Nonparametric and Semiparametric Estimation and Testing

Coenraad A.P. Pinkse

Submitted for the degree of Ph.D. in Statistics
April 1994
London School of Economics and Political Science

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

This thesis deals with certain problems in nonparametric estimation and testing.

In the first part of the thesis, we propose a method to improve nonparametric regression estimates of regression functions with a similar shape. This is achieved by first estimating the unknown parameters in the parametric relationship between the regression functions, and subsequently using the estimated transformation to pool the two data sets.

The second part is concerned with nonparametric tests for serial independence. We extend an idea by Robinson (1991a) to use the Kullback-Leibler information criterion to measure the distance between the joint and marginal densities of consecutive observations in a stationary time series, and we also propose an entirely new test in which the joint and marginal characteristic functions of afore-mentioned observations are used.

Acknowledgements

This thesis would not have been possible without the guidance and comments of Peter Robinson, for which I am very grateful. Thanks to Peter's scrutiny, my understanding of econometrics and statistics, as well as my writing style, have improved considerably over the last few years.

I would like to thank the Foreign and Commonwealth Office for their financial support, and the British Council, Trudy Kragtwijk and Marjanne van der Graaf in particular, for their role in the Foreign and Commonwealth Office awards programme. I also acknowledge the support of the Economic and Social Research Council through grants R000231441 and R000233609. Finally, I would like to thank my parents for their support throughout.

Contents

List of Tables

List of Figures

Part I

Shape-Invariant Modelling

Chapter 1

Introduction

1.1 Outline

This thesis is concerned with estimation and testing in a non- or semiparametric context. Broadly speaking, two fairly distinct topics are discussed, and we have therefore split the thesis into two parts.

The first part of this thesis deals with ways of improving nonparametric regression estimates by means of pooling them with nonparametric estimates of other regression functions, that have a similar shape, and in the second part two different nonparametric tests for serial independence are proposed.

In the remainder of the current chapter, we explain our notation, and we discuss nonparametric estimation methods, in particular kernel estimation. In Chapter 2, we describe how we can estimate the (parametric) relationship between two unknown regression functions, and how these parameter estimates can then be used to find an optimal way of pooling the nonparametric regression estimates; the sense in which the pooling method is optimal is discussed in Chapter

2, also.

Chapter 3 provides a general overview of the serial independence testing literature, both parametric and nonparametric, and discusses in some detail the issues at hand. In Chapter 4, we extend Robinson's (1991a) test for serial independence based on the Kullback-Leibler Information Criterion [cf. Kullback (1959)] in various ways, and in Chapter 5 we propose a new serial independence test based on characteristic functions.

Chapter 6, finally, summarises our achievements in this thesis.

1.2 Notation

The notation used in this thesis is fairly standard. We use capital letters for random variables, sets, and sometimes also for constants. If a caret is added to a symbol, it is an estimate of the corresponding quantity. We sometimes suppress subscripts and superscripts when this can be done without affecting clarity. P, E, V are reserved for probability, expectation and variance, respectively, where $E[X]^2 = EX^2$, and $E^2[X] = (EX)^2$.

The number N is reserved for the number of observations, and after the current chapter, k is always a kernel used in nonparametric estimation; kernels are explained later in this chapter. A superscript T is used for transposition.

Convergence in probability is denoted by $\xrightarrow{P}$, and convergence in distribution by $\xrightarrow{\mathcal{L}}$. When a convergence result holds uniformly across values of a certain variable, this is stated explicitly.

We use a combination of methods to denote derivatives. Let $m : \Re^J \to \Re$, for some finite positive integer J. If $J = 1$, m' denotes the first derivative with respect to its argument, m'' its second, etcetera. Further, $m^{(i)}$ is m's i-th derivative.

In the general case, the first partial derivative of m with respect to its j-th argument is

denoted by $\frac{\partial m}{\partial x_j}_{|x}$ or as $D_j m_{|x}$. Similarly, for any i, j,

$$\frac{\partial^2 m}{\partial x_j \partial x_i}_{|x} = \frac{\partial \frac{\partial m}{\partial x_j}}{\partial x_i}_{|x} = D_j D_i m_{|x} = D_i D_j m_{|x}.$$

Higher order variations are also used.

1.3 Nonparametric Estimation

There are various nonparametric methods to estimate densities or regression functions. We only use the kernel method, not because none of the other methods would be appropriate or relevant, but because it seemed most convenient. Indeed, we did not have any a priori information pertaining to which estimate would give us the best results, and it may well be of interest to study the same problems using other nonparametric estimation methods for a comparison.

In Subsection 1.3.1, we discuss some of the issues relating to kernel density estimation, and in Subsection 1.3.2 kernel regression estimates are discussed. Finally, in Subsection 1.3.3, we briefly review some other nonparametric estimates.

1.3.1 Kernel Density Estimation

Suppose we wish to know the distribution of a series of i.i.d. random variables $\{X_i\}$ with distribution function F. A logical estimate of F is the empirical distribution function $\hat{F}(x) = \frac{1}{N}\sum_i I(X_i \leq x)$, with N the number of observations and I the indicator function, taking a value of 1 if its argument is true, and a value of 0, otherwise. Suppose now that the distribution is continuous, and we wish to estimate the density f of X_1. Simply differentiating $\hat{F}$ does not work, as it is a simple step function with N "steps". However, $I(X_i \leq x)$ is equal to $I(\frac{x-X_i}{h} \geq 0)$ for any $h > 0$. If h, which is called the *bandwidth*, is a sequence of numbers that tends to zero, as $N \to \infty$, we can replace I by any everywhere differentiable distribution function K, because

$K(\frac{x-X_i}{h}) \to 0$, if $x < X_i$, and $K(\frac{x-X_i}{h}) \to 1$, if $x > X_i$, as $h \to 0$, and because $P[X_i = x] = 0$, for any continuously distributed random variable. Because K is differentiable, so is our new $\hat{F}(x) = \frac{1}{N}\sum_i K(\frac{x-X_i}{h})$. We can thus construct an estimate of the density at x, by differentiating $\hat{F}$ with respect to x, which leads to $\hat{f}(x) = \frac{1}{Nh}\sum_i k_h(x - X_i)$, with $k_h(x) = k(x/h)$, and k the first derivative of K. $\hat{f}(x)$ is called the *kernel density estimate* at x.

Another way of looking at $\hat{f}(x)$ is as a sum of weights. Suppose for the moment that k is even, increasing for $x < 0$, and decreasing for $x > 0$. Then, if X_i is close to x, $k_h(x - X_i)$ will be large, and if X_i is far away, it will be small. Thus, if there are many observations close to x, the density estimate at x will be relatively large, which is in line with what one would intuitively expect. However, as the number of observations increases, the kernel density estimate will put increasing emphasis on observations that are very close to x, making the estimate more precise.

There is a substantial degree of arbitrariness in the choice of both k and h. The combination of these choices determines the weight each observation gets in the density estimate at a certain point. Indeed, the choices are not separable in that there is not one single best choice of h irrespective of the choice of k. If, for instance, $k^*(x) = ck(cx)$, $c > 0$, then k^* is also a kernel, and the combination k^*, ch, will lead to the same estimates as k, h.

There has however been found little variation in performance due to the shape of k, provided that k is even and unimodal. Härdle (1990), page 138, ranks a number kernels in terms of their efficiency, and found the Epanechnikov kernel $k(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$ [cf. Epanechnikov (1969)] to be the best polynomial kernel of degree 2 on $[-1, 1]$ in terms of the mean squared error. This kernel is not everywhere differentiable, however, and it can therefore not be used in all circumstances. Indeed, in many places in this thesis we require the existence of derivatives of the kernel.

Kernel estimates are very sensitive to the bandwidth, however, or to the constant c in the

above example, for that matter. We only required h to tend to zero, if the number of observations tended to ∞, but that is hardly an adequate guide line for the choice of h in a finite sample. Choosing h large, will result in too many observations getting some weight making the density look overly smooth. Choosing h too small, however, would result in a very ragged shape of the density estimate. In more formal terms, the greater h, the larger the bias, but the smaller the variance. In fact, for the above type of kernel, the bias is of order h^2, which is easily deduced using a second order Taylor expansion, whilst the variance is of order $N^{-1}h^{-1}$.

We could try to use the mean squared error, $E[\hat{f}(x) - f(x)]^2 = E^2[\hat{f}(x) - f(x)] + V\hat{f}(x)$, to determine the bandwidth. There are however too many unknowns in its formula to derive an exact expression for the optimal h. As follows from the previous paragraph, the squared bias is of order h^4 and the variance of order $N^{-1}h^{-1}$, such that for increasing N, the optimal h should tend to zero at a rate of $N^{-\frac{1}{5}}$. For the practitioner, this is of no use, as he is faced with choosing h for one single sample of size N.

A method that does give a specific choice of h is *cross validation*. The optimal h is chosen to be *that* value of h for which

$$\sum_i \log\left\{\frac{1}{Nh}\sum_{j\neq i} k\left(\frac{X_i - X_j}{h}\right)\right\} \tag{1.1}$$

is maximised. The expression whose logarithm is taken is the kernel density estimate at X_i, based on all observations except X_i itself. The rationale is similar to that for maximum likelihood estimation. In maximum likelihood estimation, we would maximise $\sum_i \log f_\theta(X_i)$, with respect to a finite parameter vector θ, where f is known up to θ. In the nonparametric case, f itself is not known, but we do have a proxy, namely $\hat{f}$, where the unknown parameter is now h rather than θ. Maximising $\sum_i \log \hat{f}(X_i)$ would lead to $h = 0$, because X_i itself is used in the creation of $\hat{f}$; it is therefore omitted in (1.1). It has been found that the method of cross validation generally leads to a bandwidth that is too large, with the effect that the density estimate is overly smooth.

It is moreover more complicated to derive properties for data-dependent bandwidths, such as those selected by the cross validation method, than it is for bandwidths that depend solely on the sample size. Data-dependent bandwidths have been used in various articles, however, e.g. in Robinson (1991c) and Härdle, Hall and Ichimura (1993).

It is possible to reduce the bias of the kernel density estimate. If $\int k(x)x^i dx = 0$, for $i = 0, \ldots, l-1$, we say that k is an l-th order kernel. This would require k to be negative in some places, but the argument at the beginning of the present section can be easily accomodated to allow for such kernels, provided they integrate to one. It can be shown, and we prove this for instance in Lemma 4.2, that, if the density is sufficiently smooth, the bias is of order h^{2l}, whilst the variance remains of order $N^{-1}h^{-1}$, implying an optimal rate for h of $N^{-\frac{1}{2l+1}}$. Although the order of the bias is reduced, and the order of the variance remains the same, the variance nonetheless increases with an increase in the order of the kernel employed. In small and moderate samples, it is often found that choosing a higher order kernel renders the estimates no more accurate than estimates using kernels of a lower order.

1.3.2 Kernel Regression Estimation

Suppose, as earlier on in this chapter, that we wish to estimate a regression function m on the basis of the regressands $\{Y_i\}$ and regressors $\{X_i\}$. We shall for the sake of simplicity restrict our attention to univariate X_i's. So we have $Y_i = m(X_i) + \varepsilon_i$, for all i, with $\{\varepsilon_i\}$ a series of disturbances independent of the regressors.

If Y_i is generally large for X_i close to x, where x is the point at which we wish to estimate m, then it makes sense to presume that m is large at x, also, provided of course that m is continuous. One may thus consider a weighted average of the Y_i's, giving greater weight to the ones that have X_i that are close to x. Nadaraya (1964) and Watson (1964) proposed to

define $\hat{m}(x)$ as the ratio of $\hat{r}(x)$ upon $\hat{f}(x)$, with $\hat{r}(x) = \frac{1}{Nh}\sum_i k_h(x - X_i)Y_i$, and $\hat{f}(x)$ defined as before. The denominator term is needed, because $\hat{r}(x)$ estimates $r(x) = m(x)f(x)$ rather than $m(x)$ itself. $\hat{r}(x)$ is generally a consistent estimate of $r(x)$, such that $\hat{m}(x)$ is a consistent estimate of $m(x)$.

A problem with kernel regression estimates is the denominator, which may be very close to zero, or even equal to or less than zero, when higher order kernels are employed. The regression estimate does therefore not always exist. At the very least, the random denominator makes statistical treatment often rather cumbersome.

1.3.3 Other Nonparametric Estimates

As noted earlier, there are many other ways to estimate densities or regression functions in a nonparametric fashion. Indeed, many standard texts [e.g. Prakasa Rao (1983), Silverman (1985), Härdle (1990)] cover a wide variety of such estimates as do survey articles, such as Buja, Hastie, and Tibshirani (1989).

It would be well beyond the scope of this thesis to discuss every such estimate in detail, and we limit ourselves to give a short description of some of them. We moreover only examine the case of regression estimation.

1.3.3.1 k Nearest Neighbours

In contrast to the kernel estimate, the *k nearest neighbour* estimate does not weigh each observation as a function of their Euclidean distance to the point of interest, say x, but rather assigns weights according to their relative position in relation to x. The k points that lie closest to x are all assigned weights different from zero, and various weighting schemes have been used.

Essentially, the k-nearest neighbour regression estimate is

$$\hat{m}_{k-nn}(x) = \sum_i w_i(x) Y_i,$$

where $w_i(x)$ is equal to zero, when observation i is not among the k closest to x, and takes a value other than zero, otherwise. The most popular weights are the uniform weights for which $w_i(x) = k^{-1}$, when X_i is among the k closest to x. As sample size increases, the number of nearest neighbours k should increase, also, but at a slower rate than N. There are many other weighting schemes [cf. e.g. Stone (1977)], and an exhaustive list is well beyond the scope of this thesis. It is generally possible to use higher order weights for other weighting schemes, also.

1.3.3.2 Orthogonal Series

It is well-known that any Hilbert space has a complete orthonormal basis. If the space is also separable, then any complete orthonormal basis consists of countably many elements [cf. e.g. Dudley (1989)]. If μ denotes the Lebesgue measure on a compact set, then $L^2(\mu)$ is separable, such that any function $m \in L^2(\mu)$ can be written as $m(x) = \sum_{j=1}^{\infty} \alpha_j e_j(x)$, where $\{e_j\}$ is some complete orthonormal basis of $L^2(\mu)$, and $\alpha_j = \int e_j(x) m(x) d\mu(x)$.

Indeed, when m is a density, $\alpha_j = E e_j(X_1)$, for all j, which can be estimated by $\hat{\alpha}_j = \frac{1}{N} \sum_i e_j(X_i)$. In that case, a logical estimate of f at x is $\hat{f}(x) = \sum_{j=1}^{q} e_j(X_i)$, where q is some sample size dependent sequence that tends to ∞ as N, but more slowly than N.

In the case in which m is a regression function, one has

$$Y_i = \sum_{j=1}^{q} \alpha_j e_j(X_i) + \sum_{j=q+1}^{\infty} \alpha_j e_j(X_i) + \varepsilon_i,$$

for all i. For given sample size N, we can estimate $\alpha_1, \ldots, \alpha_q$ by least squares, where q is once again a cut off sequence that tends to ∞ at a slower rate than N.

1.3.3.3 Further Remarks

Other examples of nonparametric regression estimates are histogram estimates, running mean smoothers, bin smoothers, regression splines, running line smoothers, and series estimates other than orthogonal series estimates. Prakasa Rao (1983) and Härdle (1990) discuss the pros and cons of various (but not all) nonparametric density and regression estimates.

Chapter 2

Pooling Nonparametric Estimates of Regression Functions with a Similar Shape

2.1 Objective

Given two nonparametric estimates for the same regression function based on different samples, intuitively one expects that combining, or *pooling*, the two estimates will — by virtue of the larger number of observations used — lead to a gain in precision. In this chapter we wish to establish that this is indeed the case, not only for regression functions that are identical but also for those that are just similar in shape. We will also give a procedure on how to optimally combine the two kernel regression estimates.

When we say '*similar in shape*' (or when we speak of '*shape-invariant modelling*') we mean

that we know two transformations (one for the argument and one for the function value), up to a finite number of parameters, which transform one regression function into the other. If the parameters were known, one could simply transform one of the kernel estimates, such that we have two estimates of the same regression function. If the parameters are unknown, but can be estimated suitably efficiently, we show that the resulting pooled nonparametric regression estimate is asymptotically as efficient as employing known parameter values.

The related literature is not very extensive and generally has a somewhat different goal. It is in general not concerned with the improvement of the nonparametric estimates but rather with the estimation of the above parameters. It further generally assumes deterministic regressors, in particular ones with support on the unit interval and which depend on N, the number of observations, in order for the asymptotic theory to go through; the applications envisaged are mostly in physical and biometric experiments. An obvious application can be found in the estimation of human growth curves. All humans (of the same sex) have the same growth pattern, but the exact location of peaks and such may vary.

Largo et al. (1978) used smoothing splines to estimate the height growth velocity curves for boys and girls aged 4.5 to 17.75 years. Stützle et al. (1980) used an iterative approach, in which they first imposed a structural form, then estimated the transformation (and some other) parameters, which they employed to update the structural form, using B-splines, etcetera. Their approach is similar to that of Lawton et al. (1972), who modelled the air expelled from human lungs over time. Gasser et al. (1984) use kernel regression estimates to estimate individual growth curves, but are not really concerned with finding suitable parametric transformations. Kneip and Gasser (1988) use an iterative least squares procedure to estimate transformation parameters and functional form. Their result is very general in that the class of functional forms allowed is restricted, but not specified. Their paper is mathematically rigorous, in contrast to the other

authors mentioned above, and they obtain a uniform consistency result for their procedure. An entirely parametric application of growth curves can be found in Rao (1977).

In other disciplines, for instance in economics, the argument for deterministic regressors is much less convincing. Shape-invariant modelling is however also relevant in these disciplines; for instance in the estimation of so-called '*Engle-curves*', i.e. curves that represent the relationship between income and food expenditure. Another interesting setting in which shape-invariant modelling can be of interest is panel data. If, for instance, we have observed a panel with a large number of 'individuals' but over a limited number of time periods, say two, it makes sense to assume that the relationship to be estimated in both periods is similar in shape, and hence shape-invariant modelling may be of interest. In effect, practitioners usually prescribe a linear relationship between the regressand and regressors in both time periods. If the number of regressors is the same in both equations, this would imply shape invariance. If one knows that the relationship between regressands and regressors is linear, however, nonparametric regression estimates are inefficient, indeed they converge at a slower rate, and in such a setting this chapter would be of little relevance. If, on the other hand, one is uncertain about linearity, nonparametric regression estimation may be the only way to appropriately explain the relationship between the above variables. In practice, the present approach seems most useful when the regression functions of interest are identical up to scale and location parameter. To be particularly relevant in an econometric context, the results in this chapter may need to be extended to allow for more regressors. It is important to point out, though, that the precision of nonparametric estimates deteriorates rather quickly with an increase in the number of regressors.

We will use scalar random regressors and will consider two main cases, which we shall label $\mathcal{C}_1$ and $\mathcal{C}_2$. In $\mathcal{C}_1$ all regressors and disturbances are mutually independent and all regressors are identically distributed. Further, the conditional expectation of any disturbance conditional on

any regressor is zero. The regressors in both samples are also assumed to admit the same density. In $\mathcal{C}_2$ the regressors in one sample may be differently distributed from those in the other and we will allow for limited dependencies between regressors in different samples. In Section 2.2, we show that parameter estimates are $\sqrt{N}$-consistent and asymptotically normal in case of $\mathcal{C}_1$, that they converge at a rate faster than that of the nonparametric regression estimates in case of $\mathcal{C}_2$ and that whether we know or estimate the true parameters makes no difference for the pooling procedure, at least asymptotically. In Section 2.3 we demonstrate and justify optimal pooling of two kernel regression estimates for the same regression function. Section 2.4 provides some afterthoughts. Finally, in Section 2.5, we present some simulation results. The nonparametric regression methods used are kernel-based, and some useful but standard asymptotic properties of the kernel estimates, which we employ, are presented in the appendix.

2.2 Parameter Estimation

2.2.1 Models

In this subsection we will give a general outline of the models we wish to estimate. We will leave a full description of the regularity conditions to the next subsection.

We are concerned with the estimation of the following regression models:

$$\begin{aligned} Y_i &= m(X_i) + U_i, \quad i = 1, \ldots, N, \\ Z_i &= m_a(W_i) + V_i, \quad i = 1, \ldots, N_a, \end{aligned}$$

where X_i, Y_i, W_i, Z_i are scalar observables and m and m_a are nonparametric functions, with $E[U_i|X_j] = E[V_i|W_j] = 0$, almost surely for all i, j. We assume that two transformations S and T exist such that for all x

$$m(x) = S(\xi_0, m_a(T(\mu_0, x))),$$

and that these transformations are known except for the column parameter vector $\theta_0 = [\xi_0^T, \mu_0^T]^T$. Later on we will estimate θ_0 by $\hat{\theta} = [\hat{\xi}^T, \hat{\mu}^T]^T$. In Section 2.4, we briefly discuss the impact of incorrectly specifying the above parametric transformation.

Härdle and Marron (1991) suggest estimating θ_0 by minimising the following function with respect to θ:

$$\int [\hat{m}(x) - S(\xi, \hat{m}_a(T(\mu, x)))]^2 w(x) dx,$$

where w is a bounded function that is positive on the interior of a compact interval and zero elsewhere and $\hat{m} = \hat{r}/\hat{f}$ and $\hat{m}_a = \hat{r}_a/\hat{f}_a$ are Nadaraya-Watson kernel estimates for $m = r/f$ and $m_a = r_a/f_a$, respectively (where f and f_a are the densities of X_1 and W_1, respectively). When deterministic regressors are involved this approach is satisfactory, but with random regressors it is rather cumbersome, because the kernel regression estimate is then a ratio of random variables and taking expectations is awkward; indeed, they may not exist. We will only examine the most important case, namely when S is linear in both arguments, in contrast to Härdle and Marron (1991). Then we can write

$$m(x) = \xi_{01} + \xi_{02} m_a(T(\mu_0, x))$$

$$\iff f_a(T(\mu_0, x)) r(x) = \xi_{01} f(x) f_a(T(\mu_0, x)) + \xi_{02} f(x) r_a(T(\mu_0, x)),$$

for all x. We thus define $\hat{\theta}$ as the value of θ that minimises the loss function

$$L_N(\theta) := \int \Lambda_N^2(x, \theta) w(x) dx,$$

where

$$\Lambda_N(x, \theta) = \hat{f}_a(T(\mu, x)) \hat{r}(x) - \xi_1 \hat{f}(x) \hat{f}_a(T(\mu, x)) - \xi_2 \hat{f}(x) \hat{r}_a(T(\mu, x)).$$

We require w to be also twice boundedly differentiable (besides being positive only on the interior of a compact interval and zero elsewhere), to allow for a second order Taylor expansion.

