three variables will appear in the first "block" (in the equations relating to the bank's constraints) and hence this recursive structure will be broken.

3.4 Comparative Statics

The principal purpose of this paper is to construct a theoretical framework for analysing the macroeconomic consequences of liquidity regulation. To this end, I first use comparative statics to illustrate how the endogenous variables in the model respond to changes in the parameter π (which determines the amount of government bonds that banks have to hold as a fraction of deposits).

3.4.1 Parameter restrictions

In steady state, the prices R^k, R^f and R^d will be known functions of parameters. To ensure that the representative bank will still find it optimal for their regulatory capital ratio (3.3) and their liquidity ratio (3.4) to bind, I now place weak assumptions on parameters which (i) will ensure that the conditions for Proposition 1 will be satisfied, which will pin down bank behaviour; and (ii) will allow interpretation of some of the results of this section.

Assumption A1: Restrictions on parameters (Interpretation)

(i) $\phi > 1$: Banks can raise a postive amount of deposits

(ii) $\pi \in [0, 1)$: Banks are not required to hold more bonds than the amount of deposits they raise

(iii) $\chi \in [0, 1]$: The "risk-weight" on government bonds in the bank's "capital ratio" constraint (3.3) is no greater than the "risk-weight" on physical capital.

(iv) $\delta_c \geq 0$: The resource cost of raising deposits is non-negative.

(v) $\frac{1-\omega\phi}{\sigma} \geq \frac{1}{\beta} + \delta_c \frac{1}{1-\pi}$: the steady state rate of net worth accumulation for the banking sector is sufficiently high. This condition is necessary for banks to want to raise deposits.

Conditions A1(i)-(iv) have natural interpretations and ensure that banks 'look like' those in reality, by raising deposits, holding some fraction of these in liquid assets and so on. I show in Appendix 3.A.3 that together with these, additionally assuming A1(v) will be sufficient to ensure that the two condtions needed for banks to want to raise deposits ((3.6) and (3.7)) are satisfied and that the bank is optimising when both the regulatory capital ratio (3.3) and the liquidity ratio (3.4) bind. This last condition ensures the rate of return on the portfolio of bank assets are higher that the return on deposits (inclusive of any management costs, δ_c) which is a necessary condition for banks to want to raise deposits and not simply invest out of their net worth.

3.4.2 Evaluating the steady state

In the steady state of this model, productivity z is a constant and labour is supplied inelastically so changes in output will come from a change in the aggregate capital stock K^f. By the representative final good firm's first order condition (3.19), the amount of capital in steady state must be consistent with the rental rate of capital r^k and hence the gross return on physical capital R^k (by (3.5), which adjusts the rental rate by depreciation and changes in the price of physical capital) . Hence to conduct a comparative statics exercise of changing one parameter on output, we must first find how that parameter affects R^k. First, note that (3.28) and the bank policy functions (3.8)-(3.10) imply:

$$N_t = \sigma[R_t^k \psi^k + R_{t-1}^f \psi^b - (R_{t-1}^d + \delta_c)\psi^d]N_{t-1} + \omega\phi N_{t-1},$$

In the non-stochastic steady state where N is constant, this can be written:

$$\frac{1-\omega\phi}{\sigma} = [R^k \psi^k + R^f \psi^b - (\widetilde{R}^d)\psi^d], \tag{3.34}$$

Where the right-hand side of this expression are the parameters that pin down the steady-state return on equity of a bank which survives (i.e. isn't forced to exit). The consumer's first order conditions pin down the return on deposits and bonds as follows:

$$R^f = \frac{1}{\beta}, \text{ and} \tag{3.35}$$

$$\widetilde{R}^d = \frac{1}{\beta} + \delta_c. \tag{3.36}$$

The steady state R^k is implicitly defined by (3.34). This states that the return on capital - and hence the rental rate to capital - in steady state must be such that the aggregate net worth of the banking sector is constant. This is conditional on the return on bond and deposits given by the household's first-order conditions.

3.4.3 Effect of a change in the liquidity ratio parameter (π) on the steady state return on physical capital (R^k)

By rearranging (3.34) and differentiating we obtain the following expression describing how the return on capital changes as a function of changes in the required liquidity ratio in steady state:

$$\frac{dR^k}{d\pi} = \frac{\phi - 1}{(\phi(1-\pi) + \pi\chi)^2}\left[\left(\frac{1-\omega\phi}{\sigma} - \frac{1}{\beta} - \delta_c\right)\chi + \delta_c\phi\right]. \tag{3.37}$$

By examining the expression (3.37) we can see that under A1 $\frac{dR^k}{d\pi} \geq 0$.[12] This states that increasing the liquidity ratio of banks will never reduce the steady state return of captial. The following two channels can be seen clearly in (3.37):

- There is a possible **"crowding out"** effect, corresponding to the first term inside the square brackets in (3.37). If $\chi > 0$ then increasing the amount of bonds that banks have to hold reduces the amount of capital they can finance for a *given* amount of net worth. Over time, holding less capital will erode the rate at which surviving banks accrue net worth as well.

- There is also a **"financial repression"** effect: if $\delta_c > 0$ then a higher π reduces the rate at which surviving banks accrue net worth in steady-state. This is because more deposits are having to fund government bonds which have a lower return than the cost to the bank of raising deposits.

We can also see that these two channels interact to dampen the effect of the other (so that if both "crowding out" and "financial repression" are present the overall effect on R^k is less than the sum of the two effects individually). Taking the cross partial derivative of the above expression with respect to χ and δ_c:

$$\frac{\partial^2}{\partial\chi\partial\delta_c}\frac{dR^k}{d\pi} = -\frac{\phi - 1}{(\phi(1-\pi) + \pi\chi)^2}\left[1 + 2\frac{(\phi-\chi)\pi}{(\phi(1-\pi) + \pi\chi)}\right]. \tag{3.38}$$

This expression is strictly negative, which implies that the rate at which an increase in one

[12] The first term in (3.37) is positive by A1(i). In the square brackets, the first term is non-negative by A1(ii), A1(iii) and A1(v); while the second term is non-negative by A1(i) and A1(iv).

of (χ, δ_c) increases the steady state sensitivity of the cost of capital to the liquidity ratio parameter is decreasing in the level of the other. The intuition behind this "dampening" is that a higher value of χ reduces the amount of additional deposits which can be raised to in response to an increase in π, which in turn reduces the effect that a higher δ_c has on deposits.

Note also that we can clearly see the neutrality result discussed above in (3.37): $\chi = \delta_c = 0$ is clearly a sufficient condition for $\frac{dR^k}{d\pi} = 0$. Together with the additional assumption that:

$$\frac{1 - \omega\phi}{\sigma} > \frac{1}{\beta}$$

(which rules out the case when $R^k = R^d = R^f$), $\chi = \delta_c = 0$ is both necessary and sufficient for changes in π to be neutral with respect to the real economy.

3.4.4 Effect of a change in the liquidity ratio parameter (π) on steady state bank net worth (N)

Combining (3.8) and (3.29) gives:

$$k^f - k^h = \psi^k N. \tag{3.39}$$

We can think of the right-hand side of this expression as the supply of physical capital held by banks, with the left-hand side as firm demand for physical capital net of household supply. Totally differentiating this with respect to π:

$$\left[\frac{dk^f}{dR^k} - \frac{dk^h}{dR^k}\right]\frac{dR^k}{d\pi} = \frac{d\psi^k}{d\pi}N + \psi^k\frac{dN}{d\pi} \tag{3.40}$$

From (3.5), (3.19), it is straightforward that $\frac{dk^f}{dR^k} < 0$ - that a higher equilibrium return on capital implies less physical capital in steady state. Similarly, from (3.16) it is straightforward to infer that $\frac{dk^f}{dR^k} > 0$ - a higher R^k means that households will hold more capital as they are able to bear a higher management cost. As from Proposition 2 we know that $\frac{dR^k}{d\pi} \geq 0$ we know that the left-hand side of (3.39) is non-increasing in π.

Turning to the right-hand side of (3.39):

$$\frac{d\psi^k}{d\pi} = -\frac{(\phi - 1)\chi}{(1 - \pi + \pi\chi)^2} \leq 0$$

Together with the above result, this indicates that the sign of $\frac{dN}{d\pi}$ is ambiguous under A1. Putting aside the case of neutrality for the moment, higher π will increase R^k, which will reduce the demand for physical capital held by banks, which would indicate that there would be a lower N in steady state. But a higher π directly affects the amount of k that banks can hold for a given amount of N. Which of these two effects dominates is not clear for general parameter values.

Under A1, there are two special cases of note:

- $\frac{dN}{d\pi} = 0$ if $\chi = \delta_c = 0$. Under these parameter values $\frac{dR^k}{d\pi} = 0$. Note also that $\chi = 0 \implies \frac{d\psi^k}{d\pi} = 0$

- $\frac{dN}{d\pi} < 0$ if $\chi = 0$ and $\delta_c > 0$. Under these parameter values: $\frac{dR^k}{d\pi} > 0$. As $\chi = 0 \implies \frac{d\psi^k}{d\pi} = 0$

The first of these cases corresponds exactly to the neutrality result discussed above. The second case shows that if we deviate from the neutrality case by introducing a resource cost of raising deposits (but leaving the risk-weight on bonds at zero) then N would decrease if π rises, as the extra deposits needed to fund higher b would be more costly than the return on that b.

3.4.5 Effect of a change in π on steady state output and consumption

The aggregate resource constraint in steady state can be written:

$$Y = C + I + g + f(k^h) + \delta_c D^b$$

Note that $Y = (K^f)^\alpha$. In turn, K^f is determined by R^k. Hence, changes in π (the liquidity ratio) will impact output only though changes in the return on capital, and hence the steady state capital stock.

It is also the case that $I = \delta K^f$ and $D^b = \psi^d N$ so, substituting these expressions into the resource constraint and differentiating:

$$\begin{aligned} \frac{dC}{d\pi} &= (\frac{dY}{dK^f} - \frac{dI}{dK^f})\frac{dK^f}{dR^k}\frac{dR^k}{d\pi} - f'(k^h)\frac{dk^h}{dR^k}\frac{dR^k}{d\pi} - \delta_c(\frac{d\psi^d}{d\pi}N + \psi^d\frac{dN}{d\pi}) \\ &= (\alpha(K^f)^{\alpha-1} - \delta)\frac{dK^f}{dR^k}\frac{dR^k}{d\pi} - f'(k^h)\frac{dk^h}{dR^k}\frac{dR^k}{d\pi} - \delta_c\frac{d\psi^d}{d\pi}N - \delta_c\psi^d\frac{dN}{d\pi} \quad (3.41) \end{aligned}$$

Note that the first term is non-positive - the term inside parentheses must be positive under A1(v) given the definition of R^k. From proposition 2 we know that $\frac{dR^k}{d\pi} \geq 0$ and the firm's first order condition implies that $\frac{dK^f}{dR^k} < 0$.

The second term represents the increase in management costs paid by households as a consequence of the change in π. Again, it is non-positive, and strictly negative if $\frac{dR^k}{d\pi} > 0$. Intuitively, if an increase in R^k induces households to hold more capital, which means they pay a higher resource cost.

The fourth term reflects the increase in the resource costs of raising deposits for a given amount of N. This term is also non-positive, and strictly negative if $\delta_c > 0$ and $\chi < 1$. $\chi < 1$ is a sufficient condition for $\frac{d\psi^d}{d\pi} = \frac{(\phi-1)(1-\chi)}{(1-\pi+\pi\chi)^2} > 0$.

The final term is the change in the resource cost of raising deposits as a consequence of a change in steady state N. This term is of ambiguous sign. It is therefore possible that an increase in π can raise aggregate consumption (as higher π may reduce the amount of deposits held, and hence the amount of resources which need to be spent on raising deposits).

3.4.6 Comparative statics: conclusion

To sum up, in cases where the neutrality result doesn't hold, increases in the liquidity ratio π will raise the rate of return on physical capital (and hence the rental rate of capital) in steady state. This will directly lead to lower levels of output. This is even the case if steady state bank net worth increases (bank net worth may increase in steady state, depending on the relative strength of the "crowding out" and the "financial repression" channel).

Household consumption may fall by more than the fall in output, particularly if the increase in π leads to households holding a larger share of the capital stock (on which they have to pay a higher management cost). Against this, a lower steady state capital stock implies less investment is needed in steady state. In addition, if bank net worth falls it may be that the amount of deposits in steady state falls as well, which means that amount of resources spent by banks raising deposits will fall. These two effects act in the opposite direction to the higher management costs paid by households, which means that the fall in consumption may be either greater or smaller than the steady state fall in output.

3.5 Numerical Analysis

The analytical results in Section 3.4 illustrate the channels through which changes in the liquidity ratio parameter π can affect the steady state return on capital (R^k) and hence the rest of the economy. However, to gauge the quantitative importance of these channels we need to either calibrate or estimate the model's parameters. In this section, I take the first approach.

3.5.1 Calibration

Table 1 contains a list of the parameters in the model and the values of each in the baseline calibration.

The financial sector (bank) parameters are set principally by referring to US data. ϕ and

π are set by referring to historical averages in the US data using the Federal Reserve's H8 data on commercial banks. I calibrate the model to allow a 300 basis point annual spread between R^f and R^k in steady state, which approximately corresponds to the spread between US government bonds and US commercial paper. Together, this is sufficient to calculate $\frac{1-\omega\phi}{\sigma}$. The calibration of the remaining financial parameters is more suggestive. As in Gertler and Karadi (2011), σ is set so that banks survive a certain duration - here 15 periods, or just under 4 years- which allows us to recover ω.[13]

The baseline calibration assumes that $\chi = 1$ and $\delta_c = 1\%$ p.a., corresponding to a positive resource cost from raising deposits and a simple leverage ratio (in which all assets are weighted equally). These are suggestive and in the experiments below I will vary each of these by moving them "towards" the neutrality result.

I calibrate the household's management cost function parameter γ_h so that household's management costs are 1% of steady state GDP in the baseline calibration. $\overline{k}^h$ is calibrated to hit a target of households owning 2/3 of the capital stock, which corresponds to the fact that in the US household holding of corporate debt (from the flow of funds) is approximately twice as large as the stock of commercial bank lending to firms (from the Federal Reserve Board's H8 tables).

The remaining parameters in the model are standard given that the model period is a quarter - namely $\alpha, \beta, \delta,$ and ρ_z. The consumer's utility function is taken as log (which will matter for the dynamics of the model, though not for the steady state). The government spending parameter g is picked so that the ratio of government purchases to steady state output in the baseline calibration is 0.2, which corresponds to the share of government purchases in US GDP.