There is evidently any number of ways to get estimates of the unknown parameters, and the above choice of loss function is only one possible way. It allows us to limit the area over which numerical integration has to be performed and it allows us to exclude the tails in which nonparametric density and regression estimates are notoriously poor.

In $\mathcal{C}_1$ we will show that $\hat{\theta}$ is $\sqrt{N}$-consistent for θ_0 and asymptotically normal, and in $\mathcal{C}_2$ that $\hat{\theta} - \theta_0 = o_p(N^{-\frac{2}{5}})$. We have chosen this specific rate because the optimal kernel regression estimate that does not use higher order kernels converges at rate $N^{-\frac{2}{5}}$ and we only need to show that the parameter estimates converge faster. If higher order kernels were employed, we would need to show that $\hat{\theta}$ converges to θ_0 at a rate faster than the convergence rate of higher order kernel regression estimates. However, in the estimation of θ_0, we have thus far not used higher order kernels, either. Indeed, the convergence rate of $\hat{\theta}$ could generally be improved upon if higher order kernels were employed in its estimation. It should be noted, though, that higher order kernels are often less precise in moderate samples, despite their higher convergence rate.

2.2.2 Assumptions

In this subsection we will state the assumptions required for the results of the present section to go through. Throughout, a superscript (j) denotes the j-th derivative, but we will also — when we feel it improves clarity — use the usual ($'''$ is third derivative) notation.

First, we define two useful function classes.

Definition 2.1 *The class $\mathcal{G}_l$ comprises all functions g that are l times boundedly differentiable.*

The functions g that we have in mind are all multiples of a function of interest and a probability density, so many cases in which the function of interest is unbounded are included.

Definition 2.2 *The class $\mathcal{K}_l$ comprises all l times boundedly differentiable functions k that are*

even, integrate to one and satisfy

$$\int |k^{(i)}(u)|^2 du < \infty,\ i = 0, \ldots, l,\ \int |k(u)| u^2 du < \infty.$$

$$k(x) = \int \phi_k(u) e^{iux} du,$$

for all x, where

$$\int |u^i \phi_k(u)| du < \infty,\ i = 0, \ldots, l.$$

Assumption 2.A *It is assumed that all (U_i, V_i) pairs are mutually independent, and that U_i and V_i both have finite second moments. We assume $E[U_i|X_j] = E[V_i|W_j] = 0$ for all i, j. We require that the X_i's are i.i.d. with density $f \in \mathcal{G}_2$ and that the W_i's are i.i.d. with density $f_a \in \mathcal{G}_2$. We also assume that $r \in \mathcal{G}_2$, $r_a \in \mathcal{G}_2$, $q := m^2 f \in \mathcal{G}_0$ and $q_a := m_a^2 f_a \in \mathcal{G}_0$. Further, an unknown vector $\theta_0 = [\xi_0, \mu_0]^T \in \Theta$, with Θ a bounded and open set, and two functions S and T exist such that for all x, $m(x) = S(\xi_0, m_a(T(\mu_0, x)))$. We assume that S is linear in both arguments. We also assume that N_a and N increase at the same rate. In C_1 and C_2 respectively we require:*

1. *$f \equiv f_a$ and $T(\mu_0, x) = x$, for all $x \in \Xi$, where Ξ is defined in assumption 2.B below and $\{X_i\}, \{W_i\}$ are mutually independent.*

2. *X_i may depend on W_i but (X_i, W_i) is independent of (X_j, W_j) for $i \neq j$, X_1 and W_1 may have different densities but the transformation T is twice boundedly differentiable on $\Theta \times \Xi$. Moreover, a function T^{-1} exists such that for all x: $x = T^{-1}(\mu, T(\mu, x))$. It is also assumed that T_2, the partial derivative of T with respect to its second argument, is bounded away from zero on $\Theta \times \Xi$.*

The conditions in Assumption 2.A are imposed for the following reasons. That r, r_a are twice boundedly differentiable allows us to carry out a second order expansion to the loss function.

The condition on q, q_a is needed to obtain a desirable convergence rate for the variance of $\hat{r}, \hat{r}_a$. In $\mathcal{C}_1$, the conditions are stronger and therefore convergence of $\hat{\theta}$ to θ_0 is faster than in $\mathcal{C}_2$. Indeed, in the former case, we need not expand T, as $T(\mu_0, x) = x$.

To simplify notation we assume $N_a = N$, but all results go through when N_a and N are related only in the way described in the assumption.

Assumption 2.B *The twice boundedly differentiable weight function w, is non-negative and positive only on the interior of a compact interval Ξ. For all points $x \in \Xi$ we have that $f(x) > 0$ and for all $(\theta, x) \in \Theta \times \Xi$ that $f_a(T(\mu, x)) > 0$. No parameter vector $\theta \neq \theta_0$ exists such that $m(x) = S(\xi, m_a(T(\mu, x)))$ for almost all $x \in \Xi$.*

Evidently, Ξ should be a subset of the support of X_1. The condition that w is twice boundedly differentiable on Ξ is again imposed to apply an expansion to the loss function. The last condition in Assumption 2.B ensures identifiability of θ_0.

There is a variety of ways to choose w and Ξ. One way is to choose w to be polynomial on Ξ, such that w and w' are zero at the boundaries of their support. We have chosen to limit Ξ to be compact, which makes the proofs somewhat easier. If one is willing to impose certain additional conditions on w, it is likely that a similar result could be obtained for $\Xi = \Re$. In practice, as noted earlier, one will usually prefer to numerically integrate over a bounded area, though. One could dispense with numerical integration altogether by using either regressor density, instead of w. Indeed, in such a situation, one would try to estimate $E\Lambda^2(X_1, \theta)$, with $\Lambda(x, \theta) = f_a(T(\mu, x))r(x) - \xi_1 f(x) f_a(T(\mu, x)) - \xi_2 f(x) r_a(T(\mu, x))$. Such an estimate could take the form $L_{NE}(\theta) = \frac{1}{N}\sum_i \left\{\hat{f}_a(T(\mu, X_i))\hat{r}(X_i) - \xi_1 \hat{f}(X_i)\hat{f}_a(T(\mu, X_i)) - \xi_2 \hat{f}(X_i)\hat{f}_a(T(\mu, X_i))\right\}^2$, in which case $\hat{\theta}$ would be defined as the value of θ, for which $L_{NE}(\theta)$ is minimised. The obvious advantage is the ease of computation. There is however no reason why f should be a better weight function than other allowed choices of w. Indeed, one might choose w to be some other

density, and draw random numbers from that distribution. Obviously, there is any number of possibilities.

Assumption 2.C *We assume that both estimates employ the same kernel* $k \in \mathcal{K}_2$

An example of such a kernel is the Gaussian density, $k(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)$.

Assumption 2.D *We assume that both estimates employ the same bandwidth* h, *where*

$$Nh^5 \to \infty, \quad Nh^6 \to 0, \quad \text{as } N \to \infty, \quad \text{for } C_1,$$

$$Nh^{\frac{10}{3}} \to \infty, \quad Nh^5 \to 0, \quad \text{as } N \to \infty, \quad \text{for } C_2.$$

2.2.3 Convergence of Parameter Estimates

In the theorem below, we need first and second order partial derivatives of L_N with respect to θ. The partial derivative of L_N with respect to θ is

$$\frac{\partial L_N}{\partial \theta}|_\theta = \int \Lambda_N(x,\theta)\lambda_N(x,\theta)w(x)dx, \tag{2.1}$$

where (omitting arguments)

$$\lambda_N = -2 \begin{bmatrix} \hat{f}\hat{f}_a \\ \hat{f}\hat{r}_a \\ (\xi_1 \hat{f}\hat{f}'_a + \xi_2 \hat{f}\hat{r}'_a - \hat{f}'_a \hat{r})T_1 \end{bmatrix}, \tag{2.2}$$

where T_1 denotes the partial derivative of T with respect to its first argument. The Hessian of $L_N(\theta)$ is given by

$$\frac{\partial^2 L_N}{\partial\theta\partial\theta^T}|_\theta = \int \lambda_N \lambda_N^T w - 2\int \Lambda_N \begin{bmatrix} 0 & 0 & \hat{f}\hat{f}'_a T_1^T \\ 0 & 0 & \hat{f}\hat{r}'_a T_1^T \\ \hat{f}\hat{f}'_a T_1 & \hat{f}\hat{r}'_a T_1 & \lambda_N^* \end{bmatrix} w, \tag{2.3}$$

where $\lambda_N^* = (\xi_1 \hat{f}\hat{f}''_a + \xi_2 \hat{f}\hat{r}''_a - \hat{f}''_a \hat{r})T_1 T_1^T + (\xi_1 \hat{f}\hat{f}'_a + \xi_2 \hat{f}\hat{r}'_a - \hat{f}'_a \hat{r})T_{11}$, with $T_{11} = \partial^2 T/(\partial\mu\partial\mu^T)$.

Theorem 2.1 *Under Assumptions 2.A-2.D*

$$N^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Sigma), \text{ in } C_1,$$

for some finite variance matrix Σ, *and*

$$N^{\frac{2}{5}}(\hat{\theta} - \theta_0) \xrightarrow{P} 0, \text{ in } C_2.$$

2.2.4 Parameters can be Replaced by their Estimates

In this subsection we will show that it makes no difference asymptotically whether one uses the true parameter values or their estimates. Suppose we have obtained an estimate $\hat{\theta}$ for θ_0 on the basis of the procedure described in Subsections 2.2.1 to 2.2.3. By Theorem 2.1, we have $\hat{\theta} - \theta_0 = o_p(N^{-\frac{2}{5}})$. The kernel estimates used to obtain $\hat{\theta}$ need not use the same bandwidth as the kernel estimates we will use in Section 2.3 to obtain our pooled estimate. Therefore, the kernel estimates used in this section are not subject to Assumption 2.D. Indeed, we will allow $\hat{r}$ and $\hat{f}$ to use a bandwidth different from the one used by $\hat{r}_a$ and $\hat{f}_a$.

Assumption 2.E *h and h_a converge at the same rate, $Nh^3 \to \infty$ and $h \to 0$, as $N \to \infty$.*

Assumption 2.E ensures that $\hat{r}_a$ and $\hat{f}_a$ converge to r_a and f_a respectively, uniformly in $\theta \in \Theta$ and for all $x \in \Xi$, by Lemmas 2.1 and 2.2; this follows from an argument similar to that the one applying to expression (2.12).

Theorem 2.2 *Let Assumptions 2.A-2.C,2.E hold. If* $\hat{\theta} - \theta_0 = o_p(N^{-\frac{2}{5}})$, *and if* $\tilde{m}(x, \theta) = \xi_1 + \xi_2 \hat{m}_a(T(\mu, x))$, *for all* x, θ, *then*

$$\tilde{m}(x, \hat{\theta}) - \tilde{m}(x, \theta_0) = o_p(N^{-\frac{2}{5}}),$$

for all $x \in \Xi$.

2.3 Pooling Kernel Estimates

2.3.1 Setting

In the present setting we will show that pooling kernel regression estimates leads to a smaller asymptotic mean squared error. The formula for the asymptotic mean squared error of a kernel regression estimate, not involving a higher order kernel, is [cf. Mack (1981)]

$$\text{Bias}^2 + \text{Variance} = \left[\frac{c_1 h^2\{m''(x)f(x) + 2m'(x)f'(x)\}}{2f(x)}\right]^2 + \left[\frac{c_2\sigma^2(x)}{Nhf(x)}\right], \tag{2.4}$$

where $c_1 = \int k(u)u^2du$, $c_2 = \int k^2(u)du$ and $\sigma^2(x) = E[U_1^2|X_1 = x]$. After renorming, the first term in (2.4) is the squared expectation of the asymptotic distribution of $\hat{m} - m$, and the second term the variance. The bias-term is based on a Taylor expansion, where all the terms of smaller order of probability are dropped. Obviously, if $Nh^5 \to 0$, the first term in (2.4) is of lower order than the second, and as a consequence, the expectation of the asymptotic distribution is zero. By the same token, if $Nh^5 \to \infty$, the asymptotic distribution will be degenerate, because its variance is zero.

There are reasons for using the asymptotic mean squared error rather than for instance the asymptotic variance or the (normal) mean squared error ($E[\hat{m}(x) - m(x)]^2$). Because $\hat{f}$, the denominator of $\hat{m}$, can be arbitrarily close to, or even equal to, zero the expectation $E[\hat{m}(x) - m(x)]^2$ need not exist for any finite N. This makes the mean squared error undesirable as a tool in the present setting, as we wish to keep the class of allowed kernel regression estimates as large as possible. The asymptotic variance is, as we have seen above, only suitable if $Nh^5 \to 0$. This restriction on the bandwidth may not always be desirable. Indeed, for the infeasible choice of bandwidth that minimises the asymptotic mean squared error, for instance, we have that $Nh^5 \to c$, for some c with $0 < c < \infty$. There may also be other reasons to have the bandwidth converge at a slower rate, depending on the implementation of the kernel estimates in question.

If $Nh^5 \not\to 0$, however, we will need extra restrictions on the model for our results to go through, as will become evident, further below.

Evidently, if measures other than the asymptotic mean squared error are used, the optimally pooled estimate will be different. Indeed, estimates that are optimally pooled according to the asymptotic mean squared error criterion may well not be optimal according to other optimality criteria. Moreover, instead of applying the optimality criterion at each point separately, one could consider applying one across the whole support.

Because of Theorem 2.2 we can act as if the parameter vector θ_0 is fully known. Thus, if $Y_i^* = (Y_i - \xi_{01})/\xi_{02}$, $X_i^* = T(\mu_0, X_i)$ and $U_i^* = U_i/\xi_{02}$, then we can write

$$\begin{aligned} Y_i^* &= m_a(X_i^*) + U_i^* \quad i = 1, \ldots, N \\ Z_i &= m_a(W_i) + V_i \quad i = 1, \ldots, N_a. \end{aligned}$$

We define

$$\hat{m}_a^*(x) := \frac{\hat{r}_*(x)}{\hat{f}_*(x)},$$

where

$$\hat{f}_*(x) = \frac{1}{Nh}\sum_i k_h(x - X_i^*),\ \hat{r}_*(x) = \frac{1}{Nh}\sum_i k_h(x - X_i^*)Y_i^*.$$

Let x be the point at which we wish to estimate m_a. We will require that Assumptions 2.A (for C_2) and 2.C hold. We also need the following two assumptions.

Assumption 2.F *Let $f_*(u) = f(T^{-1}(\mu_0, u))$, for all u. The joint density of (X_1^*, W_1) is not degenerate and both f and f_* are strictly positive at x.*

Assumption 2.F is to ensure that $\mathrm{Cov}(\hat{m}_a^*, \hat{m}_a)$ is of lower order than both $V\hat{m}_a^*$ and $V\hat{m}_a$. Let h_a denote the bandwidth used by $\hat{m}_a$ and h that used by $\hat{m}_a^*$.

Assumption 2.G *We require that Assumptions 2.C and 2.E hold. Further, at least one of the following three conditions holds.*

C_3: *At the point x at which we wish to estimate m_a, we have that either both $h \sim h_a$ as $N \to \infty$ and $f'_*(x)/f_*(x) - f'_a(x)/f_a(x) = 0$, or that $m''_a(x)$ is known.*

C_4: *$Nh^5 \to 0$, as $N \to \infty$.*

C_5: *$Nh^5 \to \infty$, as $N \to \infty$ and m''_a satisfies a Lipschitz condition of degree one at x.*

The three conditions in Assumption 2.G will be used in Theorem 2.3.

2.3.2 Results

In Theorem 2.3 below, we derive the optimal (in terms of the asymptotic mean squared error) linear combination of $\hat{m}_a$ and $\hat{m}^*_a$. In case $f_a(x) = f_*(x)$, almost everywhere, one might also consider $\hat{m}_d(x) = \{d_1\hat{r}_*(x) + d_2\hat{r}_a(x)\}/\{d_1\hat{f}_*(x) + d_2\hat{f}_a(x)\}$, with $d_1 + d_2 = 1$. This will not generally lead to an improvement of the asymptotic mean squared error. In the special case that $N_a = N$, and $m_a(x) = m(x)$, $V[Z_i|W_i = x] = V[U_i|X_i = x]$, and all regressors are independent, within and across samples, the squared asymptotic bias will in both cases be equal to the first term on the right hand side in (2.4), and the asymptotic variance to the second, divided by two, because $\hat{m}_d$ is, under the present circumstances, just a kernel regression estimate based on twice the number of observations, whilst $V[\{\hat{m}_a(x) + \hat{m}(x)\}/2] = V[\hat{m}(x)]/2$.

Theorem 2.3 *Under Assumptions 2.F and 2.G the linear combination of $\hat{m}^*_a$ and $\hat{m}_a$ that minimises (2.4) is given by*

$$\hat{m}_{p,\Omega^*}(x) = \Omega^*\hat{m}^*_a(x) + (1 - \Omega^*)\hat{m}_a(x),$$

where

$$\Omega^* = \frac{\mathcal{V} - \mathcal{B}(\mathcal{B}_* - \mathcal{B})}{(\mathcal{B}_* - \mathcal{B})^2 + \mathcal{V}_* + \mathcal{V}}, \tag{2.5}$$

where $\mathcal{V} = V\hat{m}_a(x) = c_2\sigma_a^2(x)/(Nhf_a(x))$, $\mathcal{V}_* = V\hat{m}_a^*(x) = c_2\sigma_*^2(x)/(Nhf_*(x))$, $\mathcal{B} = E\hat{m}_a(x) - m_a(x) = c_1h^2\{m_a''(x)f_a(x) + 2m_a'(x)f_a'(x)\}/(2f_a(x))$ *and* $\mathcal{B}_* = E\hat{m}_a^*(x) - m_a(x) = c_1h^2\{m_a''(x)f_*(x) + 2m_a'(x)f_*'(x)\}/(2f_*(x))$, *where* f_* *is the density of* X_1^* *and* σ_*^2 *the variance of* U_1^* *given* $X_1^* = x$.

It may seem surprising that the optimal weights can be less than zero or greater than one. However, weights of less than zero or greater than one are found in other settings, also [cf. e.g. Samuel-Cahn (1994)]. It is not a situation that is likely to arise very often in practice, though. The asymptotic variances and biases used in Theorem 2.3 are unknown, but we may estimate them.

Theorem 2.4 *Under the conditions of the previous subsection, the weight* Ω^* *defined in Theorem 2.3 can be consistently estimated (by* $\hat{\Omega}$, *say) and the feasible pooled estimate,* $\hat{m}_{p,\hat{\Omega}}(x) = \hat{\Omega}\hat{m}_a^*(x) + (1-\hat{\Omega})\hat{m}_a(x)$, *is as efficient — in terms of the asymptotic mean squared error, as given by (2.4) — as the infeasible pooled estimate* $\hat{m}_{p,\Omega^*}(x)$.

There is one important question that remains, namely: How precise must $\hat{\Omega}$ be to improve, in the sense of leading to a reduction of the asymptotic mean squared error, on not doing any pooling at all? Or in other words, just how robust is our gain against poor estimation? We will get some insight in this from the Corollary to the Theorem below. In the Theorem, we derive, for given Ω_0, the interval, in every Ω of which the asymptotic mean squared error of the pooled estimate, $M_p(\Omega)$, is less than or equal to $M_p(\Omega_0)$. In the Corollary, we establish the circumstances under which the pooled estimate is better than both "marginal" estimates, and also better than their naive average, $\hat{m}_m(x) = (\hat{m}_a(x) + \hat{m}_a^*(x))/2$. We wish to point out that the asymptotic mean squared error ignores the impact of the estimation of θ_0. It is thus hazardous to make strong assertions on the basis of the below Theorem or its Corollary.

Theorem 2.5 *Given* Ω_0, *the set* $\Upsilon(\Omega_0)$, *consisting of all values of* $\Omega \in \Re$ *such that* $M_p(\Omega) \leq M_p(\Omega_0)$, *is given by* $\Upsilon(\Omega_0) = [\Omega_0, 2\Omega^* - \Omega_0]$, *when* $\Omega^* \geq \Omega_0$, *and by* $[2\Omega^* - \Omega_0, \Omega_0]$, *otherwise.*

Corollary 2.1 *The pooled regression estimate reduces to either marginal kernel regression estimate, when* $\Omega = 0$, *or* $\Omega = 1$, *and it is the naive pooled estimate* $\hat{m}_m$, *when* $\Omega = \frac{1}{2}$. *For* $\Omega^* \in (0,1)$, *it has an asymptotic mean squared error no higher than that of both the marginal estimates, when* $\Omega \in \Upsilon(0) \cap \Upsilon(1) = [2\Omega^* - 1, 2\Omega^*]$, *and for any* $\Omega^* \in \Re$ *it is at least as good as* $\hat{m}_m$, *if and only if* $\Omega \in \Upsilon(\frac{1}{2})$.

In the next subsection we will examine ways to estimate the conditional variance of V_1 given W_1.

2.3.3 Variance Estimation

In Theorem 2.4 we, temporarily, ignored the issue of how to estimate σ_*^2 and σ_a^2. We will now present ways to estimate $\sigma^2(x) = E[U_1^2|X_1]$, where $\sigma_a^2(x)$ and $\sigma_*^2(x)$ can be estimated in the same fashion.

A standard estimate is

$$\hat{\sigma}^2(x) = \frac{\sum_i k_h(x - X_i)(Y_i - \hat{m}(x))^2}{\sum_i k_h(x - X_i)}. \tag{2.6}$$

A disadvantage of the estimate in (2.6) is that it is rather sensitive to changes in the scale of m. This can be most easily seen when we write (omitting arguments) $\hat{\sigma}^2 = \sum k_h U^2 / \sum k_h + \sum k_h\{(m - \hat{m})^2 + 2(m - \hat{m})U\} / \sum k_h$. The latter term in the above expansion is a nuisance term and is the obvious cause of the afore-mentioned sensitivity to changes in scale of m.