3.5.2 Numerical comparative statics

Figure 3.1 assesses the degree to which the "crowding-out channel" and the "financial repression channel" identified in (3.37) matter quantitatively. Three scenarios are considered:

[13]Choosing ϕ, π and the spread to be consistent with US data requires that the survival horizon for banks is shorter than that assumed in Gertler and Karadi (2011).

Table 3.1: Calibrated parameter values

Parameter	Value	Description
Banks		
χ	1	Risk weight on government debt
δ_c	1%p.a.	Cost of raising deposits
ϕ	10	Steady state leverage
π	0.25	Liquidity ratio parameter
σ	0.933	Survival rate of banks
ω	0.003	Governs transfers to new banks
Capital Goods Firms		
γ_k	0.5	Parameter governing adjustment in aggregate capital
Final Goods Firms		
α	0.4	Curvature of production function
δ	0.02	Depreciation rate of physical capital
ρ_z	0.95	AR(1) term in productivity process
Government		
ψ_g	1	Sensitivity of taxes to govt debt
g	0.9870	Government spending
Households		
β	0.99	Discount factor
γ_h	5.2 x 10^{-4}	Scaling parameter in management cost function
$\overline{k}^h$	41.7	Physical capital that households can hold costlessly

- The baseline calibration (solid blue line) in which both the "crowding out" and "financial repression" channels are present ($\chi = 1$, $\delta_c = 1\%$p.a.)

- A second scenario (dotted red line) in which the "crowding out" channel is present, but the "financial repression" channel is not ($\chi = 0$, $\delta_c = 1\%$p.a.)

- A third scenario (dashed green line) in which the "crowding out" channel is absent, but the "financial repression" channel is present ($\chi = 1$, $\delta_c = 0\%$p.a.)

The first panel (R^k) is expressed in annualised levels, while the remaining panels (corresponding to N, Y, C) normalise the variable by the value of that variable in the baseline calibration.

The solid (blue) line shows how each of the endogenous variables (R^k, N, Y, C) changes as π varies. A 5 percentage point increase in π from its baseline calibration leads to an 19 basis point (bp) increase in R^k on an annualised basis, a 5% fall in the aggregate net worth of the banking sector and a 0.8% fall in the steady state level of output and consumption. It is worth noting that the fall in consumption is roughly the same as the fall in output, as although R^k implies higher household management costs (for owning more capital), these are roughly offset by the fall in deposit raising costs for banks. It is also worth noting that in this suggestive calibration the steady-state output cost of tigher liquidity regulation is relatively large.

The "crowding out only" scenario (dotted and dashed red line) has sharply different levels of the endogenous variables to the baseline case. Reducing δ_c from 1% p.a. to nothing has a large negative impact on R^k and large positive impact on the level of N, Y and C. The slope of the red line is somewhat smaller than that of the blue line in three out of four cases (R^k, Y and C) indicating that the effect of a change in π would be somewhat ameliorated by the absence of the financial repression channel. The level of banking sector net worth is non-monotonic in π, first increasing and then decreasing, consistent with the analytical results in the previous section that $\frac{dN}{d\pi}$ is of ambiguous sign when $(\chi > 0)$.

The "financial repression only" scenario (dashed green line) has similar levels of the endogenous variables to the baseline case. However, the slope of the green line is noticeably flatter than that of either the blue or the red line in three cases (R^k, Y, C). This implies that the absence of "crowding out" makes the effect of a given change in π noticeably smaller on these variables. This suggests that a large part of the effect of a change in π is from the "crowding out channel".

3.5.3 Transition to a new steady state

So far the analytical and numerical analysis has focused on discussion of the steady state of the model under differing parameters. However, it may also be of interest to focus on the effect of the transition from one regulatory regime to another. This section analyses such a transition.

The economy starts in steady state according to one of the three scenarios explained in the previous section ("baseline" (solid blue), "no financial repression" (dotted and dashed

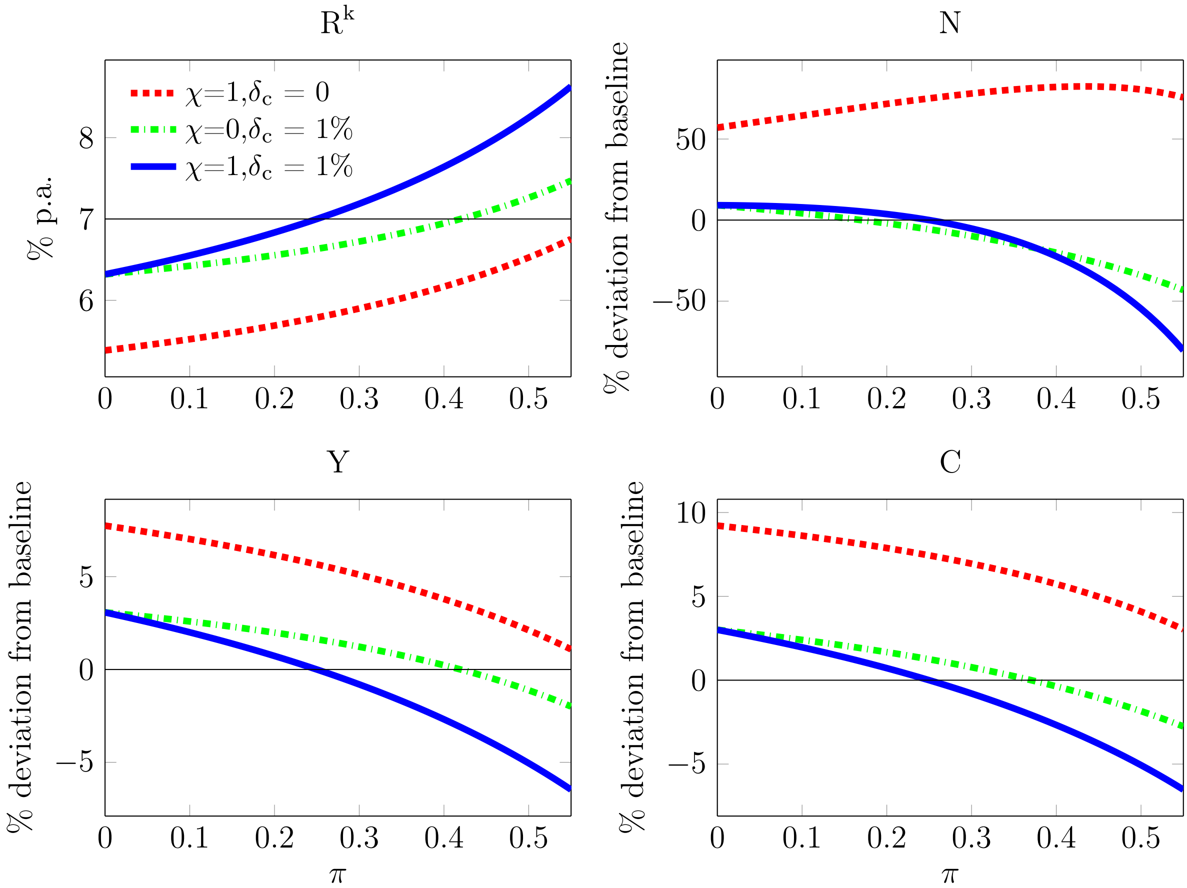

Figure 3.1: Effect of varying π on steady state

Notes: The dotted line refers to a "financial repression" only channel, the dotted-and-dashed line refers to a "crowding out" only channel and the solid line refers to the combination of the two channels.