For homoskedastic disturbances a procedure of Hall and Titterington (1986) for nonstochastic regressors can be modified to the stochastic regressor case, as follows.

Theorem 2.6 *Denote by N_D the number of X_i lying in a given interval $[a,b]$ in the support of X_1. Denote by $X_{(1)},\ldots,X_{(N_D)}$, the order statistics of the X_i within $[a,b]$ and denote by $Y_{(i)}, U_{(i)}$, the Y and U corresponding to $X_{(i)}$. Then*

$$\hat{\sigma}^2 = \frac{2}{N_D}\sum_{i=1}^{N_D}(Y_{(i)} - Y_{(i-1)})^2. \tag{2.7}$$

is a consistent estimate of σ^2.

2.4 Other Issues

An issue we have completely ignored so far, is that when the true transformation is not of the form envisaged, the resulting pooled estimates will not be consistent. One thus needs to be fairly confident of the existence of the imposed relationship between the two regression functions. It is possible to test whether the two regression functions do have the shape similarity expected. When there is shape-invariance, $L(\theta_0) = 0$. The converse is not generally true, although if w is positive on a sufficiently wide interval, a small value of $L_N(\hat{\theta})$ would be very reassuring.

Indeed, if one would extend Theorem 2.1 to include the case $\Xi = \Re$, for certain w, a consistent test could be created, i.e. a test that always rejects asymptotically whenever the null hypothesis does not hold, and always accepts when it does. For, in such a case $L(\theta_0) = 0$ if and only if there is shape invariance of the specified form, and $L_N(\hat{\theta})$ would be a consistent estimate of $L(\theta_0)$ if the above extension were feasible.

2.5 Simulations

There is an unlimited number of ways to choose m and m_a, and the procedure will not work equally well for all methods chosen.

The procedure seems particularly appealing when the number of observations is small, because the (marginal) nonparametric estimates will then be imprecise. On the other hand, when N is small, the parameter vector θ will be estimated inaccurately, also, and the pooled estimate may thus well be less accurate than either marginal estimate. When the number of observations is large, an improvement in precision is likely, but less important than in the case the number of observations were small. It is certainly true, however, that the larger the number of similarly shaped regression functions to be estimated, the more precise the pooled estimate will be. By the same token, the larger the number of parameters to be estimated, the smaller the gains will be.

The experiments are fairly modest in size. The combination of numerical integration, numerical optimisation, and nonparametric estimates rendered experiments on a large scale impossible. We have only considered the case with two regression functions to be estimated, and we have set the vector μ to 0, such that $m(x) = \xi_{01} + \xi_{02} m_a(x)$, at all x. The interval over which m and m_a were compared was close to $[-1, 1]$, and the integrals to be evaluated, were approximated numerically.

We have tried a variety of different parameter combinations, and have used sample sizes of 100 and 500 observations. The U_i's and V_i's were Gaussian, and generated by means of the Box-Müller method, and so were the X_i's and W_i's. In the first step, θ was estimated. Then $\hat{m}$ was transformed, and the naive and optimal pooled estimates determined. This procedure was replicated 2500 times for each data set, and the average mean squared error recorded for each of the four estimates of m_a.

For 100 observations, θ was always estimated rather inaccurately for at least a fraction of the samples, resulting in enormous average mean squared errors for all but $\hat{m}_a$. We have tried letting the variances of U_i and V_i get very small, but that resulted in all estimates doing well,

and there not being much difference between the performance of individual estimates.

For 500 observations, the results were rather different. We estimated three different functions, $m_a(x) = 2\exp\{-x^2\} + 3$, $m_a(x) = 2\cos(0.5\pi x) + 3$, $m_a(x) = 2\sin(0.5\pi x) + 3$, at all x, and set $\xi_{01} = -0.5$, $\xi_{02} = 0.5$. We set the standard deviations of U_1 and V_1 either to 0.5 and 2.0, or to 0.5 and 0.2, respectively. These choices and functional forms are arbitrary, and different functional forms will lead to different, quite possibly worse results.

Figures 1 to 6 in Appendix B.1 represent the changing average mean squared error we computed over our 2500 replications for the, for all estimates, 'optimal' bandwidth; optimal in the sense that for each the average mean squared error was minimal. Indeed, we computed $\text{AMSE}(x) = \frac{1}{2500}\sum_{j=1}^{2500}\{\hat{m}_{ajl}(x) - m_a(x)\}^2$, at all knot points used in the computation of the integrals, with $\hat{m}_{ajl}(x)$ is the $l = 1, 2, 3, 4$-th estimate of m_a in replication j at x.

Figures 1 to 3 correspond to models in which the variances of U_1 and V_1 are 0.25 and 4, respectively. When $Nh^5 \to 0$, as $N \to \infty$, the the variance terms in (2.5) tend to zero at a slower rate than the squared asymptotic bias terms. We have assumed this is indeed the case, which admittedly is arbitrary in view of our fixed sample size, and ignored the bias terms in (2.5). We thus estimated Ω^* by

$$\hat{\Omega} = \frac{\hat{V}V_1}{\hat{V}V_1 + \hat{V}U_1/\hat{\xi}_{02}^2},$$

where the variances were estimated by the method described in Section 2.3.3. So we should expect estimates of Ω^* of around $\frac{4}{4+0.25/0.5^2} = 0.8$.

By far the worst performer in the results represented in Figure 1 is the transformed estimate of m. The performance of the naive pooled estimate is also poor, and the optimal pooled estimate does slightly better than $\hat{m}_a$. This seems somewhat surprising if the values of $\hat{\Omega}$ are close to 0.8, as this would imply the performance of the optimal pooled estimate to be between the naive pooled estimate and the transformed estimate of m. A possible explanation is that

some of the estimates of ξ_{02} are much less than 0.5. Indeed, when $\hat{\xi}_{02} < 0.25$, and the variances of U_1 and V_1 are estimated accurately, $\hat{\Omega}$ will be less than 0.5, and the optimal pooled estimates will be between the naive pooled estimate and $\hat{m}_a$. When $\hat{\xi}_{02}$ is so far away from the true value of ξ_{02}, however, the transformed estimate of m is likely to be very imprecise, also. Although in this case, the impact of a poorly estimated $\hat{\xi}_{02}$ is not so serious, the above effect may, in a different situation, work to the optimal pooled estimate's disadvantage, also.

The results depicted in Figure 2 indicate that pooling may indeed be a good idea. For negative x's, the transformed estimate of m performs poorly, whereas for positive x's, this is the case for m_a's direct estimate. For x-values close to -1, the optimal pooled estimate's performance is not as good as the naive pooled estimate, but it performs marginally better everywhere else. A possible explanation is that the optimal pooled estimate will be more strongly influenced by the poor performance of the transformed estimate of m, than the naive pooled estimate.

In Figure 3, the transformed estimate of m performs very poorly for large values of $|x|$, but does quite well for small values. Overall, the performance of the naive pooled estimate seems to be best.

In Figures 4 to 6, the functional forms were the same as in Figures 1 to 3, but the standard deviations of U_1, V_1 were set to 0.5 and 0.2, respectively. We should now expect values of $\hat{\Omega}$ to lie around $0.2^2(0.5^2 \cdot 0.5^{-2} + 0.2^2)^{-1} \approx 0.04$, and consequently, the optimal pooled estimate to be close to $\hat{m}_a$.

This is indeed the case in Figures 4 and 5. In Figure 4, $\hat{m}$ transformed is beyond a doubt the least accurate of the four. The mean squared error curves for $\hat{m}_a$ and the optimal pooled estimate almost coincide, and seem to be somewhat better than that of the naive pooled estimate. Roughly the same pattern can be found in Figure 5.

In Figure 6, the naive pooled estimate seems to be doing best, overall. Both marginal

estimates are performing rather poorly over at least part of the domain.

It seems that on the basis of the limited number of experiments that we have carried out, there may be a basis for pooling the nonparametric estimates in sufficiently large samples, but we did not find evidence that the optimal pooled estimate performs significantly better than the naive pooled estimate. It seems certainly the case that the more precise the estimates of ξ_{01}, ξ_{02}, the more accurate the transformed estimate, and hence the more accurate the pooled estimates will be.

2.6 Summary

In this chapter, we have derived a way of estimating shape invariance parameters in the presence of random regressors. We have moreover shown that the convergence of the estimates is sufficiently fast to allow us to obtain a pooled nonparametric regression estimate that minimises the asymptotic mean squared error.

We have considered two basic cases, one in which the regressor densities are identical and there are no dependencies, and one in which regressor densities may differ and limited dependencies are allowed for. In the former case, we establish that the parameter estimates are $\sqrt{N}$-consistent with a Normal asymptotic distribution, whereas in the latter case the parameter estimates are shown to converge faster than the nonparametric regression estimates. This enabled us to substitute the estimates for the true parameter values without asymptotic significance, enabling us to use data from both data sets to estimate both regression equations simultaneously. We have also examined which was the most efficient (in terms of the asymptotic mean squared error) way of pooling the two data sets, and have pointed out under which circumstances, the (asymptotically) optimal pooled estimate is better than both "marginal" nonparametric regression estimates, and than a naive pooled estimate (i.e. their unweighted

average). We have also provided a more robust way of estimating conditional variances in a nonparametric regression setting, and have given some guide lines how a test for the correctness of the parametric specification of the transformations could be created.

Finally, we carried out some Monte Carlo experiments to find out how useful our method may be in practice. The results seemed to suggest that pooling may be a good idea in sufficiently large data sets, but that the case for using the optimal pooling rule rather than the simpler naive pooling rule is harder to establish.

Appendix

2.A Proofs of Theorems

Proof of Theorem 2.1

Applying the Mean Value Theorem to the first order partial derivative of L at $\hat{\theta}$ we get

$$\frac{\partial L_N}{\partial \theta}|_{\hat{\theta}} = \frac{\partial L_N}{\partial \theta}|_{\theta_0} + \frac{\widetilde{\partial^2 L_N}}{\partial \theta \partial \theta^T}(\hat{\theta} - \theta_0),$$

where the tilde indicates that each of the rows of $\partial^2 L_N/(\partial\theta\partial\theta^T)$ is evaluated at a (possibly different) point in $(\hat{\theta}, \theta_0)$. The quantity on the left hand side in the last displayed equation is zero by the definition of $\hat{\theta}$ and hence

$$\hat{\theta} - \theta_0 = \left[\frac{\widetilde{\partial^2 L_N}}{\partial \theta \partial \theta^T}\right]^{-1} \frac{\partial L_N}{\partial \theta}|_{\theta_0},$$

provided that the inverse exists. Thus, the theorem is proved if we can establish the following six properties.

1. An open and bounded set Θ exists to which θ_0 belongs.

2. L_N is a measurable function in the observations for all $\theta \in \Theta$. Further, $\partial L_N/\partial\theta$ (see (2.1)) exists and is continuous on Θ.

3. $L_N(\theta)$ converges to a non-stochastic function $L(\theta)$ in probability uniformly in $\theta \in \Theta$ and $L(\theta)$ attains a strict global minimum at θ_0.

4. The Hessian of L_N, which is given in (2.3), exists and is continuous on Θ.

5. The above Hessian evaluated at θ_N converges to a non-singular matrix

$$A(\theta_0) = \lim E \frac{\partial^2 L_N}{\partial \theta \partial \theta^T}|_{\theta_0}$$

in probability for any sequence θ_N that converges in probability to θ_0.

6. As $N \to \infty$,

$$N^{\frac{1}{2}} \frac{\partial L_N}{\partial \theta}|_{\theta_0} \xrightarrow{\mathcal{L}} N(0, B(\theta_0)), \text{ in } \mathcal{C}_1,$$

where

$$B(\theta_0) = \lim N E \frac{\partial L_N}{\partial \theta} \frac{\partial L_N}{\partial \theta^T}|_{\theta_0},$$

and

$$N^{\frac{2}{5}} \frac{\partial L_N}{\partial \theta}|_{\theta_0} \xrightarrow{P} 0, \text{ in } \mathcal{C}_2.$$

If all the above conditions are satisfied, then as $N \to \infty$

$$N^{\frac{1}{2}}(\hat{\theta} - \theta_0) \to N(0, A^{-1}(\theta_0)B(\theta_0)A^{-1}(\theta_0)), \text{ in } \mathcal{C}_1,$$

and

$$N^{\frac{2}{5}}(\hat{\theta} - \theta_0) \xrightarrow{P} 0, \text{ in } \mathcal{C}_2.$$

We will establish the above seven properties step by step:

1. This is assumed in Assumption 2.A.

2. The partial derivative, $(\partial L_N/\partial\theta)|_{(\theta)}$, is continuous if k and T are continuously differentiable which was assumed in Assumptions 2.C and 2.A respectively.

3. Define

$$L(\theta) = \int \Lambda^2(x, \theta) w(x) dx,$$

in probability for any sequence θ_N that converges in probability to θ_0.

6. As $N \to \infty$,

$$N^{\frac{1}{2}} \frac{\partial L_N}{\partial \theta}|_{\theta_0} \xrightarrow{\mathcal{L}} N(0, B(\theta_0)), \text{ in } \mathcal{C}_1,$$

where

$$B(\theta_0) = \lim N E \frac{\partial L_N}{\partial \theta} \frac{\partial L_N}{\partial \theta^T}|_{\theta_0},$$

and

$$N^{\frac{2}{5}} \frac{\partial L_N}{\partial \theta}|_{\theta_0} \xrightarrow{P} 0, \text{ in } \mathcal{C}_2.$$

If all the above conditions are satisfied, then as $N \to \infty$

$$N^{\frac{1}{2}}(\hat{\theta} - \theta_0) \to N(0, A^{-1}(\theta_0)B(\theta_0)A^{-1}(\theta_0)), \text{ in } \mathcal{C}_1,$$

and

$$N^{\frac{2}{5}}(\hat{\theta} - \theta_0) \xrightarrow{P} 0, \text{ in } \mathcal{C}_2.$$

We will establish the above seven properties step by step:

1. This is assumed in Assumption 2.A.

2. The partial derivative, $(\partial L_N/\partial\theta)|_{(\theta)}$, is continuous if k and T are continuously differentiable which was assumed in Assumptions 2.C and 2.A respectively.

3. Define

$$L(\theta) = \int \Lambda^2(x, \theta) w(x) dx,$$

where

$$\Lambda(x,\theta) = r(x)f_a(T(\mu,x)) - \xi_1 f(x)f_a(T(\mu,x)) - \xi_2 r_a(T(\mu,x))f(x).$$

We will first show that

$$L_N(\theta) - L(\theta) = o_p(1), \tag{2.8}$$

uniformly in θ over any bounded interval. We have

$$L_N(\theta) - L(\theta) = \int \left\{\Lambda_N^2(x,\theta) - \Lambda^2(x,\theta)\right\} w(x)dx.$$

Now, $\Lambda_N^2 - \Lambda^2 = (\Lambda_N - \Lambda)^2 + 2(\Lambda_N - \Lambda)\Lambda$, so for (2.8) it suffices to show that

$$\sup_{\theta\in\Theta}\sup_{x\in\Xi} |\Lambda_N(x,\theta) - \Lambda(x,\theta)| = o_p(1). \tag{2.9}$$

Because Θ is bounded, so are ξ_1 and ξ_2 in the expansions of Λ_N and Λ and these therefore play no role of importance. We write

$$\Lambda_N - \Lambda = (\hat{r}\hat{f}_a - rf_a) - \xi_1(\hat{f}\hat{f}_a - ff_a) - \xi_2(\hat{f}\hat{r}_a - fr_a). \tag{2.10}$$

We will now show that

$$\sup_{\theta\in\Theta}\sup_{x\in\Xi} \left|\hat{r}_a\hat{f} - r_a f\right| = o_p(1), \tag{2.11}$$

where the other terms in (2.10) can be dealt with in a similar manner. We can rewrite (2.11) as

$$\sup_{\theta\in\Theta}\sup_{x\in\Xi} \left|(\hat{r}_a - r_a)(\hat{f} - f) + (\hat{r}_a - r_a)f + (\hat{f} - f)r_a\right| = o_p(1). \tag{2.12}$$

θ appears only in the argument of r_a and $\hat{r}_a$ in (2.12) and because Lemmas 2.1 and 2.2 (all lemmas as stated and proved in the appendix) hold uniformly on $\Re$, (2.12) holds also.

That L attains a strict global minimum at θ_0 is implied by $L(\theta_0) = 0 \leq L(\theta)$ for all θ, Assumption 2.B and the obvious fact that $L(\theta)$ can only be zero if $m(x) = S(\xi, a(T(\mu, x)))$, for almost all $x \in \Xi$.

4. Existence and continuity are implied by existence and continuity of the second order derivatives of T and k and the compactness of Ξ.

5. We will show that the matrix $A(\theta_0)$ is given by

$$A(\theta_0) = \int \lambda(x, \theta_0)\lambda(x, \theta_0)^T w(x)dx,$$

where

$$\lambda(x, \theta) := -2 \begin{bmatrix} ff_a \\ fr_a \\ (\xi_1 ff_a' + \xi_2 fr_a' - f_a' r)T_1 \end{bmatrix}.$$

It is sufficient to show that

$$\int \{\lambda_N(x, \theta_N)\lambda_N^T(x, \theta_N) - \lambda(x, \theta_N)\lambda^T(x, \theta_N)\}w(x)dx = o_p(1), \tag{2.13}$$

$$\int \{\lambda(x, \theta_N)\lambda^T(x, \theta_N) - \lambda(x, \theta_0)\lambda^T(x, \theta_0)\}w(x)dx = o_p(1), \tag{2.14}$$

and if we call the matrix under the second integral in (2.3) M_N,

$$\int \{\Lambda_N(x, \theta_N)M_N(x, \theta_N) - \Lambda(x, \theta_N)EM_N(x, \theta_N)\}w(x)dx = o_p(1), \tag{2.15}$$

$$\int \Lambda(x, \theta_N)EM_N(x, \theta_N)w(x)dx = o_p(1). \tag{2.16}$$

We will prove (2.13) through (2.16) by demonstrating that the results hold for every element in the matrices. For (2.13) we need to show that $\int(\hat{f}^2\hat{f}_a^2 - f^2 f_a^2)w = o_p(1)$,

$\int(\hat{f}^2\hat{f}_a\hat{r}_a - f^2 f_a r_a)w = o_p(1), \ldots, \int\{(\hat{f}'_a)^2 r^2 - (f'_a)^2 r^2\}T_1 T_1^T w = o_p(1)$. All these conditions can be established in similar fashion. Because convergence of derivative estimates is slower than of estimates of the original function, we will show convergence of a term which includes a derivative estimate. Consider $Q := \int(\hat{f}\hat{r}\hat{f}'_a\hat{r}'_a - f r f'_a r'_a)T_1 w$. Because we know the convergence properties of the separate estimates, we will split Q up using the basic algebraic property

$$\prod_{j=1}^{4}\hat{D}_{4j} - \prod_{j=1}^{4} D_{4j} = \sum_{i_1=0}^{1}\sum_{i_2=0}^{1}\sum_{i_3=0}^{1}\sum_{i_4=0}^{3-i_1-i_2-i_3}\prod_{j} D_{4j}^{i_j}(\hat{D}_{4j} - D_{4j})^{1-i_j}.$$

We know from Lemma 2.3 and Assumption 2.D that $\sup_x|\hat{f} - f| + \sup_x|\hat{r} - r| = O_p(h^2 + N^{-\frac{1}{2}}h^{-1}) = o_p(1)$. We also know that f, r, f'_a, r'_a are all bounded. It is therefore sufficient to show that

$$\int[(\hat{f}'_a - f_a)'(\hat{r}'_a - r'_a) + (\hat{f}'_a - f'_a) + (\hat{r}'_a - r'_a)]\zeta = o_p(1),$$

for any bounded function ζ. By the inequality of Cauchy-Schwarz we only need to show that

$$\int(\hat{f}'_a - f'_a)^2\zeta = o_p(1), \int(\hat{r}'_a - r'_a)^2\zeta = o_p(1).$$

We will only show that the last of these two conditions holds. Taking the expectation leads to

$$\int E[\hat{r}'_a - r'_a]^2\zeta = \int\{V[\hat{r}'_a] + (E[\hat{r}'_a] - r'_a)^2\}\zeta = O(h^2 + N^{-1}h^{-3}),$$

by Lemmas 2.1 and 2.4.