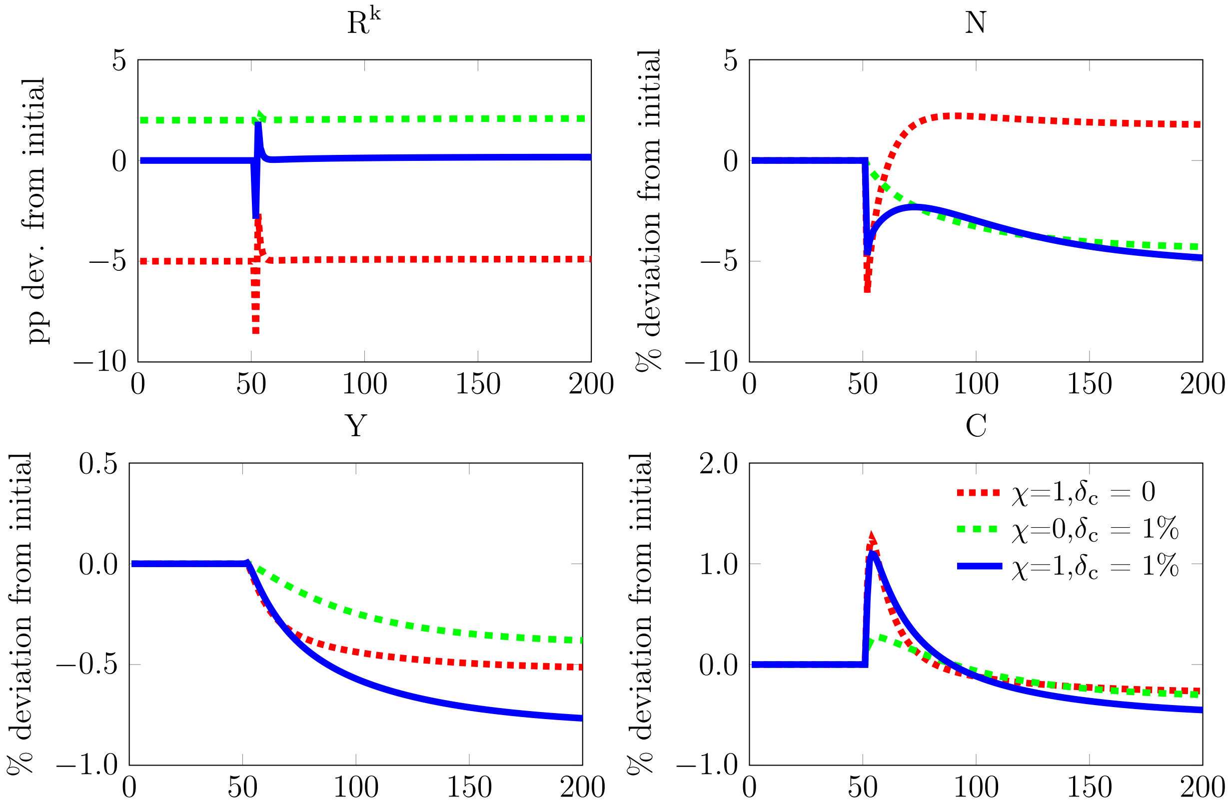

Figure 3.2: Effect of unexpected permanent increase π from $\pi = 0.25$ to $\pi = 0.30$. *Notes*: The percentage point change in the return on capital in each scenario (upper left panel) has been separated by a vertical intercept for clarity.

red) and "no crowding out" (dashed green)). Then the economy is hit by an unexpected regulatory change at t=50, in which π increases from $\pi = 0.25$ to $\pi = 0.3$ and then remains there forever. These results are calculated by linearising the model around the terminal steady state, initialising the economy at the steady state corresponding to the pre-change regime. Figure fig:transition shows the transition of key endogenous variables.

First, consider the baseline calibration (solid blue line). The increase in π forces banks to buy government bonds. As conditions (3.6) and (3.7) hold throughout this experiment, each individual bank attempts to achieve this by selling physical capital. However, physical capital can only be sold to households or scrapped (via the capital producing firms). As there are convex adjustment costs in the adjustment of aggregate capital, this means that households need to hold more physical capital. In order to induce them to buy this capital and offset the higher management costs, the price of capital q will jump down.

However, this fall in q causes the *ex post* return on physical capital to become negative for one period, and hence causes the aggregate net worth of the banking sector to jump down, further reducing the deposits that banks can gather and hence increasing the amount of physical capital that banks need to sell. The effect of the increase in π is therefore to reduce the sum of bonds and deposits available to the household sector - this leads to lower deposit interest rates - and temporarily higher consumption - throughout the initial phase of the transition.

Following the initial period after the regime change, returns on physical capital turn positive, justifying the household sector's decision to buy such capital from the banking sector. Together with the fall in deposit interest rates, this allows the banking sector's net worth to recover slightly. However, this rebound is temporary: deposit interest rates are pinned down by the rate of time preference in the steady state (which does not change) and the rate of return on capital will also diminish once the initial capital gains peter out. Under this calibration, net worth will be lower in steady state.

Given all of these changes in the rest of the economy, the behaviour of output is unremarkable, declining steadily until reaching the new steady state level (0.8% below the starting point). However, the speed of the transition is somewhat slow - indeed it takes 33 quarters for output to decline half of the way to the new steady state. Given constant Total Factor Productivity and inelastically supplied labour, this decline maps exactly from the decline in the outstanding capital stock. In turn, this is governed by the severity of the costs of capital scrapping.

Turning to the other scenarios, a number of points stand out.

- While the profile of output is similar in each case, the combination of each channel has a smaller effect than each channel separately, in line with the interaction effect discussed in section 3.4.3.

- In the no "crowding out", $\chi = 0$ scenario (dashed green) the initial change in π does not force a large amount of capital scrapping. This is because banks are able to raise more deposits in order to satisfy the tighter liquidity requirements, which they are unable to do under full crowding out. This means that the behaviour of R^k, N and the other key variables is much smoother than in either scenario in which crowding out is present.

- The no “financial repression”, $\delta_c = 0$ scenario (dashed red) in contrast has a more severe adjustment process than the baseline when the shock hits - the price of capital falls further and rebounds more vigorously than in the baseline calibration. This is because banks initially hold a much higher share of the (larger) total capital stock (around a half rather than a third in the baseline calibration) and so a given change in π forces them to sell more capital - and hence q has to fall further on impact to induce households to purchase the capital which isn’t scrapped.

3.6 Conclusion

This paper seeks to identify the conditions under which liquidity regulation may have macroeconomic consequences. To do this, I augment a standard neoclassical growth model with convex adjustment costs in capital with financial intermediaries (“banks”) of the style proposed by Gertler and Karadi (2011) and Gertler and Karadi (2013). I identify two principal channels: a “crowding out” channel, and a “financial repression” channel. Absent each of these channels, the effect of changes in liquidity requirements of banks on the rest of the economy will be zero.

Of course, the renewed policy and academic focus on financial regulation is to avoid a repeat of the financial crisis of 2008. The model presented in this paper is only one side of the story in that it doesn’t allow for such crises and focuses exclusively in understanding the potential costs of such regulation. Hence this framework cannot be used to study optimal liquidity regulation as it stands. Incorporating crises into a framework such as the one presented here is a priority for future research.

That said, the model in this paper does have some implications for the design of financial regulation. One of the two channels discussed - the “crowding out” channel - comes about if government debt held by banks counts towards their risk weighted capital requirement. The Basel III capital regulations has supplemented the Basel-II style “risk-weighted” capital ratio with a “supplementary” leverage ratio, as well as stipulated a certain level of liquid asset holding. The model presented here suggests that this is not optimal. The model would suggest that any capital ratio should place a risk-weight of zero on assets that banks are required to hold for liquidity purposes. This would eliminate the "crowding out" channel illustrated here, and substantially reduce the costs of financial regulation outside of crisis periods. There is also a case that such a policy would not materially worsen the position of banks in a crisis, as assets which form part of a "liquidity buffer" would ordinarily be government securities which are likely to remain liquid throughout a crisis.

Appendices

3.A Proofs

3.A.1 Proof of Proposition 1 (Conditions under which the bank's inequality constraints bind)

Consider the representative bank's problem. Substituting (3.2) into (3.1) and (3.4) to eliminate d we get:

$$\max_{b^b_{t+i},k^b_{t+i}\ \forall i} \mathbb{E}_t \left[\sum_{i=1}^{\infty} \Lambda_{t,t+i}(1-\sigma)\sigma^{i-1} n_{t+i}\right]$$

subject to ($\forall t$):

$$n_t = (R^k_t - \widetilde{R}^d_{t-1})q_{t-1}k^b_{t-1} + (R^f_{t-1} - \widetilde{R}^d_{t-1})b^b_{t-1} + \widetilde{R}^d_{t-1}n_{t-1}$$

$$q_t k^b_t + \chi b^b_t \le \phi n_t$$

$$b^b_t \ge \frac{\pi}{1-\pi}(q_t k^b_t - n_t)$$

where $\widetilde{R}^d_t = R^d_t + \delta_c$ is the cost to the bank of raising one unit of deposits. The interesting case in this model is when both the regulatory capital ratio and liquidity ratio (3.3) and (3.4) bind. I proceed sequentially to show conditions under which this holds.

First, assume the the liquidity ratio (3.4) doesn't bind. It is clear from the problem above that it will be optimal for regulatory capital ratio (3.3) to bind if either:

$$\mathbb{E}_t \left[R^k_{t+1}\right] \ge \widetilde{R}^d_t \tag{3.42}$$

or

$$\mathbb{E}_t \left[R^f_t\right] \ge \widetilde{R}^d_t \text{ and } \chi > 0 \tag{3.43}$$

To see why, suppose that (3.3) doesn't bind and that the first of these conditions holds. Then the bank will not reduce their expected return by increasing their deposits by a small amount and using those deposits to buy physical capital k. So if (3.42) holds and (3.4) is slack then (3.3) binds. A similar logic applies if we substitute (3.43) for (3.42), only notice that we also require $\chi > 0$ in this case (as if $\chi \leq 0$ then increasing b will never cause the leverage constraint to bind).

However, these conditions are not sufficient if the liquidity ratio (3.4) binds. Assuming that (3.4) binds, and substituting this and the balance sheet constraint to eliminate b and d the bank's problem can be re-written:

$$\max_{k^b_{t+i}\ \forall i} \mathbb{E}_t \left[\sum_{i=1}^{\infty} \Lambda_{t,t+i}(1-\sigma)\sigma^{i-1} n_{t+i} \right]$$

subject to ($\forall t$):

$$n_t = (R^k_t - \widetilde{R}^d_{t-1}) q_{t-1} k^b_{t-1} + (R^f_{t-1} - \widetilde{R}^d_{t-1}) \frac{\pi}{1-\pi} (q_{t-1} k^b_{t-1} - n_{t-1}) + \widetilde{R}^d_{t-1} n_{t-1}$$

$$q_t k^b_t \leq \frac{\phi(1-\pi) + \pi\chi}{1 - \pi + \pi\chi} n_t$$

Taking the first order condition with respect to k^b_{t+j}:

$$\begin{aligned}
\frac{d\Pi_{b,t}}{dk^b_{t+j}} &= \mathbb{E}_t \left[\Lambda_{t,t+j+1}(1-\sigma)\sigma^j \frac{dn_{t+j+1}}{dk^b_{t+j}} \right] \\
&= \mathbb{E}_t \left[\Lambda_{t,t+j+1}(1-\sigma)\sigma^j [(R^k_{t+j+1} - \widetilde{R}^d_{t+j}) q_{t+j} + (R^f_{t+j} - \widetilde{R}^d_{t+j}) \frac{\pi}{1-\pi} q_{t+j}] \right]
\end{aligned}$$

Setting $j = 0$ it is clear from the above that if the liquidity ratio constraint binds in period t, then it will be optimal for the regulatory capital ratio to bind if:

$$\begin{aligned}
(\mathbb{E}_t \left[R^k_{t+1} \right] - \widetilde{R}^d_t) + (R^f_t - \widetilde{R}^d_t) \frac{\pi}{1-\pi} &> 0 \\
(1-\pi)\mathbb{E}_t \left[R^k_{t+1} \right] + \pi R^f_t \geq \widetilde{R}^d_t & \qquad (3.44)
\end{aligned}$$

Note that condition (3.44) - is almost sufficient for one of conditions (3.42) or (3.43) to hold as well. So (3.44) together with (3.42) is sufficient for the regulatory capital ratio to bind. Another set of sufficient conditions is for (3.44) to hold when $\chi > 0$.

Now, assume that one of (3.42) or (3.43) holds alongside (3.44), so that the regulatory capital ratio (3.3) binds. We can then substitute (3.3) and (3.2) into (3.1), and then the bank's problem becomes:

$$\max_{b^b_{t+i}\ \forall i} \mathbb{E}_t \left[\sum_{i=1}^{\infty} \Lambda_{t,t+i}(1-\sigma)\sigma^{i-1} n_{t+i} \right]$$

subject to ($\forall t$):

$$n_t = (R^k_t - \widetilde{R}^d_{t-1})(\phi n_{t-1} - \chi b^b_{t-1}) + (R^f_{t-1} - \widetilde{R}^d_{t-1})b^b_{t-1} + \widetilde{R}^d_{t-1} n_{t-1}$$

$$b^b_t(1 - \pi + \pi\chi) \geq \pi(\phi - 1)n_t$$

Then, taking the first order condition with respect to b^b_{t+j}:

$$\begin{aligned} \frac{d\Pi_{b,t}}{db^b_{t+j}} &= \mathbb{E}_t \left[\Lambda_{t,t+j+1}(1-\sigma)\sigma^j \frac{dn_{t+j+1}}{db^b_{t+j}} \right] \\ &= \mathbb{E}_t \left[\Lambda_{t,t+j+1}(1-\sigma)\sigma^j [-\chi(R^k_{t+j+1} - \widetilde{R}^d_{t+j})) + (R^f_{t+j} - \widetilde{R}^d_{t+j})] \right] \end{aligned}$$

Setting $j = 0$ it is clear from the above that if the regulatory capital ratio constraint binds in period t, then it is optimal for the bank's liquidity constraint to bind if:

$$(R^f_t - \widetilde{R}^d_t) - \chi(\mathbb{E}_t \left[R^k_{t+1} \right] - \widetilde{R}^d_t)) \leq 0 \tag{3.45}$$

as the expected increase in net worth from a marginal increase in deposits will be negative, so banks will want to hold as few deposits as possible. Given that the leverage constraint binds then the minimum amount of bonds that can be held is when $b^b_t = \pi d^b_t$.