Condition (2.14) follows from Slutsky's Theorem, from the fact that $\int\lambda\lambda^T w$ is continuous in θ and from $\theta_N - \theta_0 = o_p(1)$. To show (2.15) note that

$$\Lambda_N M_N - \Lambda E M_N = (\Lambda_N - \Lambda)(M_N - EM_N) + \Lambda(M_N - EM_N) + EM_N(\Lambda_N - \Lambda). \quad (2.17)$$

We will restrict ourselves to showing convergence of the second term on the right hand side in (2.17), arguably the hardest part. The other terms follow by a similar argument. We will prove convergence of each element in M_N separately. Because second order derivatives converge slowest, we will just show convergence for an element involving a second order derivative; the rest of the elements can be handled in identical fashion. We will thus show that $\int(\hat{f}\hat{r}_a'' - E[\hat{f}\hat{r}_a''])\Lambda w = o_p(1)$. By Lemma 2.6 it suffices to show that: $\int(\hat{f}\hat{r}_a'' - E\hat{f}E\hat{r}_a'')\Lambda w = o_p(1)$ because $E[\hat{f}\hat{r}_a''] - E\hat{f}E\hat{r}_a'' = O(N^{-1}h^{-3}) = o(1)$. Because $\sup_x |\hat{f}(x) - E\hat{f}(x)| = o_p(1)$, by Lemma 2.2 and Assumption 2.D, we need to show that

$$\int(\hat{r}_a'' - E\hat{r}_a'')E\hat{f}\Lambda w = o_p(1). \tag{2.18}$$

Because $\hat{r}_a'' = \frac{1}{Nh^3}\sum_i k_h''(T(\mu_N, x) - W_i)Z_i$, we are interested in

$$\int k_h''(T(\mu_N, x) - W_i)\Lambda(\theta_N, x)E\hat{f}(x)w(x)dx = \int k_h''(u - W_i)\zeta(u)du, \tag{2.19}$$

where the equality follows from the substitution of $u = T(\mu_N, x)$, and the fact that T_2 is assumed bounded away from zero on $\Theta \times \Xi$, defining $\zeta(u)$ by

$$\zeta(u) = \frac{\Lambda(\theta_N, T^{-1}(\mu_N, u))E\hat{f}(T^{-1}(\mu_N, u))w(T^{-1}(\mu_N, u))}{T_2(T^{-1}(\mu_N, u))}.$$

Because w (and hence w') is zero outside Ξ, ζ and ζ' are zero outside $T(\mu_N, \Xi)$. We rewrite (2.19) as (using partial integration twice)

$$-h[k_h'(u - W_i)\zeta(u)]_{-\infty}^{\infty} + h^2[k_h(u - W_i)\zeta'(u)]_{-\infty}^{\infty} + h^2\int k_h(u - W_i)\zeta''(u)du. \tag{2.20}$$

As noted before ζ and ζ' are zero outside the integration area, so the first two terms in (2.20) are zero. Substitution of $v = (u - W_i)/h$ in the last term in (2.20) gives

$$h^3\int k(v)\zeta''(W_i - hv)dv = h^3\rho(W_i),$$

which implicitly defines ρ. By the boundedness of ζ'', $\rho(W_i)$ is bounded. The left hand side in (2.18) now reads

$$\begin{aligned}
&\frac{1}{Nh^3}\sum_i \int k_h''(T(\mu_N,x)-W_i)\Lambda(x,\theta_N)E\hat{f}(x)w(x)dxZ_i \\
&- E\left[\frac{1}{Nh^3}\sum_i \int k_h''(T(\mu_N,x)-W_i)\Lambda(x,\theta_N)E\hat{f}(x)w(x)dxZ_i\right] \\
&= \frac{1}{N}\sum_i \rho(W_i)Z_i - E\left[\rho(W_1)Z_1\right] = O_p(N^{-\frac{1}{2}}),
\end{aligned}$$

because the $\rho(W_i)Z_i$ are i.i.d. with finite variance. So (2.18) holds and so does (2.15). All that is left is (2.16). By Lemma 2.1, $E\hat{r}_a'' = O(1)$ uniformly in x and hence so is EM_N. So all we really need to do is to show that

$$\int \Lambda(x,\theta_N)w(x)dx - \int \Lambda(x,\theta_0)w(x)dx = o_p(1), \tag{2.21}$$

noting that $\Lambda(x,\theta_0) = 0$ for all x. But because $\int \Lambda(x,\theta)w(x)dx$ is a continuous function and because $\theta_N - \theta_0 = o_p(1)$ Slutsky's Theorem gives (2.21).

6. For C_2 we need to prove that:

$$\int \Lambda_N(x,\theta_0)\lambda_N(x,\theta_0)w(x)dx = o_p(N^{-\frac{2}{5}}) \tag{2.22}$$

We shall do this by proving the following three conditions that are together sufficient for (2.22).

$$\int \Lambda_N(x,\theta_0)\{\lambda_N(x,\theta_0)-\lambda(x,\theta_0)\}w(x)dx = o_p(N^{-\frac{2}{5}}), \tag{2.23}$$

$$E\int \Lambda_N(x,\theta_0)\lambda(x,\theta_0)w(x)dx = o(N^{-\frac{2}{5}}), \tag{2.24}$$

$$\int \Lambda_N(x,\theta_0)\lambda(x,\theta_0)w(x)dx - E\left[\int \Lambda_N(x,\theta_0)\lambda(x,\theta_0)w(x)dx\right] = o_p(N^{-\frac{2}{5}}). \tag{2.25}$$

We first consider (2.23). We define $D_1 = [r, f, f]^T$, $D_2 = [f_a, f_a, r_a]^T$, $d_1 = [1, -\xi_{01}, -\xi_{02}]^T$, $D_3 = [f, f, f, f, r]^T$, $D_4 = [f_a, r_a, f_a', r_a', f_a']^T$, $d_2 = [-2, -2, \xi_{01}||T_1||, \xi_{02}||T_1||, -||T_1||]^T$, so that we can write

$$\begin{aligned}
\Lambda_N - \Lambda &= \sum_{j=1}^{3} d_{1j}(\hat{D}_{1j}\hat{D}_{2j} - D_{1j}D_{2j}), \\
\lambda_N - \lambda &= \\
&\left[d_{21}(\hat{D}_{31}\hat{D}_{41} - D_{31}D_{41}), d_{22}(\hat{D}_{32}\hat{D}_{42} - D_{32}D_{42}), \sum_{j=3}^{5} d_{2j}(\hat{D}_{3j}\hat{D}_{4j} - D_{3j}D_{4j}) \right]^T.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left| \int \Lambda_N(\lambda_N - \lambda)w \right| \\
&\leq \int \sqrt{(\Lambda_N - \Lambda)^2 ||\lambda_N - \lambda||^2} w \\
&\leq C \int \sqrt{\sum_{j=1}^{3} \left\{ (\hat{D}_{1j} - D_{1j})^2(\hat{D}_{2j} - D_{2j})^2 + D_{1j}^2(\hat{D}_{2j} - D_{2j})^2 + D_{2j}^2(\hat{D}_{1j} - D_{1j})^2 \right\}} \\
&\times \sqrt{\sum_{j=1}^{5} \left\{ (\hat{D}_{3j} - D_{3j})^2(\hat{D}_{4j} - D_{4j})^2 + D_{4j}^2(\hat{D}_{3j} - D_{3j})^2 + D_{3j}^2(\hat{D}_{4j} - D_{4j})^2 \right\}} w,
\end{aligned}$$

for some large $C > 0$. Because $\sup_{x,j} |\hat{D}_{1j} - D_{1j}| + \sup_{x,j} |\hat{D}_{2j} - D_{2j}| + \sup_{x,j} |\hat{D}_{3j} - D_{3j}| + \sup_{x,j} |\hat{D}_{4j} - D_{4j}| = o_p(1)$, $(\hat{D}_{1j} - D_{1j})^2(\hat{D}_{2j} - D_{2j})^2$ and $(\hat{D}_{3j} - D_{3j})^2(\hat{D}_{4j} - D_{4j})^2$ are of lower order than $D_{1j}^2(\hat{D}_{2j} - D_{2j})^2 + D_{2j}^2(\hat{D}_{1j} - D_{1j})^2$ and $D_{3j}^2(\hat{D}_{4j} - D_{4j})^2 + D_{4j}^2(\hat{D}_{3j} - D_{3j})^2$ respectively. So we only need to look at

$$\begin{aligned}
&C \int \sqrt{\sum_{j=1}^{3} \left\{ D_{1j}^2(\hat{D}_{2j} - D_{2j})^2 + D_{2j}^2(\hat{D}_{1j} - D_{1j})^2 \right\}} \\
&\times \sqrt{\sum_{j=1}^{5} \left\{ D_{4j}^2(\hat{D}_{3j} - D_{3j})^2 + D_{3j}^2(\hat{D}_{4j} - D_{4j})^2 \right\}} w. \qquad (2.26)
\end{aligned}$$

By Lemmas 2.1 and 2.4 we know that

$$\begin{aligned}
\sup_{x,j} E[\hat{D}_{1j} - D_{1j}]^2 + \sup_{x,j} E[\hat{D}_{2j} - D_{2j}]^2 &= O(h^4 + (Nh)^{-1}), \\
\sup_{x,j} E[\hat{D}_{3j} - D_{3j}]^2 + \sup_{x,j} E[\hat{D}_{4j} - D_{4j}]^2 &= O(h^2 + N^{-1}h^{-3}),
\end{aligned}$$

and hence the expectation of (2.26) is of order $O(h^3 + N^{-1}h^{-2}) = o(N^{-\frac{2}{5}})$. Now consider (2.24). Lemma 2.6 implies that we can proceed as if the kernel estimates are all based on completely independent samples. Using again $\Lambda(x, \theta_0) = 0$ for all x we write (using Lemma 2.6)

$$E\Lambda_N - \Lambda = E\hat{r}E\hat{f}_a - rf_a - \xi_{01}(E\hat{f}E\hat{f}_a - ff_a) - \xi_{02}(E\hat{f}E\hat{r}_a - fr_a) + O(N^{-1}h^{-1}), \quad (2.27)$$

uniformly in x. We will deal with each of the terms in (2.27) separately. We will just demonstrate the procedure for the first term, where the proof for the other terms follows trivially. We write

$$E\hat{r}E\hat{f}_a - rf_a = (E\hat{r} - r)(E\hat{f}_a - f_a) + r(E\hat{f}_a - f_a) + f_a(E\hat{r} - r). \quad (2.28)$$

By Lemma 2.1, (2.28) is $O(h^2)$, uniformly in x, which with Assumption 2.D implies that (2.28) is $o(N^{-\frac{2}{5}})$. So (2.24) holds. Expression (2.25) is fairly straightforward to deal with. We will limit ourselves to proving

$$\int \left\{\hat{r}(x)\hat{f}_a(T(\mu_0, x)) - E[\hat{r}(x)\hat{f}_a(T(\mu_0, x))]\right\} \lambda(x, \theta_0)w(x)dx = O_p(N^{-\frac{1}{2}}), \quad (2.29)$$

where the result for the other terms can be obtained in the same fashion. Lemma 2.6 states that $E[\hat{r}\hat{f}_a] - E\hat{r}E\hat{f}_a = O(N^{-1}h^{-1})$. Further,

$$\hat{r}\hat{f}_a - E\hat{r}E\hat{f}_a = (\hat{r} - E\hat{r})(\hat{f}_a - E\hat{f}_a) + (\hat{f}_a - E\hat{f}_a)E\hat{r} + (\hat{r} - E\hat{r})E\hat{f}_a. \quad (2.30)$$

The first term on the right hand side in (2.30) is of lower order as seen when verifying (2.23). We will now deal with the last term on the right hand side in (2.30) where the middle term can be dealt with in the very same way.

Note that for any bounded function ζ,

$$\int k_h(x - X_i)\zeta(x)dx = h \int k(u)\zeta(X_i + hu)du = h\rho(X_i),$$

which implicitly defines ρ. We define ζ by $\zeta(x) = \lambda(x, \theta_0)w(x)E\hat{f}_a(T(\mu_0, x))$. Hence

$$\begin{aligned}
&\int \{\hat{r}(x) - E\hat{r}(x)\}E\hat{f}_a(T(\mu_0, x))\lambda(x, \theta_0)w(x)dx \\
&= \int \{\hat{r}(x) - E\hat{r}(x)\}\zeta(x)dx \\
&= \frac{1}{Nh}\sum_i \int k_h(x - X_i)\zeta(x)dxY_i - E\left[\frac{1}{Nh}\sum_i \int k_h(x - X_i)\zeta(x)dxY_i\right] \\
&= \frac{1}{N}\sum_i \rho(X_i)Y_i - E[\rho(X_1)Y_1] = O_p(N^{-\frac{1}{2}}). \qquad (2.31)
\end{aligned}$$

This procedure can be applied in turn to all terms in the expansion of the left hand side of (2.25) and so (2.25) holds. This concludes the proof of (2.22). The proof for $\mathcal{C}_1$ is not very different. We have to show that the left hand sides in (2.23) and (2.24) are $o(N^{-\frac{1}{2}})$ and that the left hand side in (2.25) times $\sqrt{N}$ converges to a normal distribution.

The proof of the first of these three conditions is simple; just apply the bandwidth restrictions of $\mathcal{C}_1$ to the proof of (2.23) for case $\mathcal{C}_2$. The second is not hard either. Note that

$$E\hat{r} = \frac{1}{Nh}\sum_i E[k_h(x - X_i)Y_i] = \frac{1}{h}E[k_h(x - X_1)\{\xi_{01} + \xi_{02}m_a(X_1)\}] = \xi_{01}E\hat{f} + \xi_{02}E\hat{r}_a,$$

and hence

$$E\Lambda_N(x, \theta_0) = E\hat{r}E\hat{f}_a - \xi_{01}E\hat{f}E\hat{f}_a - \xi_{02}E\hat{f}E\hat{r}_a = 0.$$

For the third we again refer to the proof for $\mathcal{C}_2$. From (2.31) and the discussion preceding it, it follows that the left hand side of (2.25) times $\sqrt{N}$ can be written

$$\frac{1}{\sqrt{N}}\sum_i \left\{\sum_{j=1}^{J}(\rho_j(X_i) - E\rho_j(X_1))\right\} \qquad (2.32)$$

where J is some finite positive integer. By the Lindeberg-Levy Central Limit Theorem, (2.32) is asymptotically normal with a finite variance.

Q.E.D.

Proof of Theorem 2.2

Because Θ is an open set and $\hat{\theta}$ is a consistent estimate of $\theta_0 \in \Theta$, $\hat{\theta}$ will (for sufficiently large N) lie in Θ a.s.. So we can assume that $\hat{\theta} \in \Theta$ and that no $\theta \notin \Theta$ exists such that $||\theta - \theta_0|| < ||\hat{\theta} - \theta_0||$.

Because $S, T, \hat{m}_a$ are all differentiable we use the Mean Value Theorem to obtain

$$\tilde{m}(x, \hat{\theta}) - \tilde{m}(x, \theta_0) = (\hat{\theta} - \theta_0)^T \begin{bmatrix} S_1(\xi^*, \hat{m}_a(T(\mu^*, x))) \\ T_1(\mu^*, x)\hat{m}_a'(T(\mu^*, x))S_2(\xi^*, \hat{m}_a(T(\mu^*, x))) \end{bmatrix}, \quad (2.33)$$

where subscripts indicate to which argument the partial derivative was taken and θ^* — which may depend on x — lies between $\hat{\theta}$ and θ_0 and hence in Θ. As a result of the argument just preceding this Theorem and of Slutsky's Theorem, the second factor on the right hand side in (2.33) is bounded in probability. As assumed, the first factor is $o_p(N^{-\frac{2}{5}})$.

Q.E.D.

Proof of Theorem 2.3

The asymptotic mean squared error, as defined in (2.4), of the pooled estimate, as a function of Ω reads

$$\begin{aligned} M_p(\Omega) &= \mathcal{B}_p^2(\Omega) + \mathcal{V}_p(\Omega) \\ &= [\Omega\mathcal{B}_* + (1-\Omega)\mathcal{B}]^2 + \Omega^2\mathcal{V}_* + (1-\Omega)^2\mathcal{V} + 2\Omega(1-\Omega)\mathrm{AsCov}(\hat{m}_a^*, \hat{m}_a). \end{aligned}$$

The asymptotic covariance in the last displayed equation is of lower order than the two variances,

which are $O(N^{-1}h^{-1})$; we will not prove this, but a heuristic argument is easy to give. The principal reason that the above statement is true, is that $E[k_h^2(x - X_1)] = O(h)$ and $E[k_h(x - X_1)k_h(x - W_1)] = O(h^2)$, unless the density of (X_1, W_1) is degenerate. Thus ignoring the last term, expanding the last displayed equation leads to

$$M_p(\Omega) = [\mathcal{V}_* + \mathcal{V} + (\mathcal{B}_* - \mathcal{B})^2]\Omega^2 + 2[\mathcal{B}(\mathcal{B}_* - \mathcal{B}) - \mathcal{V}]\Omega + [\mathcal{B}^2 + \mathcal{V}] \tag{2.34}$$

Minimising $M_p(\Omega)$ with respect to Ω gives (2.5).

Q.E.D.

Proof of Theorem 2.4

Note first that $\mathcal{V}$ and $\mathcal{V}_*$ can be consistently estimated if f_a, f_*, σ_a^2 and σ_*^2 can, where σ_a^2 and σ_*^2 correspond to σ^2 in (2.4). f_* and f_a can be consistently estimated, as follows from Lemma 2.5. The estimation of σ_a^2 and σ_*^2 is discussed in subsection 2.3.3. In case $\mathcal{C}_3$, $(\mathcal{B}_* - \mathcal{B}) = 0$, and hence Ω^* is a function of $\mathcal{V}_*$ and $\mathcal{V}$ only, which are estimable. In case $\mathcal{C}_4$, h^4 tends to zero faster than $(Nh)^{-1}$ and hence the squared bias terms are of lower order than the variance terms and again Ω^* depends (for large N) only on the variances. In case $\mathcal{C}_5$, $f_a, f_a', f_a'', r_a, r_a', r_a'', f_*, f_*', f_*'', r_*, r_*', r_*''$ can all be estimated consistently in view of Lemma 2.5 and hence so can m_a, m_a'', m_*, m_*'' and thence $\mathcal{B}, \mathcal{B}_*, \mathcal{V}, \mathcal{V}_*$.

It remains to be shown that a consistent estimate for Ω^* automatically leads to the same efficiency. We have

$$\begin{aligned}
\hat{m}_{p,\hat{\Omega}}(x) &= \{\hat{\Omega}(x) - \Omega^*(x)\}\hat{m}_a^*(x) + \{\Omega^*(x) - \hat{\Omega}(x)\}\hat{m}_a(x) + \hat{m}_{p,\Omega^*}(x) \\
&= \{\hat{\Omega}(x) - \Omega^*(x)\}\{\hat{m}_a^*(x) - \hat{m}_a(x)\} + \hat{m}_{p,\Omega^*}(x) \\
&= \hat{m}_{p,\Omega^*}(x) + o_p(|\hat{m}_a^*(x) - m_a(x)| + |\hat{m}_a(x) - m_a(x)|),
\end{aligned}$$

because $\hat{m}_a^*$ and $\hat{m}_a$ converge at the same rate to m_a and $\hat{\Omega}$ is consistent for Ω^*.

Q.E.D.

Proof of Theorem 2.5

Because M_p (defined as in (2.34)) is quadratic in Ω, there are two (possibly coinciding) solutions to $M_p(\Omega) = M_p(\Omega_0)$, one of them being $\Omega = \Omega_0$, where we shall call the other one Ω_e. Thus, $M_p(\Omega) - M_p(\Omega_0) = 0$ is equivalent to

$$\{\mathcal{V}_* + \mathcal{V} + (\mathcal{B}_* - \mathcal{B})^2\}\{\Omega^2 - \Omega_0^2\} + 2\{\mathcal{B}(\mathcal{B}_* - \mathcal{B}) - \mathcal{V}\}\{\Omega - \Omega_0\} = 0$$

$$\Leftrightarrow \quad \{\mathcal{V}_* + \mathcal{V} + (\mathcal{B}_* - \mathcal{B})^2\}\{\Omega - \Omega_0\}\{\Omega + \Omega_0\} + 2\{\mathcal{B}(\mathcal{B}_* - \mathcal{B}) - \mathcal{V}\}\{\Omega - \Omega_0\} = 0,$$

such that $\Omega_e = 2\Omega^* - \Omega_0$. If $\Omega^* \geq (\leq)\ \Omega_0$, $2\Omega^* - \Omega_0 \geq (\leq)\ \Omega_0$, and because $\Omega^* \in [\Omega_0, 2\Omega^* - \Omega_0]$ ($\Omega^* \in [2\Omega^* - \Omega_0, \Omega_0]$), $M_p(\Omega)$ is less than $M_p(\Omega_0)$, in all points on the afore-mentioned interval.

Q.E.D.

Proof of Theorem 2.6

Expanding (2.7), we have

$$\begin{aligned}\hat{\sigma}^2 &= \frac{2}{N_D}\sum_{i=1}^{N_D}\{m(X_{(i)}) - m(X_{(i-1)})\}^2 \\ &+ \frac{4}{N_D}\sum_{i=1}^{N_D}\{m(X_{(i)}) - m(X_{(i-1)})\}\{U_{(i)} - U_{(i-1)}\} \\ &+ \frac{2}{N_D}\sum_{i=1}^{N_D}\{U_{(i)} - U_{(i-1)}\}^2. \qquad (2.35)\end{aligned}$$

By $E[U_i|X_j] = 0$, for all i, j, and independence across U_i, the last term in (2.35) is $\sigma^2 + o_p(1)$, as $N \to \infty$. On the other hand, by the Mean Value Theorem, the first term on the right of (2.35) is

$$O\left(\frac{1}{N_D}\sum_{i=1}^{N_D}(X_{(i)} - X_{(i-1)})^2\right) = O_p(N_D^{-1}),$$

and if f is positive on $[a, b]$ this is $O_p(N^{-1})$. From these properties and the Cauchy inequality, the intermediary term in (2.35) is $O_p(N_D^{-\frac{1}{2}})$. It follows that

$$\hat{\sigma}^2 \xrightarrow{P} \sigma^2, \text{ as } N \to \infty.$$

Q.E.D.

2.B Technical Lemmas

The lemmas in this appendix are fairly standard. The assumptions made in the main body of this article, in so far as they concern r, k and f and the conditions on the various variables and the way they are related apply here as well. In this section Ξ should just be read as any compact set on which f is bounded away from zero.

Lemma 2.1 *We have*

$$\sup_{x\in\Re} |E\hat{r}^{(l)}(x) - r^{(l)}(x)| = O(h^{2-l}),\ l = 0, 1, 2.$$

Proof:

We write

$$\sup_x |E\hat{r}^{(l)}(x) - r^{(l)}(x)| = \sup_x \left|\frac{1}{h^{l+1}} E\left[k_h^{(l)}(x - X_1)m(X_1)\right] - r^{(l)}(x)\right|$$

$$= \sup_x \left| \frac{1}{h^{l+1}} \int k_h^{(l)}(x-u) r(u) du - r^{(l)}(x) \right|.$$

For $l > 0$ the partial integration rule can be applied to obtain

$$\sup_x \left| -\frac{1}{h^l} [k_h^{(l-1)}(x-u) r(u)]_{-\infty}^{\infty} + \frac{1}{h^l} \int k_h^{(l-1)}(x-u) r^{(1)}(u) du - r^{(l)}(x) \right|. \tag{2.36}$$

Because $r \in \mathcal{G}_2$, r is bounded and because k integrates to one, $k_h(-\infty) = k_h(\infty) = 0$. Therefore, the first term in (2.36) is zero. For $l = 2$ this step can be repeated. This leaves

$$\sup_x \left| \frac{1}{h} \int k_h(x-u) r^{(l)}(u) du - r^{(l)}(x) \right| = \sup_x \left| \int k(v) \left[r^{(l)}(x - hv) - r^{(l)}(x) \right] dv \right|. \tag{2.37}$$

If $l = 2$ the boundedness of r'' implies that the above expression is $O(1)$. For $l = 0$ we get by a first order Taylor expansion

$$\sup_x \left| \int k(v) [hvr'(x) + h^2 v^2 r''(x - hv; x)] dv \right|,$$

where $(x - hv; x)$ is some number between $x - hv$ and x. Because $k \in \mathcal{K}_2$, $\int k(v) v dv = 0$ and thus by $k \in \mathcal{K}_2$ and $r \in \mathcal{G}_2$ the last displayed expression is $O(h^2)$. For $l = 1$ we write using the Mean Value Theorem

$$h \sup_x \left| \int k(v) [vr''(x - hv; x)] dv \right|.$$

Again, by the assumed boundedness of r'' the last displayed expression is $O(h)$.