Finally, I show that if both (3.44) and (3.45) hold, then this implies either (3.42) or (3.43) holds as well, if $\pi \in [0,1]$ and $\chi \geq 0$.

First note that if $\pi \geq 0$ (3.44) implies that either $\mathbb{E}_t \left[R^k_{t+1} \right] \geq \widetilde{R}^d_t$ and/or $R^f_t \geq \widetilde{R}^d_t$. If the first of these is true then (3.42) holds. If the second of these is true, (3.43) holds *unless*

$\chi = 0$ (as we assumed that $\chi \in [0, 1]$).

So if $\pi \in [0, 1]$, $\chi \in [0, 1]$ and (3.44) holds then there is only one case in which neither (3.42) nor (3.43) can hold: when $\mathbb{E}_t \left[R^k_{t+1}\right] < \widetilde{R}^d_t < R^f_t$ and $\chi = 0$. However, $\chi = 0$ together with (3.45) implies that:

$$\widetilde{R}^d_t \geq R^f_t \tag{3.46}$$

which contradicts $\widetilde{R}^d_t < R^f_t$. Hence, if both (3.44) and (3.45) hold and $\pi \geq 0$ and $\chi \geq 0$, then this is sufficient to imply that either (3.42) or (3.43) holds. In turn, this implies that (3.44) and (3.45) hold and $\pi \geq 0$ and $\chi \geq 0$ always imply that the regulatory capital ratio (3.3) and the liquidity ratio (3.4) will always bind.

3.A.2 Proof of Proposition 2 (The neutrality result)

Define the household's holdings of safe assets: Δ_t:

$$\Delta_t = b^h_t + d_t \tag{3.47}$$

Where I have imposed market clearing in the market for deposits ($d_t = d^b_t = d^h_t$). We can use this definition to rewrite market clearing in the market for bonds:

$$b_t = b^b_t + \Delta_t - d_t \tag{3.48}$$

Using the definition of Δ_t, and noting that $R^f_t = R^d_t$ in equilibrium, the consumer's budget constraint (3.11) becomes:

$$c_t + \Delta_t + q_t k^h_t + f(k^h_t) = w_t + R^f_{t-1}\Delta_{t-1} + R^k_t q_{t-1} k^h_{t-1} - T_t + \pi^b_t + \pi^f_t + \pi^k_t \tag{3.49}$$

We can use the second expression together with the definition of Δ to eliminate b^b_t from the bank's constraints (3.1)-(3.4). These become:

$$n_t = R^k_t q_{t-1} k^b_{t-1} + R^f_{t-1}(b_{t-1} - \Delta_{t-1}) - \delta_c d_t \tag{3.50}$$

$$q_t k^b_t + b_t = n_t + \Delta_t \tag{3.51}$$

$$q_t k^b_t + \chi(b_t - \Delta_t + d_t) = \phi n_t \tag{3.52}$$

$$b_t + (1 - \pi) d_t = \Delta_t \tag{3.53}$$

Again, assuming that the conditions stipulated in Proposition 1 hold, and so (3.52) and (3.53) bind.

If $\chi = \delta_c = 0$ then the first three of these don't contain d_t. Re-written, the model becomes:

Block 1: Banks:

$$N_t = \sigma[R^k_t q_{t-1} K^b_{t-1} + R^f_{t-1}(B_{t-1} - \Delta_{t-1})] + \omega\phi N^{t-1}$$

$$q_t K^b_t + B_t = N_t + \Delta_t$$

$$q_t K^b_t = \phi N_t$$

$$\pi^b_t = (1 - \sigma) R^k_t q_{t-1} K^b_{t-1} + R^f_{t-1}(B_{t-1} - \Delta_{t-1})] - \omega\phi N^{t-1}$$

Household:

$$c_t + \Delta_t + q_t k^h_t + f(k^h_t) = w_t + R^f_{t-1}\Delta_{t-1} + R^k_t q_{t-1} k^h_{t-1} - T_t + \pi^b_t + \pi^k_t$$

$$1 = \mathbb{E}_t \left[\Lambda_{t,t+1}\right] R^f_t$$

$$1 + f'(k^h_t) = \mathbb{E}_t \left[\Lambda_{t,t+1} R^k_{t+1}\right]$$

$$\Lambda_{t,t+1} = \frac{\beta u'(c_{t+1})}{u'(c_t)}$$

$$f(k^h_t) = \frac{\gamma_h}{2}(k^h_t - \overline{k}^h)^2$$

$$R^k_t = \frac{r^k_t + (1 - \delta) q_t}{q_{t-1}}$$

Firms:

$$r_t^k = \alpha z_t (k_t^f)^{\alpha-1}$$

$$w_t = (1-\alpha) z_t (k_t^f)^{\alpha}$$

$$I_t^n = (K_t^b + K_t^h) - (K_{t-1}^b + K_{t-1}^h)$$

$$I_t = (K_t^b + K_t^h) - (1-\delta)(K_{t-1}^b + K_{t-1}^h)$$

$$q_t = 1 + g(\frac{I_t^n + I^{SS}}{I_{t-1}^n + I^{SS}}) + \frac{I_t^n + I^{SS}}{I_{t-1}^n + I^{SS}} g'(\frac{I_t^n + I^{SS}}{I_{t-1}^n + I^{SS}}) - \mathbb{E}_t \left[\Lambda_{t,t+1} (\frac{I_{t+1}^n + I^{SS}}{I_t^n + I^{SS}})^2 g'(\frac{I_{t+1}^n + I^{SS}}{I_t^n + I^{SS}}) \right]$$

$$\pi_t^k = \Big((q_t - 1) I_t^n - g(\frac{I_t^n + I^{SS}}{I_{t-1}^n + I^{SS}})(I_t^n + I^{SS}) \Big)$$

Government:

$$g + R_{t-1}^f B_{t-1} = B_t + T_t$$

$$T_t = g + \psi_g (B_t - b^{SS})$$

Market Clearing and Aggretation:

$$K_t^f = K_{t-1}^b + K_{t-1}^h$$

$$Y_t = z_t (K_t^f)^{\alpha}$$

Block 2:

$$B_t + (1-\pi) D_t = \Delta_t$$

$$B_t = B_t^b + \Delta_t - D_t$$

$$\Delta_t = B_t^h + D_t$$

Notice that the equations listed in block 1 do not contain B_t^b, B_t^h or D_t. Therefore block 1 can be solved irrespective of the values that these variables take. Notice also that block 1 does not contain the parameter π. Hence changes in π will alter B_t^b, B_t^h or D_t (in block 2) but will not alter any of the variables in block 1. Hence if $\chi = \delta_c = 0$ then changes in π are "neutral" with respect to all variables in the economy.