Q.E.D.

A remark that should be made is that if r'' is first order Lipschitz-continuous, (2.37) is $O(h)$, and therefore so is $E\hat{r}''(x) - r''(x)$.

Obviously, the above lemma can just as well be applied to $\hat{f}$. This also holds for the following uniform convergence result:

Lemma 2.2 *We have*

$$\sup_x \left| \hat{r}^{(l)}(x) - E\hat{r}^{(l)}(x) \right| = O_p\left(N^{-\frac{1}{2}} h^{-l-1} \right), \; l = 0, 1, 2. \tag{2.38}$$

Proof:

We write the left hand side of (2.38) as

$$\sup_x \left| \frac{1}{Nh^{l+1}} \sum_j \left[k_h^{(l)}(x - X_j)Y_j - E\left(k_h^{(l)}(x - X_j)Y_j \right) \right] \right|.$$

Using $k(x) = \int \phi_k(u) e^{iux}$, we obtain

$$\sup_x \left| \frac{1}{Nh^{l+1}} \sum_j \int u^l \phi_k(u) \left\{ e^{iu\frac{x-X_j}{h}} Y_j - E\left(e^{iu\frac{x-X_j}{h}} Y_j \right) \right\} du \right|. \tag{2.39}$$

After substitution of $v = u/h$, (2.39) can be bounded by

$$\int \left| v^l \phi_k(hv) \right| \sup_x \left| e^{ivx} \right| \left| \frac{1}{N} \sum_j \left\{ e^{-ivX_j} Y_j - E\left(e^{-ivX_j} Y_j \right) \right\} \right| dv. \tag{2.40}$$

The expression in (2.40) is non-negative, so it suffices to show that its expectation is $O(N^{-\frac{1}{2}} h^{-l-1})$ or

$$\int \left| v^l \phi_k(hv) \right| \sup_x \left| e^{ivx} \right| E \left| \frac{1}{N} \sum_j \left\{ e^{-ivX_j} Y_j - E\left(e^{-ivX_j} Y_j \right) \right\} \right| dv = O(N^{-\frac{1}{2}} h^{-l-1}). \tag{2.41}$$

But $\sup_x \left| e^{ivx} \right| = 1$ and

$$\int \left| v^l \phi_k(hv) \right| dv = \frac{1}{h^{l+1}} \int \left| u^l \phi_k(u) \right| du = O(h^{-l-1}),$$

by the assumption that $k \in \mathcal{K}_2$, so we only need to show that the expectation in (2.41) is $O(N^{-\frac{1}{2}})$ uniformly in v. Thus, by the inequality of Cauchy-Schwarz

$$\begin{aligned} & \sup_v E \left| \frac{1}{N} \sum_j \left\{ e^{-ivX_j} Y_j - E\left(e^{-ivX_j} Y_j \right) \right\} \right| \\ \leq & \sup_v \sqrt{E \left| \frac{1}{N} \sum_j \left\{ e^{-ivX_j} Y_j - E\left(e^{-ivX_j} Y_j \right) \right\} \right|^2} \\ = & \; O(N^{-\frac{1}{2}}). \end{aligned}$$

Q.E.D.

Lemma 2.3 *We have:*

$$\sup_{x\in\Xi}\left|\hat{r}^{(l)}(x)-r^{(l)}(x)\right| = O_p\left(N^{-\frac{1}{2}}h^{-l-1}+h^{2-l}\right).$$

Proof:

Is a trivial combination of Lemmas 2.1 and 2.2.

Q.E.D.

Lemma 2.4 *We have*

$$\sup_{x\in\Xi} V\hat{r}^{(l)}(x) = O\left(N^{-1}h^{-2l-1}\right),\ \sup_{x\in\Xi} V\hat{f}^{(l)}(x) = O\left(N^{-1}h^{-2l-1}\right).$$

Proof:

(We will only show the first result). We have

$$\begin{aligned} V\hat{r}^{(l)}(x) &= E[\hat{r}^{(l)}(x)]^2 - E^2\hat{r}^{(l)}(x) \\ &= \frac{1}{N^2h^{2l+2}}\sum_i\left\{E\left[k_h^{(l)}(x-X_i)\left(m(X_i)+U_i\right)\right]^2\right. \\ &- \left.E^2\left[k_h^{(l)}(x-X_i)\left(m(X_i)+U_i\right)\right]\right\}. \end{aligned} \tag{2.42}$$

The last expectation on the right hand side in the above equation is by Lemma 2.1, $O(h^2)$.

Because $E[U_1|X_1]=0$, (2.42) is

$$\frac{1}{Nh^{2l+2}}\left\{E\left[\left(k_h^{(l)}(x-X_1)\right)^2 m^2(X_1)\right]+E\left[\left(k_h^{(l)}(x-X_1)\right)^2 E[U_1^2|X_1]\right]\right\}+O(N^{-1}h^{-2l}).$$

Expanding the first expectation in the above expression we obtain (using the conditions on q and k)

$$\int[k_h^{(l)}(x-v)]^2m^2(v)f(v)dv = h\int[k^{(l)}(u)]^2q(x-hu)du = O(h).$$

Q.E.D.

Lemma 2.5 *We have at any point $x\in\Xi$*

$$\hat{r}^{(l)}(x)-r^{(l)}(x) = O_p\left(N^{-\frac{1}{2}}h^{-l-\frac{1}{2}}+h^{2-l}\right)$$

and if r'' is first order Lipschitz-continuous, $\hat{r}^{(2)}(x) - r^{(2)}(x) = O_p(N^{-\frac{1}{2}}h^{-\frac{5}{2}} + h)$.

Proof:

Follows immediately from Lemmas 2.1 (and the remark immediately after it) and 2.4.

Q.E.D.

Of course Lemma 2.5 holds also for $\hat{f}$ with respect to f.

Lemma 2.6 *Suppose we have for all x that $\hat{r}_\omega(x) = \hat{r}^{(\omega_1)}(x)\hat{r}^{(\omega_2)}(x)$, where $\omega_1, \omega_2 \leq 2$. Let $\tilde{\omega} = \omega_1 + \omega_2 + 2$. Let $r \in \mathcal{G}_2$ and let the kernel be $k \in \mathcal{K}_2$. Then*

$$E\hat{r}_\omega(x) - E\hat{r}^{(\omega_1)}(x)E\hat{r}^{(\omega_2)}(x) = O(N^{-1}h^{1-\tilde{\omega}}),$$

for all x.

Proof:

$$\begin{aligned}
& E\left[\hat{r}^{(\omega_1)}(x)\hat{r}^{(\omega_2)}(x)\right] - E\hat{r}^{(\omega_1)}(x)E\hat{r}^{(\omega_2)}(x) \\
&= \frac{1}{N^2h^{\tilde{\omega}}}\sum_i\sum_j\left\{E\left[k_h^{(\omega_1)}(x-X_i)Y_i k_h^{(\omega_2)}(x-X_j)Y_j\right] - E\left[k_h^{(\omega_1)}(x-X_i)Y_i\right]E\left[k_h^{(\omega_2)}(x-X_j)Y_j\right]\right. \\
&= \frac{1}{N^2h^{\tilde{\omega}}}\sum_i \mathrm{Cov}\left(k_h^{(\omega_1)}(x-X_i)Y_i, k_h^{(\omega_2)}(x-X_i)Y_i\right) \\
&= O(N^{-1}h^{1-\tilde{\omega}}).
\end{aligned}$$

Q.E.D.

Part II

Serial Independence Testing

Chapter 3

Serial Independence Testing

3.1 Principles

This chapter serves as an introduction to Chapters 4 and 5, in which we extend one serial independence test, and put forward another.

In this chapter, depending on the context, we shall either be discussing tests for independence or for *serial* independence. We shall give a detailed explanation of these concepts in Subsection 3.1.1. In the first case, we assume to have observed an i.i.d. sequence $\{(X_{1t}, X_{2t})\}$, and we wish to test whether X_{11} and X_{21} are dependent upon one another. Thus, we wish to test

$$\begin{aligned} H_0 &: X_{11} \text{ is independent of } X_{21}, \\ &\text{versus} \\ H_1 &: X_{11} \text{ is not independent of } X_{21}. \end{aligned} \tag{3.1}$$

In the case of serial independence testing against serial dependence of order $J-1$, we test for a stationary series $\{X_t\}$, i.e. a series for which the distribution of $(X_{t+s_1}, \ldots, X_{t+s_J})$ for any

$s_1 < \cdots < S_J$; $J > 0$ does not depend on t, whether

$$H_0 \ : \ \{X_t\} \text{ are i.i.d.,}$$

versus

$$H_1 \ : \ X_1, \ldots, X_J \text{ are not all independent of one another.} \tag{3.2}$$

There are two considerations in selecting an independence test, or any other test for that matter. These are consistency and efficiency.

Definition 3.1 *A test is called* consistent *against a certain alternative if the probability that the test rejects the null if the alternative is correct tends to one, when the number of observations tends to* ∞.

A test that is not consistent against a certain alternative, may still have power greater than the significance level against that alternative, but the power will not tend to one, asymptotically.

The *efficiency* of a consistent test against a certain alternative is related to how fast the power tends to one, as the sample size increases, if the alternative hypothesis holds. The faster power increases as sample size increases, the more efficient the test is.

Definition 3.2 *We call an independence test* parametric, *if under the alternative, the relationship between the two variables whose independence is to be tested can be expressed in a finite number of parameters; it is called nonparametric if such a relationship would require an infinite number of parameters.*

Parametric correlation tests, for instance, are related to linear alternatives, whilst a (first order) Autoregressive Conditional Heteroskedasticity (ARCH) alternative takes the form $X_t = \varepsilon_t \sqrt{1 + \theta X_{t-1}^2}$, with $0 < \theta < 1$ and $\{\varepsilon_t\}$ white noise, the latter example being relevant only in a time series context. Parametric tests will generally be consistent against a much wider range

of alternatives than the parametrised alternative mentioned above, but they will generally not be consistent against quite as wide a class of alternatives as nonparametric tests. However, nonparametric tests are generally not as efficient as parametric tests with respect to alternatives against which the latter type is consistent.

Nonparametric tests generally require some additional regularity conditions, mostly related to the dependence structure under the alternative hypothesis. Mixing conditions are particularly relevant in this context; they are discussed in Section 3.1.2.

In Section 3.2, we discuss some parametric tests, whilst nonparametric tests are reviewed in Section 3.3. Section 3.4 discusses some specification tests and the relevance of nuisance parameters. Finally, Section 3.5 discusses the choice of test in a particular situation.

3.1.1 Independence versus Uncorrelatedness

In Chapters 4 and 5, independence and uncorrelatedness play an important role. We therefore discuss them in the current subsection.

Definition 3.3 *The random variables X_1 and X_2 are called* independent *if for all (Borel-) measurable sets A and B, $P[X_1 \in A, X_2 \in B] = P[X_1 \in A]P[X_2 \in B]$.*

A generally more practical yet equivalent definition is that the joint distribution function F_{12} of (X_1, X_2) is everywhere equal to the product of F_1 and F_2, their marginal distribution functions. We relax this restriction to F_{12} being *almost* everywhere equal to $F_1 \times F_2$. The difference is obviously irrelevant in practice. For continuous distributions, independence also implies that the joint density f_{12} of (X_1, X_2) is equal to the product of the marginal densities of X_1 and X_2, f_1 and f_2, for almost all values of the argument.

Definition 3.4 *Let $EX_1^2 + EX_2^2 < \infty$. The random variables X_1 and X_2 are called* uncorrelated *when $E[(X_1 - EX_1)(X_2 - EX_2)] = 0$.*

If X_1 and X_2 have finite second moments, independence implies uncorrelatedness, but not vice versa. If second moments do not exist, uncorrelatedness loses its meaning. This is particularly relevant in finance, as many financial time series have been found to have fat-tailed distributions.

An example in which X_1 and X_2 are uncorrelated but not independent, is the case where $X_2 = \varepsilon\sqrt{1+\theta X_1^2}$, for some $\theta > 0$, with ε some arbitrarily distributed random variable with zero mean and finite variance, that is independent of X_1. Indeed, $E[X_2\{X_1 - EX_1\}] = E[\varepsilon]E[\{X_1 - EX_1\}\sqrt{1+\theta X_1^2}] = 0$. However, if X_1 and X_2 are *jointly* Gaussian, their uncorrelatedness does imply their independence.

In time series analysis, we are often interested in serial independence and serial uncorrelatedness.

Definition 3.5 *A stationary time series* $\{X_t\}$ *is called* serially independent of order $J-1$, *if* $X_1, \ldots, X_J$ *are mutually independent.*

Definition 3.6 *Let* $\rho_j = Cov[X_1, X_{1+j}]/VX_1$, *for all* j, *exist for all* j. *A covariance stationary time series* $\{X_t\}$ *is called* serially uncorrelated of order $J-1$, *if* $\rho_j = 0$, *for all* $j = 1, \ldots, J-1$.

Again, if second moments exist, serial independence of order $J-1$ implies serial uncorrelatedness of order $J-1$, but not vice versa. Obviously, serial independence of order $J-1$ does not imply serial independence of order J, although the converse is true. An i.i.d. series is thus serially independent of infinite order.

We shall call random variables dependent (correlated) when they are not independent (uncorrelated). This may sometimes be a bit awkward as $X_t = \theta X_{t-2} + \varepsilon_t$, with $\{\varepsilon_t\}$ white noise, is not only serially dependent of order two, but also of any order greater than two.

3.1.2 Mixing Conditions

The conditions described in the current subsection relate to the way dependence between two elements in a stationary time series decreases as the elements are farther apart, time-wise. We begin with the discussion of some standard mixing conditions.

Definition 3.7 *A stationary series* $\{X_t\}$ *is called* strong mixing, *if a sequence of 'mixing' numbers* $\alpha(t)$, *with* $\alpha(t) \to 0$, *as* $t \to \infty$, *exists, such that for all* $t, s > 0$

$$\sup_{A \in \mathcal{M}_0^t, B \in \mathcal{M}_{t+s}^\infty} |P[AB] - P[A]P[B]| \leq \alpha(s), \tag{3.3}$$

where $\mathcal{M}_t^s$ *is the* σ*-algebra generated by* $\{X_t\}$ *in periods* t *to* s.

As two events A and B are independent when $P[AB] = P[A]P[B]$, $|P[AB] - P[A]P[B]|$ may be viewed as a measure of their dependence. If s in (3.3) is large, A and B are events related to (combinations of) elements that are far apart in time. The greater s, the less dependent events A and B are allowed to be, and when s tends to ∞, dependence should disappear altogether. Strong mixing is originally due to Rosenblatt (1956); an extensive discussion of both strong mixing and uniform mixing, which is introduced further below, can be found in Ibragimov and Linnik (1971). Rosenblatt (1956) showed that strong mixing series allow a central limit theorem. Davydov (1968) obtained the following very useful result for strong mixing processes, which can in a different guise also be found in Ibragimov and Linnik (1971), Lemma 1.

Lemma 3.1 (Davydov's Inequality) *Let* $\{X_t\}$ *be strong mixing with mixing numbers* $\{\alpha(t)\}$, *and let* Y_1 *and* Y_2 *be measurable with respect to* $\mathcal{M}_0^t$ *and* $\mathcal{M}_{t+s}^\infty$, *respectively. Assume that a* $p > 1$ *and a* $0 < C < \infty$ *exist such that* $E|Y_1|^p < \infty$, *and* $|Y_2| < C$ *a.s.. Then*

$$|E[Y_1Y_2] - EY_1EY_2| \leq 6E^{\frac{1}{p}}|Y_1|^p \alpha^{\frac{p-1}{p}}(s). \tag{3.4}$$

Proof:

See Davydov (1968).

There are however relevant cases for which the strong mixing condition does not hold. Borges (1991) (page 7) showed that a series $\{X_t\}$ defined by $X_t = \varepsilon_t + \sum_{s=1}^{\infty} s^{-\theta}\varepsilon_{t-s}$, for $\theta \in (\frac{1}{2}, 1)$, with $\{\varepsilon_t\}$ white noise, is not strong mixing. Another example, in this case for a discrete autoregressive process of finite order was given by Andrews (1984). Another example of a non-strong mixing process can be found in Rosenblatt (1961). It has long been known that stationary and invertible Gaussian Autoregressive Moving Average (ARMA) models are strong mixing with exponentially decaying mixing numbers $\{\alpha(t)\}$ [cf. e.g. Ibragimov and Rozanov (1978)].

Another condition that is frequently applied is absolute regularity.

Definition 3.8 *A stationary series* $\{X_t\}$ *is called* absolutely regular *with mixing numbers* $\{\beta(t)\}$, *if*

$$\sup_t E\left[\sup_{B \in \mathcal{M}_{t+s}^{\infty}} |P[B|\mathcal{M}_0^t] - P[B]|\right] \leq \beta(s),$$

where $\beta(s) \to 0$, *as* $s \to \infty$.

Absolute regularity is apparently due to Kolmogorov. Volkonskii and Rozanov (1961) were the first to study the properties of processes satisfying the conditions of Definition 3.8, however. Absolute regularity is somewhat stronger than strong mixing. It is frequently used to obtain asymptotic results for U- or V-statistics (see also Subsection 3.1.3). As absolute regularity is stronger than strong mixing, the same counter examples apply. Pham and Tran (1985) discuss conditions under which Gaussian autoregressive moving average (ARMA) type processes of finite order are absolutely regular.

Yoshihara (1976), Lemma 1, [cf. also Denker and Keller (1983), Lemma 6] proved a very important result for absolutely regular processes, which we reproduce in an adapted and simplified form below.

Lemma 3.2 *For any absolutely regular series $\{X_t\}$ with mixing numbers $\beta(t)$, and any combination of function g and $\delta > 0$, for which all expectations below exist,*

$$|Eg(X_t, X_s) - E_I g(X_t, X_s)|$$
$$\leq 4\left[\max\left\{E|g(X_t, X_s)|^{1+\delta}, E_I|g(X_t, X_s)|^{1+\delta}\right\}\right]^{\frac{1}{1+\delta}} \beta^{\frac{\delta}{1+\delta}}(|t-s|),$$

where E_I denotes the expectation under independence of X_t and X_s.

Proof:

See Yoshihara (1976), or Denker and Keller (1983).

Lemma 3.2 establishes a relationship between the rate at which $E[g(X_t, X_s)]$ for absolutely regular $\{X_t\}$ converges to the same expectation for i.i.d. $\{X_t\}$, and the rate at which the mixing numbers tend to zero. Indeed, if we set t to zero, Lemma 3.2 implies that a sequence $\{C(s)\}$ exists, such that $|Eg(X_0, X_s) - E_I g(X_0, X_s)| \leq C(s) \propto \beta^{\frac{\delta}{1+\delta}}(s)$, for $s > 0$, where $\propto$ means 'is proportional to'. If g were bounded, one could let $\delta \to \infty$, thus implying that $|Eg(X_0, X_s) - E_I g(X_0, X_s)|$ is bounded by a quantity which is proportional to the s-th mixing number.

A mixing condition much stronger than absolute regularity is *uniform mixing*. Indeed, Ibragimov and Linnik (1971), Theorem 17.3.2, show that if a Gaussian series is uniform mixing, dependence can only be of finite order, in the sense that some s^* exists, such that X_t does not depend on X_{t-s}, for any $s > s^*$, and any t (this additional explanation is needed in view of our definition of serial dependence of certain order at the end of Section 3.1.1).

Definition 3.9 *A stationary series $\{X_t\}$ is uniform mixing if and only if a sequence of mixing numbers $\{\tilde{\phi}(t)\}$ exists, such that*

$$\sup_{t, A \in \mathcal{M}_0^t, B \in \mathcal{M}_{t+s}} |P[B|A] - P[B]| \leq \tilde{\phi}(s),$$

for all s.

Another condition we shall use is one we have coined "trigonometric mixing".

Definition 3.10 *A stationary series* $\{X_t\}$ *is called* trigonometric mixing *with mixing numbers* $\alpha(t)$, *if*

$$\sup_{u,v}\{|Cov\{\cos(uX_1), \cos(vX_{1+s})\}| + |Cov\{\sin(uX_1), \sin(vX_{1+s})\}|\} \leq \alpha(s), \tag{3.5}$$

with $\alpha(s) \to 0$, *as* $s \to \infty$.

Trigonometric mixing is closely linked to strong mixing, and we have therefore used the same symbol for its mixing number sequence. Strong mixing implies trigonometric mixing. This is an immediate result of Lemma 3.1.

To show that strong mixing implies trigonometric mixing it suffices to show that for any strong mixing series, (3.5) holds for a certain sequence of (trigonometric) mixing numbers. Applying Lemma 3.1 to the sines and cosines in (3.5) is adequate for this purpose.

We have no evidence indicating how often a series is trigonometric mixing, but not strong mixing. We are not even sure any such series exist. Moreover, many authors have used strong mixing, where a weaker assumption would have sufficed. As trigonometric mixing is at least as weak as strong mixing, we shall nevertheless assume trigonometric mixing rather than strong mixing in Chapters 4 and 5.

Another mixing condition was used by Robinson (1991b). Robinson left his mixing condition nameless. As it is defined in terms of characteristic functions we shall call it CF-mixing.

Definition 3.11 *A stationary series* $\{X_t\}$ *is called* CF-mixing, *with mixing numbers* $v(t)$, *if for all* $t > 0$

$$\int |Ee^{iu(X_1 - X_{1+t})} - |Ee^{iuX_1}|^2|du \leq v(t), \tag{3.6}$$

where $v(t) \to 0$, *as* $t \to \infty$.

Definition 3.11 is likely to be weaker than the other mixing conditions discussed in this subsection. Definition 3.11 restricts an average rather than the supremum of the absolute difference of the two characteristic functions, as Definition 3.10 does.

3.1.3 U- and V-statistics

U-statistics are originally due to Hoeffding (1948), and are very popular in the nonparametric estimation and testing literature.

Definition 3.12 *A* U-statistic *is a statistic of the form*

$$U = \begin{pmatrix} N \\ J \end{pmatrix}^{-1} \sum_{t_1} \sum_{t_2 > t_1} \cdots \sum_{t_J > t_{J-1}} g(X_{t_1}, \ldots, X_{t_J}). \tag{3.7}$$

Here, g is called the U-statistic kernel and is often assumed to be symmetric in its arguments. Let $g_1(x) = E_I[g(x, X_2, \ldots, X_J)]$. Then g is called degenerate *(for the distribution used in the afore expectation and the hereafter used variance) if $Vg_1(X_1) = 0$. A* V-statistic *is a statistic of the form*

$$V = N^{-J} \sum_{t_1} \cdots \sum_{t_J} g(X_{t_1}, \ldots, X_{t_J}).$$

An example of a degenerate kernel is $g(x_1, x_2) = x_1 x_2$, which is degenerate for all distributions with $EX_1 = 0$. The restriction that g be symmetric is not important, as if g is not symmetric, we may replace $g(X_{t_1}, \ldots, X_{t_J})$ in (3.7) by $\frac{1}{J!}\{g(X_{t_1}, \ldots, X_{t_J}) + g(X_{t_1}, \ldots, X_{t_{J-2}}, X_{t_J}, X_{t_{J-1}}) + \ldots + g(X_{t_J}, \ldots, X_1)\}$.

Of course, U- and V-statistics are very similar, and are usually treated simultaneously. The following theorem is due to Hoeffding (1948), and can also be found in Serfling (1980), Theorem A on page 192.

Theorem 3.1 *If* $\{X_t\}$ *are i.i.d.,* $Eg^2(X_1,\ldots,X_J) < \infty$, *and* $Vg_1(X_1) > 0$, *then*

$$\frac{\sqrt{N}}{J}\{U - Eg(X_1,\ldots,X_J)\} \xrightarrow{\mathcal{L}} N(0, Vg_1(X_1)).$$

Proof:

See Hoeffding (1948).

A simple example is the U-statistic with $g(x,y) = (x-y)^2$. Because $g_1(x) = E[g(x,X_1)] = x^2 - 2xEX_1 + EX_1^2$, Theorem 3.1 implies that $\sqrt{N}(U - E[X_1 - X_2]^2) \xrightarrow{\mathcal{L}} N(0, 4\mu_4 - 16\mu_3\mu_1 + 32\mu_2\mu_1^2 - 4\mu_2^2 - 16\mu_1^4)$, where $\mu_j = EX_1^j$, provided that $EX_1^4 < \infty$.

A similar result is available for absolutely regular processes (Definition 3.8). The following is due to Denker and Keller (1983), Theorem 1 (c).

Theorem 3.2 *Let* g *be a non-degenerate kernel. Let* $\{X_t\}$ *be absolutely regular with mixing numbers* $\beta(t)$ *such that a* $\delta > 1$ *exists for which* $\sum_t \beta^{\frac{\delta}{1+\delta}}(t) < \infty$, *and* $\sup_{t_1<t_2<\cdots<t_J} E|g(X_{t_1},\ldots,X_{t_J})|^{1+\delta} < \infty$. *Let* $\sigma^2 = Vg_1(X_1) + 2\sum_{t>1} Cov\{X_1, X_t\} \neq 0$. *Then*

$$\frac{\sqrt{N}}{J}\{U - E_I[g(X_1,\ldots,X_J)\} \xrightarrow{\mathcal{L}} N(0,\sigma^2).$$

Proof:

See Denker and Keller (1983).

3.2 Parametric Tests

As indicated in Subsection 3.1.1, we present some parametric independence tests. In Subsection 3.2.1, we discuss some parametric tests for uncorrelatedness and in Subsection 3.2.2 we examine the Lagrange Multiplier test, the Likelihood Ratio test, and the Wald test.

3.2.1 Tests of Uncorrelatedness

The first correlation test dates back to the end of the last century. The basis for the first serial correlation test, or rather serial uncorrelatedness test, we know of, was laid by von Neuman (1941). His idea was based on the mean square successive difference of von Neuman et al. (1941), which is given by $\frac{1}{N-1}\sum_t(X_{t+1}-X_t)^2$. Von Neuman et al. were particularly interested in Gaussian series, for which they wished to determine the presence of trends. They noted that half the mean square difference estimates the population variance, if the elements in the series are identically distributed and uncorrelated, which for Gaussian series is equivalent to them being i.i.d., as we noted in Subsection 3.1.1. However, if the observations are identically distributed but have non-zero first order autocorrelations, then half the mean square successive difference does not estimate the variance, and the ratio of mean square successive difference to variance may thus well serve as a basis for testing for uncorrelatedness.

Von Neuman (1941) obtained expressions for the distribution of the afore-mentioned ratio for Gaussian series. We are interested in testing for uncorrelatedness (H_0) against first order serial correlation (H_1). We assume throughout that $\{X_t\}$ is stationary and ergodic and that X_1 has an arbitrary distribution with $EX_1 = 0$ and $0 < VX_1 < \infty$. Von Neuman's ratio is under these conditions given by

$$\hat{\tau}_{VN} = \frac{\sum_t(X_{t+1}-X_t)^2}{\sum_t X_t^2}.$$

Under the above conditions, $\hat{\tau}_{VN} \xrightarrow{P} 2-2\rho_1$, where $\rho_1 = E[X_1X_2]/EX_1^2$, which follows immediately with the ergodic theorem.

Strictly speaking, the stationarity condition is stronger than required; we could obtain a result for non-stationary martingale differences, which are defined below, under certain additional conditions. Two of these conditions are that $\{X_tX_{t+1}\}$ is a martingale difference sequence, and that $V[X_1X_2] = V^2X_1$, both under the null.

Definition 3.13 *$\{X_t\}$ is a sequence of martingale differences if* $E[X_t|\mathcal{M}_0^{t-1}] = 0$.

$\{X_tX_{t+1}\}$ are martingale differences if $\{X_t\}$ are martingale differences with $P[X_t = 0] = 0$, for all t, as $E[X_tX_{t+1}|X_{t-1}X_t, \ldots, X_1X_2] =$

$E[X_tE\{X_{t+1}|X_t, X_{t-1}X_t, \ldots, X_1X_2\}|X_{t-1}X_t, \ldots, X_1X_2] = 0$, because the inner expectation is equal to $E[X_{t+1}|X_t, \ldots, X_1]$, as $P[X_t = 0] = 0$, for all t.

McLeish's (1974) central limit theorem for martingale difference sequences implies that $N^{-\frac{1}{2}}\sum_t X_tX_{t+1} \xrightarrow{\mathcal{L}} N(0, V[X_1X_2])$. As we assumed that $V^2X_1 = V[X_1X_2]$, in the paragraph preceding Definition 3.13, $N^{-\frac{1}{2}}\sum_t X_tX_{t+1} \xrightarrow{\mathcal{L}} N(0, V^2X_1)$, and hence $\sqrt{N}(\hat{\tau}_{VN} - 2) \xrightarrow{\mathcal{L}} N(0, 1)$.

The condition $V[X_1X_2] = V^2X_1$ is necessary to get asymptotically valid test results. This condition does not hold for all $\{X_t\}$ that are serially uncorrelated, however. Indeed, if $\{X_t\}$ is Autoregressive Conditionally Heteroskedastic (ARCH), i.e. $X_{t+1} = \varepsilon_{t+1}\sqrt{1 + \theta X_t^2}$, with $\{\varepsilon_t\}$ white noise and $0 \leq \theta < 1$, then $E[X_1^2X_2^2] = E\varepsilon_1^2E[(1+\theta X_1^2)X_1^2] \neq V^2X_1$, although $E[X_1X_2] = E\varepsilon_1E[\sqrt{1+\theta X_1^2}X_1] = 0$.

Durbin and Watson (1950, 1951, 1971) applied Von Neuman's statistic to the residuals of linear models, as we shall see in Section 3.4. Box and Pierce (1970) and Ljung and Box (1978) were also primarily interested in the problem studied by Durbin and Watson, and their efforts are therefore also discussed there.

It should be noted that tests for uncorrelatedness are particularly powerful against linear alternatives. Nevertheless, there are many other alternatives against which they are also consistent, but the class is more limited than that of nonparametric tests. It is important to notice, that in order for uncorrelatedness tests to make sense, VX_1 must exist. Particularly in financial data, one often encounters leptokurtic distributions. In settings like these, one should therefore be very careful with applying an uncorrelatedness test.

3.2.2 The Lagrange Multiplier Test, the Likelihood Ratio Test, and the Wald Test

We now turn to test for serial independence against a certain parametric alternative.

The three tests in the title of this subsection are all defined in the context of maximum likelihood estimation, although the Lagrange multiplier principle can easily be extended to models in which parameters are estimated by other extremum estimates, and the Wald test can be generalised to cover a still wider variety of situations. However, when parameters are estimated by maximum likelihood estimation, all three tests are efficient, provided that the distribution specified is correct.

3.2.2.1 Wald Test

If the restriction to be tested is $a(\theta_0) = 0$, with θ_0 the true parameter vector, then the Wald (1943) test examines $a(\hat{\theta})$, where $\hat{\theta}$ is the unconstrained maximum likelihood estimate of θ_0. When $\hat{\theta}$ is not a maximum likelihood estimate, the test is not necessarily efficient, but, as we show below, provided that $N^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{V})$, with $\mathcal{V}$ some positive definite and finite variance matrix, and that a is totally differentiable at θ_0, we still have $N^{\frac{1}{2}}a(\hat{\theta}) \xrightarrow{\mathcal{L}} N(0, \mathcal{V}^*)$, for a positive definite and finite matrix $\mathcal{V}^*$.

By Slutsky's theorem, $a(\hat{\theta}) \xrightarrow{P} a(\theta_0)$, whenever a is continuous. Indeed, if a is totally differentiable, and $N^{\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \mathcal{V})$, then $N^{\frac{1}{2}}\{a(\hat{\theta}) - a(\theta_0)\} \xrightarrow{\mathcal{L}} N\left(0, \frac{\partial a}{\partial \theta^T}|_{\theta_0} \mathcal{V} \frac{\partial a}{\partial \theta}|_{\theta_0}\right)$, such that, under the null,

$$\hat{\tau}_W = N a^T(\hat{\theta}) \left(\frac{\partial a}{\partial \theta^T}|_{\theta_0} \mathcal{V} \frac{\partial a}{\partial \theta}|_{\theta_0}\right)^{-1} a(\hat{\theta}) \xrightarrow{\mathcal{L}} \chi^2_s, \tag{3.8}$$

where s is the number of restrictions in a. A simple example would be to test whether $\theta_0 = 0$ in $X_{t+1} = \theta_0 X_t + \varepsilon_{t+1}$, assuming $-1 < \theta_0 < 1$. The maximum likelihood estimate for θ_0 is under Gaussianity asymptotically equivalent to the least squares estimate, which is $\hat{\theta} =$

$\sum_t X_t X_{t+1} / \sum_t X_t^2$. Under the null, $N^{-\frac{1}{2}} \sum_t X_t X_{t+1} \xrightarrow{\mathcal{L}} N(0, E^2 X_1^2)$, and the AR specification could then simply be tested by $\sum_t X_t X_{t+1} / \sum_t X_t^2$, which is the basis for the Box Pierce (1970) test, discussed in Section 3.4.

Although the Wald test is very simple, it is often criticised for the degree of arbitrariness, resulting from the specification of a. Indeed, testing $\theta_0 = 0$, is equivalent to testing $\theta_0^2 = 0$, but the test statistic will attain different values [cf. e.g. Lafontaine and White (1986)].

3.2.2.2 Likelihood Ratio Test

The likelihood ratio test compares the values of the loglikelihood under the null and that of the loglikelihood when no restrictions are applied. If $L(\theta)$ denotes the likelihood at θ, then the likelihood ratio test statistic is defined as

$$\hat{\tau}_{LR} = 2 \log \frac{L(\hat{\theta})}{L(\tilde{\theta})},$$

where $\hat{\theta}$ is the unrestricted maximum likelihood estimate of the true parameter vector θ_0, and $\tilde{\theta}$ is that under the restrictions applying under the null hypothesis. $\hat{\tau}_{LR}$ is always greater than or equal to zero, because $L(\hat{\theta}) \geq L(\tilde{\theta})$. Indeed, under the null hypothesis, $\hat{\tau}_{LR} \xrightarrow{\mathcal{L}} \chi_1^2$, provided that certain additional regularity conditions are satisfied [cf. Godfrey (1988), Section 1.4].

A disadvantage of the likelihood ratio test is that one needs to know the distribution of X_1 to be able to specify the likelihood equation. The Lagrange multiplier test may, in principle, also be applied in settings other than those involving maximum likelihood estimation, and it does not require the actual maximum likelihood estimates to be computed.

3.2.2.3 Lagrange Multiplier Test

The Lagrange multiplier test is conceptually more difficult, but generally easier to implement, than the likelihood ratio test. It is applied to models, in which the parameters are estimated

by means of maximum likelihood, but one could easily extend this to other settings. Indeed, any model with parameter estimates that can be defined as optimising an objective function is in principle suitable, although certain regularity conditions are required. We shall however examine the test in the original format, where the objective function is the loglikelihood.

We shall not list all the regularity conditions required for the Lagrange multiplier test to apply, but refer to Godfrey (1988), page 6–7, for the interested reader. One very important condition, however, is that the loglikelihood be strictly concave (convex for minimisation problems) near θ_0.

Let f denote the density of X_1. Then the loglikelihood function, for a time series model, is given by $L_N(\theta) = \sum_{t=1}^{N} \log f(X_t|\mathcal{M}_0^{t-1};\theta)$, where, as before, $\mathcal{M}_0^{t-1}$ denotes the sigma algebra in periods 0 through $t-1$. As before, we wish to test the restriction $a(\theta_0) = 0$. Godfrey's (1988) regularity conditions only apply to linear a, but nonlinear a are in principle also possible.

If the null hypothesis holds, imposing the restriction $a(\hat{\theta}) = 0$, will asymptotically have no effect, provided that $\hat{\theta}$ is consistent for θ_0, which is ensured by the afore-mentioned regularity conditions. Hence, under the null, we could also estimate θ_0 by

$$\hat{\theta} = \text{argmax}_{\theta,\lambda}\{L_N(\theta) - \lambda a(\theta)\}.$$

The first order conditions are

$$\begin{aligned} \frac{\partial L_N}{\partial \theta}|_{\hat{\theta}} - \frac{\partial a}{\partial \theta^T}|_{\hat{\theta}}\hat{\lambda} &= 0, \\ a(\hat{\theta}) &= 0. \end{aligned}$$

In a linear regression context one usually lets $a(\theta_0) = R\theta_0 - r$, for some matrix of constants R, and some vector of constants r.

If $a(\theta) = \theta - \theta^*$, with θ and θ^* scalars, then $\hat{\lambda}$ is just the first derivative of the loglikelihood at θ^*. Because in this case $\theta_0 = \theta^*$ under the null, $\hat{\lambda}$ should then converge to zero, in probability.

If the test is defined in terms of the value of $\hat{\lambda}$, one has the proper Lagrange multiplier form suggested by Aitchinson and Silvey (1958,1960). Closely related is the *score test*, which is often also referred to as the Lagrange multiplier test. Originally due to Rao (1948), it examines the behaviour of $\hat{\lambda}^T \frac{\partial a}{\partial \theta}|_{\hat{\theta}}$. The score test is equivalent to the Lagrange multiplier test, but it is usually easier to obtain, particularly if a contains more than one restriction.

The first derivative of the loglikelihood at θ is

$$L'_N(\theta) = \sum_t \frac{\partial f}{\partial \theta}|_{(X_t|\mathcal{M}_0^{t-1};\theta)} f^{-1}(X_t|\mathcal{M}_0^{t-1};\theta).$$

Under the null, and under suitable regularity conditions [cf. Godfrey (1988), page 6–7, for those for linear a], $\frac{1}{N}L'_N(\hat{\theta})$, where $\hat{\theta}$ is again the restricted maximum likelihood estimate, converges in probability to 0, and $N^{-\frac{1}{2}}L'_N(\hat{\theta}) \xrightarrow{\mathcal{L}} N(0, -E[\lim_{N\to\infty} N^{-1}L''_N(\theta^*)])$. If we call the variance matrix of the limiting normal distribution $\mathcal{V}$,

$$\hat{\tau}_{LM} = \hat{\lambda}^T \frac{\partial a}{\partial \theta}|_{\hat{\theta}} \mathcal{V}^{-1} \frac{\partial a}{\partial \theta^T}|_{\hat{\theta}} \hat{\lambda} \xrightarrow{\mathcal{L}} \chi^2_s, \tag{3.9}$$

where $1 \leq s < \infty$ is the number of restrictions in a.

Many tests have been based on the Lagrange multiplier test, and many that were not, originally, can also be written in that form. Suppose for instance that $f(X_{t+1}|\mathcal{M}_0^t;\theta_0) \sim N(\theta_0 X_t, \sigma^2)$, $-1 < \theta_0 < 1$, which corresponds to a first order AR specification, $X_{t+1} = \theta_0 X_t + \varepsilon_{t+1}$ with $\{\varepsilon_t\}$ Gaussian white noise with $V\varepsilon_1 = \sigma^2(1-\theta_0)^2$. We want to test $\theta_0 = 0$, such that $a(\theta) = \theta$. Now, if $c = \log\sqrt{2\pi}$, then $L_N(\theta_0) = -Nc - N\log\sigma - \sum_t (X_{t+1} - \theta_0 X_t)^2/(2\sigma^2)$, such that $L'_N(0) = \frac{1}{\sigma^2}\sum_t X_t X_{t+1}$, and $L''_N(0) = -\frac{1}{\sigma^2}\sum_t X_t^2$. Under the null, $N^{-\frac{1}{2}}L'_N(0) \xrightarrow{\mathcal{L}} N(0,1)$, as expected, and indeed $N^{-1}\{L'_N(0)\}^2 \xrightarrow{\mathcal{L}} \chi^2_1$. σ^2 is not observed, but can be estimated by $\frac{1}{N}\sum_t X_t^2$, leading to

$$N\left(\frac{\sum_t X_t X_{t+1}}{\sum_t X_t^2}\right)^2 \xrightarrow{\mathcal{L}} \chi^2_1,$$

which is, as we have seen earlier, the basis for the Box Pierce (1970) test, discussed in Section 3.4.

Many other tests for parametric hypotheses have been based on the Lagrange multiplier principle, amongst others Engle's (1982) ARCH test. Indeed, most parametric tests are based on the Lagrange multiplier principle.

3.2.2.4 Concluding Remarks

It is a well-known fact [cf. e.g. Godfrey (1988), page 17] that the Wald test, the likelihood ratio test, and the Lagrange multiplier test are all asymptotically efficient, against the parametrised alternative. If all three are defined, always $\hat{\tau}_W \geq \hat{\tau}_{LR} \geq \hat{\tau}_{LM}$, where the inequality obviously disappears, asymptotically. None of these tests is consistent against as wide a class of alternatives as most nonparametric tests, but they generally do have power against other alternatives than the specified one. It should also be noted that all three are used in the context of maximum likelihood estimation, which requires knowledge about the distribution of X_1. Indeed, an incorrect specification of this distribution may lead to incorrect results.

3.3 Nonparametric Tests

As mentioned above, nonparametric tests neither require distributional assumptions, and are usually consistent against a very wide range of alternatives. Parametric tests, on the other hand, are commonly based upon a specified parametric alternative, and require distributional assumptions, at least to be efficient.

There are many ways to test for independence, nonparametrically. The first type of nonparametric independence tests were rank tests, which are discussed in Subsection 3.3.4. One may also compare the joint distribution function to the product of the marginal distribution

functions; such tests are discussed in Subsection 3.3.1. If the distributions in question are continuous, one may alternatively compare the joint density to the product of the marginal density, as the tests that are discussed in Subsection 3.3.2 do. The tests in Subsection 3.3.3 compare the characteristic function of the joint distribution to the product of the characteristic functions of the marginal distributions. Brock, Dechert, and Scheinkman's (1987) correlation dimension test, which is discussed in Subsection 3.3.5, is based upon the correlation dimension, which is also discussed in that subsection. Finally, there exist some nonparametric tests that have a null hypothesis under which each of the X_t's is a (possibly infinite) linear combination of i.i.d. white noise, whereas under the alternative the relationship may be nonlinear, also. One test of this type is discussed in Subsection 3.3.6.

3.3.1 Distribution Function Based Tests

Distribution function based tests are the oldest nonparametric independence tests. The first to propose such a test was Hoeffding (1948). He examined the quantity

$$\int \{F_{12}(x,y) - F_1(x)F_2(y)\}^2 dF_{12}(x,y), \tag{3.10}$$

which is zero, if $F_{12}(x,y) = F_1(x)F_2(y)$, almost everywhere, and is greater than zero, otherwise. (3.10) can be estimated by

$$\hat{\tau}_{HO} = \frac{1}{N}\sum_t \left[\frac{1}{N}\sum_s I(X_{1s} \leq X_{1t})\left\{I(X_{2s} \leq X_{2t}) - \frac{1}{N}\sum_u I(X_{2u} \leq X_{2t})\right\}\right]^2, \tag{3.11}$$

which is an asymmetric V-statistic. However, its kernel (when symmetrised) is degenerate, such that Theorem 3.1 can not be applied. The theory for degenerate kernels [cf. Serfling (1980), Theorem B, page 193] implies that $N\hat{\tau}_{HO}$ is asymptotically distributed as $\sum_t \sum_s \lambda_{ts} A_{ts}$, where the A_{ts}'s are mutually independent χ^2_1-random variables. The λ_{ts}'s are obtainable and are, for continuous distributions, not dependent on the distribution, and the asymptotic distribution

can therefore be tabulated. Hoeffding (1948) only obtained the characteristic function of the asymptotic distribution; Blum, Kiefer, and Rosenblatt (1961) suggested the above representation.