3.A.3 Parameters under which the assumptions of Proposition 1 hold in steady state

Expressions for spreads $R^k - \widetilde{R}^d$ and $\widetilde{R}^d - R^f$ are:

$$\widetilde{R}^d - R^f = \delta_c$$

$$\begin{aligned}
R^k - \widetilde{R}^d &= \frac{1}{\psi^k}\left[\frac{1-\omega\phi}{\sigma} - \frac{1}{\beta}\psi^b + (\frac{1}{\beta}+\delta_c)\psi^d\right] - (\frac{1}{\beta}+\delta_c) \\
&= \frac{1}{\psi^k}\left[\frac{1-\omega\phi}{\sigma} - \frac{1}{\beta}\psi^b + (\frac{1}{\beta}+\delta_c)\psi^d - \psi^k(\frac{1}{\beta}+\delta_c)\right] \\
&= \frac{1}{\psi^k}\left[\frac{1-\omega\phi}{\sigma} - \frac{1}{\beta}\psi^b + (\frac{1}{\beta}+\delta_c)(\psi^d - \psi^k)\right] \\
&= \frac{1}{\psi^k}\left[\frac{1-\omega\phi}{\sigma} - \frac{1}{\beta}(\psi^k + \psi^b - \psi^d) + \delta_c(\psi^d - \psi^k)\right] \\
&= \frac{1-\pi+\pi\chi}{\phi(1-\pi)+\pi\chi}\left[\frac{1-\omega\phi}{\sigma} - \frac{1}{\beta} - \delta_c\frac{1-\pi\phi+\pi\chi}{1-\pi+\pi\chi}\right]
\end{aligned}$$

With our definitions for steady state asset returns and spreads, recall that necessary conditions for the bank's two inequality constraints ((3.6) and (3.7)) to be binding (in the non-stochastic steady state) are:

$$(1-\pi)R^k + \pi R^f \geq \widetilde{R}^d$$

$$\chi(R^k - \widetilde{R}^d) \geq (R^f - \widetilde{R}^d)$$

Rearranging the first of these:

$$
\begin{aligned}
(1-\pi)(R^k - \widetilde{R}^d) + \pi(R^f - \widetilde{R}^d) &\geq 0 \\
(R^k - \widetilde{R}^d) + \pi(R^f - R^k) &\geq 0 \\
(R^k - \widetilde{R}^d) &\geq \frac{\pi}{1-\pi}(\widetilde{R}^d - R^f) \\
\frac{1-\pi+\pi\chi}{\phi(1-\pi)+\pi\chi}\left[\frac{1-\omega\phi}{\sigma} - \frac{1}{\beta} - \delta_c \frac{1-\pi\phi+\pi\chi}{1-\pi+\pi\chi}\right] &\geq \frac{\pi}{1-\pi}\delta_c \\
\frac{1-\omega\phi}{\sigma} &\geq \frac{1}{\beta} + \delta_c\left[\frac{1-\pi\phi+\pi\chi}{1-\pi+\pi\chi} + \frac{\pi(\phi(1-\pi)+\pi\chi)}{(1-\pi)(1-\pi+\pi\chi)}\right] \\
\frac{1-\omega\phi}{\sigma} &\geq \frac{1}{\beta} + \delta_c \frac{1}{1-\pi} \qquad (3.54)
\end{aligned}
$$

Notice that if $\chi \geq 0$ and $\delta_c \geq 0$, the second condition will be satisfied if $R^k - \widetilde{R}^d \geq 0$, which will be true if:

$$\frac{1-\omega\phi}{\sigma} \geq \frac{1}{\beta} + \delta_c \frac{1-\pi(\phi+\chi)}{1-\pi(1+\chi)} \qquad (3.55)$$

If $\pi \in [0,1]$ and $\phi > 1$ then

$$\frac{1-\pi(\phi+\chi)}{1-\pi(1+\chi)} < 1 \leq \frac{1}{1-\pi}$$

So A1(i)-(iv) and (3.54) are a sufficient conditions for (3.6) and (3.7) to hold in steady state.

Conclusion

This thesis presented three chapters in the broad field of Applied Macroeconomics. The first chapter argued that firm stock price movements on days on which firms release financial information (such as quarterly earnings reports) were likely to reflect firm-specific, rather than aggregate information. I used this in conjunction with an adaptation of a new econometric technique – identifying Vector Autoregressions (VARs) to recover estimates of shocks which originate at the firm level, rather than the response of firm productivity to aggregate or sectoral developments. I then compared how economic aggregates responded to this measure of firm shocks relative to the measure proposed in the original paper of Gabaix (2011), and found that this new measure suggests that these "granular" shocks are somewhat less important for aggregate fluctuations than these earlier estimates. I also provide evidence that the high share of aggregate fluctuations "explained" by the earlier measure are likely to come in part from the fact that the very largest firms are more cyclical than the average firm.

The second chapter examined the empirical claim that GDP is best modelled with a random walk, and hence a 1% fall in GDP should be associated with a 1% fall in long term economic forecasts. We show that recoveries are in fact predictable in previous UK and US recessions using multivariate models, and argue that this is because GDP is an aggregate of nonstationary and stationary components. This implies that while the "correct" univariate representation of GDP is nonstationary, richer models can identify predictable changes where these exist. An implication of the first chapter (and indeed the third chapter) is that many shocks can matter for aggregate fluctuations and as such multivariate models may be beneficial for forecasting, even if the latter involves estimating considerably more parameters.

The final chapter examines the consequences of tightening liquidity regulation on financial intermediaries on the aggregate economy. Using a general equilibrium model, I illustrate two different channels: a "crowding out" channel under which higher liquidity requirements reduce the capacity of financial intermediaries to intermediate deposits for a given level of net worth, and a "financial repression" channel through which higher liquidity requirements erode financial intermediary net worth, and hence raise the cost of capital for firms, over time. I demonstrate a neutrality result such that in the absence of either of these channels liquidity regulation does not raise the cost of capital and adversely affect aggregate output or consumption. I go on to calibrate the model to US data and attempt to quantify the effect of higher liquidity requirements on interest rates, financial sector net worth, output and consumption when one or both of the channels is present.

Bibliography

Acharya, V. V., H. S. Shin, and T. Yorulmazer (2011): "Crisis Resolution and Bank Liquidity," *Review of Financial Studies*, 24, 2166–2205.

Acharya, V. V. and S. Viswanathan (2011): "Leverage, Moral Hazard, and Liquidity," *Journal of Finance*, 66, 99–138.

Admati, A. and M. Hellwig (2013): *The Bankers' New Clothes: What's Wrong with Banking and What to Do about It*, vol. 1 of *Economics Books*, Princeton University Press.

Adrian, T. and N. Boyarchenko (2013): "Liquidity Policies and Systemic Risk," Staff Reports 661, Federal Reserve Bank of New York.

Allen, F., E. Carletti, and D. Gale (2009): "Interbank market liquidity and central bank intervention," *Journal of Monetary Economics*, 56, 639–652.

Angeloni, I. and E. Faia (2013): "Capital regulation and monetary policy with fragile banks," *Journal of Monetary Economics*, 60, 311–324.

Atalay, E. (2014): "How Important Are Sectoral Shocks," Working Papers 14-31, Center for Economic Studies, U.S. Census Bureau.

Blanchard, O., R. Perotti, et al. (2002): "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output," *The Quarterly Journal of Economics*, 117, 1329–1368.

Blanchard, O. J., J.-P. L'Huillier, and G. Lorenzoni (2013): "News, Noise, and Fluctuations: An Empirical Exploration," *American Economic Review*, 103, 3045–3070.

CAMPBELL, J. Y. AND N. G. MANKIW (1987): "Are Output Fluctuations Transitory?" *The Quarterly Journal of Economics*, 102, 857–880.

CARVALHO, V. AND B. GRASSI (2015): "Large Firm Dynamics and the Business Cycle," Working Papers 824, Barcelona Graduate School of Economics.

CHARI, V. V., P. J. KEHOE, AND E. R. MCGRATTAN (2007): "Business cycle accounting," *Econometrica*, 75, 781–836.

CHAUVET, M. AND S. POTTER (2013): *Forecasting Output*, Elsevier, vol. 2 of *Handbook of Economic Forecasting*, chap. 0, 141–194.

DI GIOVANNI, J., A. A. LEVCHENKO, AND I. MEJEAN (2014): "Firms, Destinations, and Aggregate Fluctuations," *Econometrica*, 82, 1303–1340.

DOMAR, E. D. (1961): "On the measurement of technological change," *The Economic Journal*, 71, 709–729.

EDGE, R. M. AND R. S. GURKAYNAK (2010): "How Useful Are Estimated DSGE Model Forecasts for Central Bankers?" *Brookings Papers on Economic Activity*, 41, 209–259.

EDGE, R. M., M. T. KILEY, AND J.-P. LAFORTE (2010): "A comparison of forecast performance between Federal Reserve staff forecasts, simple reduced-form models, and a DSGE model," *Journal of Applied Econometrics*, 25, 720–754.

FAIR, R. C. AND R. J. SHILLER (1990): "Comparing Information in Forecasts from Econometric Models," *American Economic Review*, 80, 375–389.

FARHI, E. AND I. WERNING (2013): "A Theory of Macroprudential Policies in the Presence of Nominal Rigidities," NBER Working Papers 19313, National Bureau of Economic Research, Inc.

FERNALD, J. ET AL. (2012): "A quarterly, utilization-adjusted series on total factor productivity," *Federal reserve bank of San Francisco working paper*, 19, 2012.

FOERSTER, A. T., P.-D. G. SARTE, AND M. W. WATSON (2011): "Sectoral versus Aggregate Shocks: A Structural Factor Analysis of Industrial Production," *Journal of Political Economy*, 119, pp. 1–38.

Freeman, S. (1987): "Reserve requirements and optimal seigniorage," *Journal of Monetary Economics*, 19, 307–314.

Gabaix, X. (2011): "The Granular Origins of Aggregate Fluctuations," *Econometrica*, 79, 733–772.

Gertler, M. and P. Karadi (2011): "A model of unconventional monetary policy," *Journal of Monetary Economics*, 58, 17–34.

——— (2013): "QE 1 vs. 2 vs. 3. . . : A Framework for Analyzing Large-Scale Asset Purchases as a Monetary Policy Tool," *International Journal of Central Banking*, 9, 5–53.

——— (2015): "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics*, 7, 44–76.

Gertler, M. and N. Kiyotaki (2010): "Financial Intermediation and Credit Policy in Business Cycle Analysis," in *Handbook of Monetary Economics*, ed. by B. M. Friedman and M. Woodford, Elsevier, vol. 3 of *Handbook of Monetary Economics*, chap. 11, 547–599.

——— (2013): "Banking, Liquidity and Bank Runs in an Infinite-Horizon Economy," NBER Working Papers 19129, National Bureau of Economic Research, Inc.

Gertler, M., N. Kiyotaki, and A. Queralto (2012): "Financial crises, bank risk exposure and government financial policy," *Journal of Monetary Economics*, 59, S17–S34.

Goodhart, C. A., A. K. Kashyap, D. P. Tsomocos, and A. P. Vardoulakis (2012): "Financial Regulation in General Equilibrium," NBER Working Papers 17909, National Bureau of Economic Research, Inc.

Granger, C. W. and M. J. Morris (1976): "Time series modelling and interpretation," *Journal of the Royal Statistical Society. Series A (General)*, 246–257.

Granger, C. W. J. (1980): "Long memory relationships and the aggregation of dynamic models," *Journal of Econometrics*, 14, 227–238.

Guellec, D. and B. V. P. de la Potterie (2001): "Measuring productivity-OECD Manual: Measurement of Aggregate and Industry-Level Productivity Growth," .

Haan, W. J. D., S. W. Sumner, and G. M. Yamashiro (2011): "Bank Loan Components and the Time-varying Effects of Monetary Policy Shocks," *Economica*, 78, 593–617.

Hopenhayn, H. A. (1992): "Entry, Exit, and Firm Dynamics in Long Run Equilibrium," *Econometrica*, 60, 1127–50.

Hulten, C. R. (1978): "Growth accounting with intermediate inputs," *The Review of Economic Studies*, 45, 511–518.

Imrohoroglu, A. and S. Tüzel (2014): "Firm-level productivity, risk, and return," *Management Science*, 60, 2073–2090.

Jovanovic, B. (1987): "Micro Shocks and Aggregate Risk," *The Quarterly Journal of Economics*, 102, 395–409.

Kilian, L. (2009): "Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market," *The American Economic Review*, 99, 1053–1069.

Long, John B, J. and C. I. Plosser (1983): "Real Business Cycles," *Journal of Political Economy*, 91, 39–69.

MacKinlay, C. A. (1997): "Event Studies in Economics and Finance," *Journal of Economic Literature*, 35, 13–39.

Mertens, K. and M. O. Ravn (2013): "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States," *American Economic Review*, 103, 1212–47.

Nelson, C. R. (1972): "The prediction performance of the FRB-MIT-PENN model of the US economy," *The American Economic Review*, 62, 902–917.

Nuño, G. and C. Thomas (2013): "Bank leverage cycles," Working Paper Series 1524, European Central Bank.

Olley, G. S. and A. Pakes (1996): "The Dynamics of Productivity in the Telecommunications Equipment Industry," *Econometrica*, 64, 1263–97.

on Banking Supervision, B. C. (2012): "Results of the Basel III monitoring exercise as of 31 December 2011," Tech. rep., Bank for International Settlements.

——— (2013): "Basel III: The Liquidity Coverage Ratio and liquidity risk monitoring tools," Tech. rep., Bank for International Settlements.

Ramey, V. A. (2011): "Identifying Government Spending Shocks: It's all in the Timing," *The Quarterly Journal of Economics*, 126, 1–50.

Romer, C. D. and D. H. Romer (1989): "Does monetary policy matter? A new test in the spirit of Friedman and Schwartz," in *NBER Macroeconomics Annual 1989, Volume 4*, MIT Press, 121–184.

——— (2004): "A New Measure of Monetary Shocks: Derivation and Implications," *American Economic Review*, 94, 1055–1084.

Smets, F. and R. Wouters (2007): "Shocks and frictions in US business cycles: A Bayesian DSGE approach," *The American Economic Review*, 97, 586–606.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stella, A. (2015): "Firm dynamics and the origins of aggregate fluctuations," *Journal of Economic Dynamics and Control*, 55, 71 – 88.

Stock, J. H. and M. W. Watson (2002): "Macroeconomic forecasting using diffusion indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

——— (2012): "Disentangling the Channels of the 2007-2009 Recession," NBER Working Papers 18094, National Bureau of Economic Research, Inc.

Summers, L. H. (1986): "Some skeptical observations on real business cycle theory," *Quarterly Review*, 23–27.

Terry, S. J. (2015): "The macro impact of short-termism," Tech. rep.

Wagner, A., R. J. Zeckhauser, and A. Ziegler (2017): "Company Stock Reactions to the 2016 Election Shock: Trump, Taxes and Trade," NBER Working Papers 23152, National Bureau of Economic Research, Inc.