Blum, Kiefer, and Rosenblatt (1961) extended Hoeffding's idea, and also suggested (3.10) could be replaced by a measure of the form

$$\sup_{x,y} |F_{12}(x,y) - F_1(x)F_2(y)|, \tag{3.12}$$

which could be estimated by

$$\hat{\tau}_{BKR} = \sup_{x,y} \left| \frac{1}{N} \sum_t I(X_{1t} \leq x) \left\{ I(X_{2t} \leq y) - \frac{1}{N} \sum_s I(X_{2s} \leq y) \right\} \right|. \tag{3.13}$$

Using the Kolmogorov-Smirnov representation of (3.12) and (3.13) instead of the Cramér-von Mises one used in (3.10) and (3.11), would not alter the mathematical treatment very much. However, in practical terms, $\hat{\tau}_{HO}$ may perhaps be preferred over $\hat{\tau}_{BKR}$, because it is easier to compute.

Skaug and Tjøstheim (1992b) extended the work of Hoeffding (1948) and Blum, Kiefer and Rosenblatt (1961) to the framework of serial independence testing. They showed that $\int \{F_{12}(x,y) - F(x)F(y)\}^2 dF_{12}(x,y)$, the equivalent of (3.10) in the context of serial independence testing, can be estimated by

$$\hat{\tau}_{ST} = \frac{1}{N} \sum_t \left[\frac{1}{N} \sum_s I(X_s \leq X_t) \left\{ I(X_{s+1} \leq X_{t+1}) - \frac{1}{N} \sum_u I(X_{u+1} \leq X_{t+1}) \right\} \right],$$

and showed $N\hat{\tau}_{ST}$ to have an asymptotic distribution of the same type as that of Hoeffding (1948), and Blum, Kiefer and Rosenblatt (1961). They also obtained expressions for the λ_{ts}'s, and suggested a test for serial independence against serial dependence of order $J-1$, that uses

$$\hat{\tau}_{ST}^{(J)} = \sum_{j=1}^{J-1} \frac{1}{N} \sum_t \left[\frac{1}{N} \sum_s I(X_s \leq X_t) \left\{ I(X_{s+j} \leq X_{t+j}) - \frac{1}{N} \sum_u I(X_{u+j} \leq X_{t+j}) \right\} \right],$$

which they showed to have an asymptotic distribution of the afore type, albeit with the A_{ts}'s mutually independent χ^2_{J-1}.

Delgado (1993), in an effort independent of that of Skaug and Tjøstheim (1992b), obtained the same result as Skaug and Tjøstheim for the case of testing against serial dependence of order one, but chose an approach that is somewhat closer to the original one, for testing against higher order alternatives. He suggested to examine

$$\hat{\tau}_{DE}^{(J)} = \frac{1}{N}\sum_t \left[\frac{1}{N}\sum_{s_0} I(X_{s_0} \le X_t)\left\{\prod_{j=1}^{J-1} I(X_{s_0+j} \le X_{t+j}) - \sum_{s_1}\cdots\sum_{s_{J-1}}\prod_{j=1}^{J-1} I(X_{s_j} \le X_{t+j})\right\}\right].$$

He obtained an expression for the asymptotic distribution of $N\hat{\tau}_{DE}^{(J)}$, and also suggested a bootstrapping method to improve the rate of convergence of his statistic.

3.3.2 Density Function Based Tests

All tests in this subsection are based on distance measures, that take the value zero, when $f_{12}(x,y) = f_1(x)f_2(y)$, almost everywhere, and a positive value, if this is not the case. The four measures that have been used are the L_2-distance, the L_1-distance, the Kullback-Leibler (1961) [see also Kullback (1959)] information criterion, and the expected difference, which are (in the present setting)

$$\mathcal{I}_{L_2} = \int \{f_{12}(x,y) - f_1(x)f_2(y)\}^2 dxdy, \tag{3.14}$$

$$\mathcal{I}_{L_1} = \int |f_{12}(x,y) - f_1(x)f_2(y)| dxdy, \tag{3.15}$$

$$\mathcal{I}_{KL} = \int f_{12}(x,y)\{\log f_{12}(x,y) - \log f_1(x) - \log f_2(y)\} dxdy, \tag{3.16}$$

$$\mathcal{I}_{ST}^{*} = \int f_{12}(x,y)\{f_{12}(x,y) - f(x)f(y)\} dxdy. \tag{3.17}$$

Evidently, (3.14) to (3.17) could be used to test a variety of hypotheses other than (serial) independence. It is obvious that both $\mathcal{I}_{L_2}$ and $\mathcal{I}_{L_1}$ are zero under the null, and greater than zero, otherwise. Skaug and Tjöstheim (1992a) showed that this is not necessarily the case for

$\mathcal{I}^*_{ST}$. It is true, however, for $\mathcal{I}_{KL}$, which is not obvious. Kullback (1959) provided the following lemma [Kullback (1959), Theorem 3.1].

Lemma 3.3 *For any two bounded densities f and f^* that have the same support, $\int f(x)\log\{f(x)/f^*(x)\}dx \geq 0$, and there is equality if and only if $f(x) = f^*(x)$, almost everywhere.*

Proof:

Let $g(x) = f(x)/f^*(x)$, for all x, for which $f^*(x) > 0$. Then, by the mean value theorem,

$$\begin{aligned}\int f(x)\log\frac{f(x)}{f^*(x)}dx &= \int f^*(x)g(x)\log g(x)dx \\ &= \int f^*(x)\left[\{g(x)-1\} + \frac{\{g(x)-1\}^2}{2\{g(x);1\}}\right]dx \geq 0,\end{aligned}$$

as the first term in the last displayed equation is zero, and because $0 < \{g(x);1\} < \infty$, the second is non-negative, and zero only if $g(x) = 1$, in almost all x, which is equivalent to $f(x) = f^*(x)$, in almost all x.

Q.E.D.

Rosenblatt (1975) replaced the densities in (3.14) with kernel density estimates. Wahlen (1991) extends this idea in his Ph.D. thesis [published in reduced form as Rosenblatt and Wahlen (1992)], in that no longer the bandwidth necessarily needs to converge to zero at a rate of $N^{-\frac{1}{5}}$. They showed that

$$\hat{\tau}_{RW} = Nh^2\int\{\hat{f}_{12}(x,y) - \hat{f}_1(x)\hat{f}_2(y)\}^2dx,$$

where the $\hat{f}$'s are kernel density estimates, and h a bandwidth, behaves asymptotically as a $N(A_1 - hA_2, h^2\sigma^2)$-distributed random variable, where $A_1, A_2, \sigma^2 > 0$, or, differently stated, that $h^{-1}(\hat{\tau}_{RW} - A_1 + hA_2) \xrightarrow{\mathcal{L}} N(0,\sigma^2)$ [see Rosenblatt and Wahlen (1992), Theorem]. The mean $A_1 - hA_2$ is not directly estimable, and one may therefore need to use simulated critical values.

A problem associated with all nonparametric density estimation techniques is the need for the choice of a sample-size-dependent bandwidth parameter. The bandwidth usually has a strong impact on performance. See Subsection 1.3.1 for a short discussion of bandwidth selection procedures. Kernel estimates moreover require the choice of a kernel, and this issue is also discussed in Subsection 1.3.1. It is probably feasible to generalise the work of Rosenblatt and Wahlen to serial independence testing.

Chan and Tran (1992) estimated $\mathcal{I}_{L_1}$ in (3.15) using histogram density estimates [cf. e.g. Prakasa Rao (1983), page 93 ff.]. For densities limited to a bounded support, they showed that their test based on

$$\hat{\tau}_{CT} = \int |\hat{f}_{12}(x,y) - \hat{f}(x)\hat{f}(y)| dx dy,$$

where the $\hat{f}$'s are histogram based density estimates, is consistent against all departures from the null of serial independence against serial dependence of order one. For densities with infinite support, their statistic will reject any alternatives for which the integral in (3.15), taken over a bounded and practitioner-chosen interval, is not equal to zero. If the histogram density estimates are replaced by kernel density estimates, it is, under certain smoothness conditions, probably possible to prove that tests based on $\hat{\tau}_{CT}$ are generally consistent against all serial dependence of order alternatives, for densities with infinite support, also. They failed to obtain a convergence rate or indeed a limiting distribution for their statistic, relying on a resampling scheme to provide the critical values for their test.

The Kullback-Leibler (1961) information criterion has been used by various authors. When testing for serial independence against serial dependence of order one, one may use $\int \hat{f}_{12}(x,y)\{\log \hat{f}_{12}(x,y) - 2\log \hat{f}(x)\} dx dy$, or a statistic based on $\frac{1}{N}\sum_t \{\log \hat{f}_{12}(X_t, X_{t+1}) - 2\log \hat{f}(X_t)\}$. Robinson (1991a) examined the latter type, but found that if the $\hat{f}$'s were chosen to be kernel density estimates, its asymptotic distribution was intractable. He therefore chose

to introduce a weighting sequence $\{c_t\}$, with $c_t = 1 + \gamma\{1 - 2I(t \text{ even})\}$, for some $\gamma > 0$. His statistic,

$$\hat{\tau}_R = \frac{1}{\gamma\sqrt{2N\hat{V}\log f(X_1)}} \sum_{t \in S} c_t \{\log \hat{f}_{12}(X_t, X_{t+1}) - 2\log \hat{f}(X_t)\}, \tag{3.18}$$

where $\hat{V}\log f(X_1) = \frac{1}{N}\sum_t \log^2 \hat{f}(X_t) - \{\frac{1}{N}\sum_t \log \hat{f}(X_t)\}^2$, has an asymptotic $N(0,1)$ distribution, under the null hypothesis, provided that f and f_{12} have compact support. The set S in (3.18) serves to trim out observations for which $\hat{f}_{12}(X_t, X_{t+1})$ or $\hat{f}(X_t)$ is less than, or equal to, zero. The choice of γ is quite important: choosing γ large will lead to a good normal approximation with reduced power, whilst choosing γ small will result in the power being large but the normal approximation being poor. The reason is that

$$\tilde{\tau}_R = \frac{1}{\gamma\sqrt{2NV\log f(X_1)}} \sum_t c_t \{\log f(X_{t+1}) - \log f(X_t)\}$$

will for any positive γ approximate the asymptotic $N(0,1)$, reasonably well, but that $\hat{\tau}_{TR} - \tilde{\tau}_{TR}$ is relatively large for small γ, and indeed small for large γ, although asymptotically this term will vanish altogether. Drost and Werker (1993) used a large scale Monte Carlo study to show that for the categorical data equivalent of $\hat{\tau}_R$, this is indeed a serious problem. One may set $\gamma = 0$, replacing $\hat{\tau}_R$ by

$$\hat{\tau}_R^* = \sum_{t \in S} \{\log \hat{f}_{12}(X_t, X_{t+1}) - \log \hat{f}(X_t)\}.$$

One would obviously have to use simulated critical values to draw conclusions on the basis of $\hat{\tau}_R^*$, as its distribution is intractable, even (after proper rescaling) asymptotically.

Robinson's (1991a) proofs had a minor flaw in that they did not take into account any boundary effects regarding the kernel density estimates. His results are none the less valid, if the densities have infinite support, restricting the comparison of $f_{12}(x,y)$ and $f(x)f(y)$ to a compact set. This would imply that the consistency against all departures from serial independence

against serial dependence of order one were lost, but for an appropriately chosen compact set, this should in practice not make any difference.

Guerre (1991) generalised Robinson's idea to densities with infinite support. He needed a special kernel, which only takes rational values and has a large number of discontinuity points, and technical and fairly strong conditions on f and f_{12}. His paper is, however, particularly concerned with the estimation of entropies, i.e. quantities of the form $\int g(x) \log g(x) dx$, under weak dependence conditions.

In Chapter 4 we generalise Robinson's (1991a) test to densities with infinite support, and we obtain also a nuisance parameter result (see Section 3.4).

Skaug and Tjøstheim (1992a) estimated $\mathcal{I}_{ST}^*$ by

$$\hat{\tau}_{ST}^* = \frac{1}{N} \sum_t \{\hat{f}_{12}(X_t, X_{t+1}) - \hat{f}(X_t)\hat{f}(X_{t+1})\},$$

where the $\hat{f}$'s are all kernel estimates. They showed $\sqrt{N}\hat{\tau}_{ST}^*$ to have an asymptotic normal distribution. The same remarks regarding the choice of a bandwidth sequence apply as for other kernel density estimate based statistics, and the test statistic is again only available for series whose elements are continuously distributed. Skaug and Tjøstheim (1992a) argued, using Taylor series approximations, that their statistic's behaviour was likely to be similar to that of Robinson's (1991a), for $\gamma = 0$, against most alternatives.

A disadvantage of the present test, is that it is not consistent against all departures from the null against a serial dependence of order one alternative. Indeed, as noted earlier in this subsection, $\mathcal{I}_{ST}^*$ may be less than or equal to zero, under the alternative hypothesis, and a test based on $\hat{\tau}_{ST}^*$ may thus not be consistent against all departures from $f_{12}(x, y) = f(x)f(y)$, almost everywhere. An advantage is its limiting $N(0, \sigma^2)$, $0 < \sigma^2 < \infty$, distribution without the need for weights of the kind Robinson's (1991a) test requires to achieve it.

3.3.3 Characteristic Function Based Tests

It is widely known that two distributions are identical, if and only if their characteristic functions are the same [cf. e.g. Lukacs (1970), Theorem 3.1.1]. One may therefore base a test on the joint and the product of the marginal characteristic functions of (X_1, X_2).

Csörgö (1985) proposed a test based on

$$\hat{\tau}_{CS} = \hat{\psi}(\hat{u}, \hat{v})\hat{\Sigma}^{-1}(\hat{u}, \hat{v})\hat{\psi}(\hat{u}, \hat{v}), \tag{3.19}$$

where $\hat{\psi}(u, v) = \frac{1}{N}\sum_t e^{i(uX_{1t}+vX_{2t})} - \frac{1}{N}\sum_t e^{iuX_{1t}}\frac{1}{N}\sum_t e^{ivX_{2t}}$, and $(\hat{u}, \hat{v})$ is an estimate of the point at which $\psi(u, v) = Ee^{i(uX_{1t}+vX_{2t})} - Ee^{iuX_{1t}}Ee^{ivX_{2t}}$ is most variable, and $\hat{\Sigma}$ is just a weighting matrix to ensure that $\hat{\tau}_{CS}$ has an asymptotic χ^2_1 distribution. Evidently, $\hat{\psi}(u, v)$ is the difference between the joint and marginal empirical characteristic functions of (X_1, X_2) at (u, v), and estimates $\psi(u, v)$, the difference between the joint and marginal characteristic functions. Under independence, $\psi(u, v) = 0$, at all u, v, whilst under dependence, $\psi(u, v) \neq 0$, at some (u, v). Hence, under independence $\hat{\tau}_{CS}$ is expected to attain relatively small values, whilst under dependence $\hat{\tau}_{CS}$ is unbounded in probability.

We were not aware of Csörgö's (1985) test, when work on Chapter 5 began. There are however many differences between Csörgö's (1985) work and the test in Chapter 5. First, the test we propose tests a *serial* independence hypothesis, whilst Csörgö's test does not. However, it is probably feasible to generalise his test to a time series framework. Another difference is, that we do not make explicit use of the empirical characteristic function, and we integrate over the squared difference of the characteristic functions, rather than taking a supremum. Moreover, we actually use a measure that is bounded from below by the afore-mentioned integral, for performance reasons. We also require, at least for distributions with characteristic functions that are not square integrable, that a weighting function g be specified, whose choice is arbitrary. The statistic put forward in Chapter 5 also has a limiting χ^2_1 distribution under the null.

There is, as far as we are aware, no evidence on the performance of Csörgö's test, nor do we know of any empirical applications. One may also need to establish ways to determine the supremum in (3.19).

3.3.4 Rank Tests

Rank based tests have frequently been used to test for independence. An example of such a rank-based test is Spearman's rank correlation test, dating from the beginning of this century. Let there be two i.i.d. samples of equal length, both ordered in ascending order, and let the values in the original samples are replaced by their ranks. Then, the correlation between the rank numbers in both samples gives some indication of the dependence of corresponding elements in the original series. Indeed, if high ranks in one sample correspond to high ranks in the other, the correlation will be greater than zero, and it will be less than zero, if the converse is true. If the ranking in one sample is unrelated with that in the other, the correlation will be close to zero.

This is however not equivalent to corresponding elements in the sample being independent. The following example may seem somewhat contrived, but there are undoubtedly many cases in which Spearman's (or any other rank test, for that matter) will not reject any departures from the null hypothesis. Suppose that the X_{2t}'s have an even density with compact support $[-M, M]$. Suppose further that

$$X_{1t} = \begin{cases} \{M - |X_{2t}|\}\text{sgn}(X_{2t}), & |X_{2t}| \leq Q_{0.75} \\ \{|X_{2t} - M|\}\text{sgn}(X_{2t}), & |X_{2t}| > Q_{0.75} \end{cases},$$

where $\text{sgn}(x) = 2I(x > 0) - 1$, and $Q_{0.75}$ is the third quartile of X_{2t}'s distribution. Certainly, X_{1t} is not independent of X_{2t}. However, as one can easily establish, Spearman's rank correlation test will not reject this alternative.

An interesting example of a rank test for serial independence, was given by Dufour (1981). Assuming the distributions of all X_t's are even and continuous (they need indeed not be identically distributed), his test rejects any alternative for which $\text{Med}(X_1X_2) \neq 0$. Under the null hypothesis, $\text{Med}(X_1X_2) = 0$, because $P[X_1X_2 \leq 0] = P[X_1 \leq 0, X_2 \geq 0] + P[X_1 \geq 0, X_2 \leq 0] = 2P[X_1 \leq 0]P[X_1 \geq 0] = \frac{1}{2}$. If $r_N(|X_tX_{t+1}|)$ is the relative rank $|X_tX_{t+1}|$ has when the first N elements of $\{|X_tX_{t+1}|\}$ are sorted in ascending order, then Dufour's test statistic takes the form

$$\hat{\tau}_{DU} = \sum_t I(X_tX_{t+1} \geq 0)\Omega(r_N(|X_tX_{t+1}|)),$$

where Ω is some non-negative and increasing score function. Let us, for the sake of the argument assume that $\Omega(x) = x$, such that $\hat{\tau}_{DU} = \sum_t I(X_tX_{t+1})r_N(|X_tX_{t+1}|)$. If positive X_tX_{t+1}'s tend to be greater in absolute value than negative ones, $\hat{\tau}_{DU}$ will be large, whereas it will be small if the converse is true. Dufour (1981) derives an expression for the characteristic function of $\hat{\tau}_{DU}$ under the null, which of course depends on Ω, but does *not* depend on the distribution of X_1, irrespective of the sample size. This is a major advantage of rank based tests, as for most tests this is only the case, asymptotically, or is a direct result of certain parametric conditions. Moreover, as mentioned before, the X_t's need not necessarily be identically distributed under the null, as long as each of their distributions has a median that is equal to zero.

Other references in this area are Bartels (1982), Hallin, Ingenbleek and Puri (1985), Hallin and Mélard (1989), Hallin and Puri (1989), and Knoke (1977).

3.3.5 Correlation Dimension Test

A test that is widely used in finance to test for non-linearities is the correlation dimension test of Brock, Dechert and Scheinkman (1987) [cf. also Brock, Dechert, Scheinkman and LeBaron (1987)]. It has its origins in the chaos literature, and is based on the correlation dimension,

which is (for $J = 2$)

$$P[|X_1 - Y_1| \leq \lambda, |X_2 - Y_2| \leq \lambda] - P^2[|X_1 - Y_1| \leq \lambda], \tag{3.20}$$

for some practitioner-chosen $\lambda > 0$, where $\{Y_t\}$ is an independent replication of $\{X_t\}$. It is easy to see that (3.20) can also be written as $\text{Cov}[I(|X_1 - Y_1| \leq \lambda), I(|X_2 - Y_2| \leq \lambda)]$. Brock, Dechert and Scheinkman (1987) suggested to estimate (3.20) by

$$\begin{aligned} \hat{\tau}_{CD} &= \frac{24}{N(N-1)(N-2)(N-3)} \sum_t \sum_{s>t} \sum_{u>s} \sum_{v>u} \\ &\quad I(|X_t - X_s| \leq \lambda)\{I(|X_{t+1} - X_{s+1}| \leq \lambda) - I(|X_u - X_v| \leq \lambda)\}, \end{aligned}$$

and showed $\sqrt{N}\hat{\tau}_{CD}$ to have a limiting normal distribution under serial independence, which is actually a direct consequence of Theorem 3.2. They made an attempt at obtaining a nuisance parameter result in Brock, Dechert, Scheinkman and LeBaron (1987), but were not quite able to prove it.

Under serial independence, (3.20) equals zero, but this may also be the case under serial dependence of order one, even if trivial choices of λ, such as $\lambda = 0$, or a λ for which $F^*(\lambda/2) - F^*(-\lambda/2) = 1$, where F^* is the distribution function of X_1, are excluded. Indeed, suppose that F_{12} and F are the distribution functions of $(X_1 - Y_1, X_2 - Y_2)$ and $(X_1 - Y_1)$, respectively. The corresponding distributions need necessarily be symmetric. Then, $F(-x) = 1 - F(x)$, $F_{12}(-x, -y) = 1 + F_{12}(x, y) - F(x) - F(y)$, and $F_{12}(x, -y) = F_{12}(-x, y)$, where x and y should be greater than or equal to zero. Thus, $P[|X_1 - Y_1| \leq \lambda, |X_2 - Y_2| \leq \lambda] = F_{12}(\lambda, \lambda) + F_{12}(-\lambda, -\lambda) - F_{12}(\lambda, -\lambda) - F_{12}(-\lambda, \lambda)$, which is $2 + 2F_{12}(\lambda, \lambda) - 4F(\lambda) - 2F_{12}(\lambda, -\lambda)$. Under independence, the last expression is $2 + 2F^2(\lambda) - 4F(\lambda) - 2F(\lambda)\{1 - F(\lambda)\} = 4F^2(\lambda) - 6F(\lambda)$, such that the correlation dimension is zero, if and only if

$$F_{12}(\lambda, \lambda) - F_{12}(\lambda, -\lambda) = 2F^2(\lambda) - F(\lambda), \tag{3.21}$$

which may happen for many combinations of F_{12} and F. Indeed, suppose that $F_{12}(x,y) = \frac{1}{2}\{A(x)A(y)+B(x)B(y)\}$, where A and B are distribution functions corresponding to symmetric distributions, such that $F(x) = \frac{1}{2}\{A(x)+B(x)\}$. Suppose also that $A(\lambda) = B(\lambda)$. Clearly, there is any number of possibilities to choose A and B satisfying this condition. We also require that $F_{12}(x,y) \neq F(x,y)$ holds in a subset of $\Re^2$ of positive measure.

Suppose, for instance, that A is any distribution function corresponding to a symmetric distribution. Let λ have been selected such that $\frac{1}{2} < A(\lambda) < 1$. Let B be the double-exponential distribution function with parameter θ, i.e. $B(x) = \frac{1}{2}\{e^{\theta x}I(x<0)+(2-e^{-\theta x})I(x \geq 0)\}$, for all x. Set $\theta = -\log\{2-2A(\lambda)\}/\lambda$, such that

$$B(x) = \begin{cases} \frac{1}{2}\{2-2A(\lambda)\}^{-\frac{x}{\lambda}}, & x < 0, \\ 1-\frac{1}{2}\{2-2A(\lambda)\}^{\frac{x}{\lambda}}, & x \geq 0. \end{cases}$$

Clearly, $B(\lambda) = A(\lambda)$. Therefore, $F_{12}(\lambda,\lambda)-F_{12}(\lambda,-\lambda)-2F^2(\lambda)+F(\lambda) = \frac{1}{2}\{2A^2(\lambda)+2B^2(\lambda)-A(\lambda)-B(\lambda)\} - \frac{1}{2}[\{A(\lambda)+B(\lambda)\}^2 - A(\lambda) - B(\lambda)\} = 0$. One would still have to verify that F_{12}, F could indeed be the joint distribution and marginal distribution of $(X_1 - Y_1, X_2 - Y_2)$ and $X_1 - Y_1$, respectively, but it seems highly unlikely, that no combination of A and B exists for which this is the case. Although somewhat beyond the scope of this thesis, one would need to verify that the characteristic function of $(X_1 - Y_1, X_2 - Y_2)$, say $\psi_{12}(u,v)$ could be written as $\psi_{12}(u,v) = \psi(u,v)\psi(-u,-v)$, where ψ is again a characteristic function, indeed the characteristic function of (X_1, X_2). Various methods to verify whether a function is a characteristic function can be found in Lukacs (1970), Chapter 4.

3.3.6 Tests for Linearity of Processes

Hinich (1982) was interested in verifying whether

$$X_t = \sum_{s=0}^{\infty} \theta_s \varepsilon_{t-s}, \tag{3.22}$$

for all t, with $\{\varepsilon_t\}$ white noise. Both moving average and autoregressive processes can be written in the above form, and hence (3.22) is essentially a test for linearity of the process $\{X_t\}$.

If the ε_t's are Gaussian, so are the X_t's. If they are not, the X_t's will not be Gaussian, either. Similarly, when the above process is nonlinear, the X_t's will be non-Gaussian, even if the ε_t's are. Further, when the ε_t's are mean zero Gaussian, they all have $E\varepsilon_t^3 = 0$, and then the third order cumulants $E[X_t X_{t+s} X_{t+u}]$ are equal to 0, for all t, s, u, also. So when the third order cumulants are not all equal to zero, either the ε_t's are non-Gaussian or the process is nonlinear.

Hinich (1982) thus proposes a test for Gaussianity using the bispectrum assuming linearity, and a test for linearity assuming that the ε_t's are Gaussian, or rather: a test for symmetry assuming linearity, and a test for linearity assuming symmetry.

3.4 Specification Testing and Nuisance Parameters

Suppose we have formulated a model

$$A_t = g(Z_t; \theta_0) + X_t, \tag{3.23}$$

for all t, where the vector of regressors Z_t may include past values of A. There are many ways to test whether (3.23) is the correct specification.

One usually examines the structure of $\{X_t\}$. Indeed, if $\{X_t\}$ is i.i.d., there is little reason to suspect that the model is misspecified. If the X_t's are dependent, however, it is quite likely that a lagged dependent variable has been omitted, and if the X_t's prove to be heteroskedastic, one may well wish to model the form of the heteroskedasticity. Indeed, ordinary least squares depends on homoskedasticity for its efficiency properties, and there are tests that are not asymptotically valid under heteroskedasticity.

However, disturbances are not observed, except in trivial cases like the random walk model. One may sometimes use the residuals $\{Y_t\}$, with $Y_t = A_t - g(Z_t; \hat{\theta})$, with $\hat{\theta}$ a consistent estimate of θ_0, instead, however. Indeed, one can in certain circumstances prove that applying a test to the residuals will asymptotically lead to the same result as if the test had been applied to the disturbances. Problems of this type are called *nuisance parameter problems*, and θ_0 is called the nuisance parameter.

Durbin and Watson (1950, 1951, 1971) proposed to used von Neuman's (1941) mean square successive difference to variance ratio to test whether the disturbances of a standard linear regression model without lagged regressands are serially independent, where the alternative is serial correlation of order one.

The Box-Pierce test (1970) uses the Lagrange multiplier test against autoregressive alternatives obtained in Subsubsection 3.2.2.3, and uses

$$\hat{\tau}_{BP} = N \sum_{j=1}^{J-1} \left(\frac{Y_t Y_{t+j}}{\sum_t Y_t^2} \right)^2,$$

which is asymptotically χ^2_{J-1-l}, with l the dimension of θ_0 in (3.23), if the X_t's are i.i.d.. If (3.23) is a finite order invertible AR model, its results are asymptotically still valid. The model needs to be linear, however.

Originally suggested by Box and Pierce (1970), who apparently preferred the simpler form of $\hat{\tau}_{BP}$, the Ljung-Box (1978) test appeared in practice more effective than the Box-Pierce test, particularly against ARMA alternatives [cf. Box and Pierce (1970)]. It is given by

$$\hat{\tau}_{LB} = N(N+2) \sum_{j=1}^{J-1} \frac{1}{N-j} \left(\frac{Y_t Y_{t+j}}{\sum_t Y_t^2} \right)^2,$$

which has the same asymptotic distribution as $\hat{\tau}_{BP}$.

Engle (1982) showed that the Lagrange multiplier principle implies that the sum of squared errors of the regression of Y_t^2 on $Y_{t-1}^2, \ldots, Y_{t-J+1}^2$ is asymptotically χ^2_{J-1}, under Gaussianity. Again, g needs to be linear.

We show for both the entropy based test of Chapter 4, and the characteristic function based test of Chapter 5, one may, under certain conditions, use the $\{Y_t\}$ and still obtain asymptotically reliable results regarding the $\{X_t\}$. It should be noted that in both these tests, the alternative is serial dependence subject to certain regularity conditions, instead of a specific parametric alternative as is the case for the other tests discussed in this subsection. There are very few nonparametric serial independence tests for which nuisance parameter results have been obtained. Brock, Dechert and Scheinkman (1987) made an attempt, but did not quite manage to prove it to hold for their correlation dimension test. An advantage of the results we obtain over those discussed earlier in this subsection, is that they are still valid under a variety of nonlinear model specifications.

It should be noted that there are various other ways to test the specification of a model other than by examining the residuals of the model. It would be well beyond the scope of this chapter, indeed this thesis, however, to discuss such tests in detail, and we therefore refrain from doing so.

3.5 Which Test to Choose

In this chapter, we discussed a wide variety of tests for independence. Faced with a particular data set, the choice of the most suitable test is difficult, if not impossible.

From a practical perspective, independence tests are most useful when applied to the residuals of a time series or regression model. In case only linear dependence, i.e. correlation, is of importance, correlation tests such as the Durbin-Watson (1950, 1951, 1971) test, the Box-Pierce test (1970) or the Ljung-Box test (1978) are most appropriate, provided the model itself is linear and existence of second moments can be assumed. In other cases, nonparametric tests are more appropriate. There are however few nonparametric tests that allow for the presence of nuisance

parameters. The tests proposed in the next two chapters are consistent against any nonlinear dependence structure and can moreover be applied to the residuals of nonlinear models.

Chapter 4

Entropy Based Testing Revisited

4.1 Introduction

A considerable number of independence tests was described in the previous chapter. In the current chapter, we extend a result of Robinson (1991a) using the Kullback-Leibler information criterion (3.16) as the basis for a serial independence test. We extend Robinson's (1991a) test in three directions: we allow for unbounded support, we allow for alternatives of any fixed and finite order, and we provide a nuisance parameter result.

As noted in Chapter 3, Robinson's (1991a) test does not allow for unbounded support. It is possible to alter his set up slightly, such that equivalence of joint and marginal densities, that have infinite support, is verified over a compact set. Indeed, such an extension would be quite straightforward. If the compact set chosen covers a sufficiently large part of the support of the afore-mentioned densities, the probability that the joint and marginal densities only differ outside this compact set is small. However, as we show in this chapter, it is possible to test the equality of joint and marginal densities, almost everywhere.

Robinson concentrated on testing serial independence against serial dependence of order one. He did suggest a test statistic for serial independence against serial dependence of any finite order, but he did not prove its consistency or its asymptotic normality. We consider the more general case of higher order alternatives throughout.

As we have seen in Chapter 3, in many situations in which serial independence tests are of interest, nuisance parameters are present. To deal with this situation, we have proved that, under certain additional conditions, the proposed test is still consistent against serial dependence of fixed order and is also still asymptotically normal under the null, when nuisance parameters are present.

The outline of this chapter is as follows. In Section 4.2 we introduce notation and put forward our test statistic. Section 4.3 discusses the conditions required and also contains the results for the case without nuisance parameters. Section 4.4 has the same form as Section 4.3, but covers the case with nuisance parameters. In Section 4.5, we present some results regarding the size properties of our test statistic (without nuisance parameters) under the null hypothesis; a power comparison under the alternative is contained in Section 5.5 of Chapter 5. Finally, Section 4.6 summarises the results of this chapter.

4.2 Test

In this section, we introduce notation, and we describe our test statistic. Formal assumptions and results are postponed until Sections 4.3 and 4.4, for the cases with and without nuisance parameters, respectively.

We observe a stationary time series $\{X_t\}$. We denote the density of X_1 by f, and that of $X_{J1} = (X_1, \ldots, X_J)^T$ by f_J.. Notation thus slightly deviates from that used in Chapter 3.

The null hypothesis is serial independence, whilst the alternative is serial dependence of

order $1 \leq J-1 < \infty$, where $J \geq 2$ is chosen by the practitioner. Under serial independence,

$$f_{J\cdot}(x) = \prod_{j=1}^{J} f(x_j), \text{ at almost all } x \in \Re^J. \tag{4.1}$$

Condition (4.1) does itself not imply serial independence. Indeed, for a process with $X_t = \theta X_{t-J} + \varepsilon_t$, for every t, with $\{\varepsilon_t\}$ white noise and $|\theta| < 1$, (4.1) will hold. It does however imply the absence of serial dependence of order $J-1$. The test proposed in this chapter is consistent against all departures from (4.1) subject to certain regularity conditions.

Our test statistic is, like Robinson's (1991a), based on the Kullback-Leibler information criterion [cf. also (3.16)], given by

$$\mathcal{I}_{KL} = \int f_{J\cdot}(x) \log \frac{f_{J\cdot}(x)}{\prod_{j=1}^{J} f(x_j)} dx, \tag{4.2}$$

which is zero if and only if (4.1) holds; otherwise it is greater than zero. We could substitute nonparametric density estimates for all densities in (4.2), but, as Robinson, we prefer to rewrite (4.2) as $E \log f_{J\cdot}(X_{J1}) - E^J \log f(X_1)$, and estimate it by

$$\hat{\mathcal{I}} = \hat{U}_J - \hat{U}_1, \tag{4.3}$$

where $\hat{U}_J = \frac{1}{N} \sum_{t \in S} c_{Jt} \log \hat{f}_{Jt}$, and $\hat{U}_1 = \frac{1}{N} \sum_{t \in S} c_{1t} \sum_{j=0}^{J-1} \log \hat{f}_{t+j}$. The set $S \subseteq \{1, \ldots, N\}$ serves to trim out certain observations; a more detailed explanation follows in Section 4.3. The restrictions on the weight sequences $\{c_{Jt}\}$ and $\{c_{1t}\}$ will also be explained in Section 4.3; their function is explained further below in this section. Weight sequences were also used by Robinson (1991a), although he used a different set of weights. The density estimates in the definitions of $\hat{U}_J$ and $\hat{U}_1$ were chosen to be kernel density estimates, which were introduced in Chapter 1.

To prove $\hat{\mathcal{I}} \xrightarrow{P} \mathcal{I}_{KL}$, we shall, like Robinson, make use of an intermediary quantity, $\tilde{\mathcal{I}}$, defined by

$$\tilde{\mathcal{I}} = U_J - U_1, \tag{4.4}$$

with $U_J = \frac{1}{N}\sum_{t\in S} c_{Jt}\log f_{Jt}$, and $U_1 = \frac{1}{N}\sum_{t\in S} c_{1t}\sum_{j=0}^{J-1}\log f_{t+j}$.

In Section 4.3, we show that $\hat{\mathcal{I}} \xrightarrow{P} \mathcal{I}_{KL}$, where $\mathcal{I}_{KL}$ was defined in (4.2). As noted before, $\mathcal{I}_{KL}$ is zero if and only if (4.1) holds, and is otherwise greater than zero. It is therefore natural to base our test on $\hat{\mathcal{I}}$. Let $\hat{\mathcal{V}} = \frac{1}{N}\sum_{t\in S}\log^2 \hat{f}_t - (\frac{1}{N}\sum_{t\in S}\log \hat{f}_t)^2$ be an estimate of the variance of $\log f_1$ (in the proof to Theorem 4.1, we show that $\hat{\mathcal{V}} \xrightarrow{P} V\log f_1$). Let

$$\hat{\tau} = \sqrt{\frac{N}{2J\hat{\mathcal{V}}^*}}\hat{\mathcal{I}}, \tag{4.5}$$

where $\hat{\mathcal{V}}^* = \sum_{l=0}^{2J-1}\sum_{j=0}^{J-1}\tilde{c}_{4J-j-l}^2\hat{\mathcal{V}}$, with $\tilde{c}_t = c_{Jt} - c_{1t}$, for all t.

Under the null hypothesis, $f_{Jt} = \sum_{j=1}^{J-1} f_{t+j}$, for all t, such that $\tilde{\mathcal{I}} = \frac{1}{N}\sum_{t\in S}\tilde{c}_t\log f_{Jt}$. Hence, if c_{Jt} and c_{1t} are identical at all t, $\tilde{\mathcal{I}} = 0$, and the limiting distribution is determined by the nonparametric approximation of $\tilde{\mathcal{I}}$, which is rather hard to obtain. Of course, one could still use a bootstrap procedure, but this is evidently more involved than using the quantiles of a known distribution.

Aside from arbitrariness, the introduction of weights has disadvantages in other respects, also. The introduction of weights makes the test less efficient. Indeed, the reason that we can not find an asymptotic distribution for $\hat{\mathcal{I}}$ when all weights are set to one, is that $\hat{\mathcal{I}}$'s convergence to $\tilde{\mathcal{I}}$ is slower than that of $\tilde{\mathcal{I}}$ to $\mathcal{I}_{KL}$. In other words, in that case the slowest converging part is the nonparametric approximation of $\tilde{\mathcal{I}}$. The weights in effect slow down the convergence of $\tilde{\mathcal{I}}$ to $\mathcal{I}_{KL}$ to make *that* the slowest converging part. Another problem with choosing weights different from one is that they imply a greater degree of arbitrariness; different weights lead to different results.

In Section 4.3, we restrict the choice of weights, and these restrictions do not allow for the weights to all be equal to one. However, the proofs of the theorems in Section 4.3 imply that in the case all weights are equal to one, $\hat{\mathcal{I}} = o_p(N^{-\frac{1}{2}})$, when $\{X_t\}$ are i.i.d.. Indeed, the convergence rate can be made much greater still, depending on the smoothness of f.

If the weights are chosen under the restrictions of Section 4.3, and the other conditions made therein are also satisfied, we show (in Theorem 4.2) that $\hat{\tau} \xrightarrow{\mathcal{L}} N(0,1)$. Our test can then be formulated as:

"Reject serial independence if and only if $\hat{\tau} > C$",

where C is a critical value based upon the quantiles of the standard normal distribution.

4.3 Standard Case

In this section, we establish conditions under which the proposed test is consistent against departures from (4.1), and also establish its asymptotic validity. Most conditions could be relaxed if others were strengthened, and we have tried to find conditions that are relatively easy to express.

Assumption 4.A *The stationary and ergodic series* $\{X_t\}$ *is trigonometric mixing with summable mixing numbers* $\{\alpha(t)\}$.

Trigonometric mixing was extensively discussed in Chapter 3. It is imposed in addition to serial dependence of order $J-1$, and is required for certain convergence results.

Definition 4.1 $\mathcal{G}_\mu$, $\mu > -1$, *is the class of functions* g *satisfying:* g *is* ς *times partially differentiable for* $\varsigma \leq \mu \leq \varsigma+1$; *for some* $\rho_g > 0$, $\sup_{x \in \mathcal{S}_{y\rho_g}} ||g(x)-g(y)-Q_{g\varsigma}(x,y)||/||x-y||^{\mu+1} \leq \xi_{g\varsigma}(y)$ *for all* y, *where the neighbourhood* $\mathcal{S}_{y\rho_g}$ *is defined by* $\mathcal{S}_{y\rho_g} = \{x : 0 < ||x-y|| < \rho_g\}$; $Q_{g\varsigma}$ *is the* ς*-th order Taylor expansion of* g, *and* $\xi_{g\varsigma}$ *is bounded.*

Definition 4.1 is a slightly modified version of a definition used by Robinson (1988) and Hidalgo (1993). It enables us to deal with Taylor series expansions in a more straightforward manner. We only apply Definition 4.1 to densities, for which it is implied by the density being $\varsigma + 1$

times boundedly differentiable. Let $g : \Re \to \Re$ be some function in $\mathcal{G}_1$. Then $\varsigma = \mu = 1$, such that g is twice boundedly differentiable, and $Q_{g\varsigma}(x,y) = (x-y)g'(y)$. As a consequence, $g(x) - g(y) - Q_{g\varsigma}(x,y) = (x-y)^2 g''(x;y)/2$, where $(x;y)$ is some number between x and y. As g'' is twice boundedly differentiable, $|g(x) - g(y) - Q_{g\varsigma}(x,y)|/|x-y|^2 \leq |g''(x;y)/2| < \infty$, provided $x \neq y$.

Assumption 4.B *For some integer $r > 0$ and some $\omega \in (0,1)$, the density f of X_1 is in $\mathcal{G}_{r+\omega}$, and $f_{J\cdot}$, the density of X_{J1}, is in $\mathcal{G}_{r+\omega}$. $E\log f(X_1)$, $E\log^2 f(X_1)$ and $E\log f_{J\cdot}(X_{J1})$ exist, and there are two sequences L and B that both tend to ∞ as $N \to \infty$, for which $\sup_{\|x\|_\infty \leq B} 1/f_{J\cdot}(x) < L$.*

The existence of $E\log f(X_1)$ and $E\log f_{J\cdot}(X_{J1})$ ensures that $\mathcal{I}_{KL}$ in (4.2) exists. The numbers r and ω in Assumption 4.B relate to the smoothness of the densities. Assumption 4.B does itself not impose any conditions on r and ω, and hence it does not impose any smoothness conditions on f and $f_{J\cdot}$, either. However, r and ω are restricted further below. The rationale for separating $r+\omega$ into an integer and a remainder part is that $r+1$ is the order of the kernels employed (cf. Assumption 4.C, below).

Kernel density estimates may be close to or equal to zero. If higher order kernels are employed, as in Assumption 4.C, they may even take negative values. As indicated in the previous section, we are particularly interested in the logarithms of the densities. It is most natural to estimate these by the logarithms of the estimated densities. However, if the estimated densities can take zero or even negative values, their logarithms will not be defined. We therefore restrict the density estimates used to those with arguments lying in a certain interval. This interval is controlled by the sequence B, which is chosen by the practitioner. As implied by Assumption 4.B, the interval grows with sample size, such that asymptotically the equivalence of joint and marginal densities is verified everywhere. The sequence L then represents the rate at which the

inverse of the joint density of $(X_1, \ldots, X_J)$ tends to ∞ for a given choice of B. For a finite number of observations and some choice of B, the positiveness of all density estimates is not implied. However, the probability of any of the density estimates, that are employed in the construction of $\hat{I}$, being non-positive decreases rapidly with sample size under the conditions imposed below.

In practice, one knows neither the density of $(X_1, \ldots, X_J)$ nor that of X_1. In the conditions imposed in this and subsequent sections, it is assumed that the practitioner has some knowledge about the thickness of the tails of the afore-mentioned densities. The conditions are constructed for densities with infinite support, although the conditions can probably be altered for the case of densities with finite support. It is not necessary to know the exact rate at which the density at x tends to zero as $|x| \to \infty$, but it is necessary to assume a bound. If the conditions below necessitate $L \sim N^q$, for any $q > 0$, for instance, and the tails of f and $f_{J\cdot}$ resemble that of a Gaussian density, $B \sim \sqrt{\log N}$ will satisfy the relevant conditions of Assumption 4.B.

Assumption 4.C *For r and ω used in Assumption 4.B, the univariate kernel $k \in \mathcal{G}_{r+\omega}$, is an $r+1$-th order kernel, i.e. $\int k(x)dx = 1$, $\int k(x)x^l dx = 0$, $l = 1, \ldots, r$, $\int k(x)x^{r+1}dx \neq 0$. k can be written as $k(x) = \int \phi_k(u)e^{iux}du$, for all x, with $\int |\phi_k(u)|du < \infty$. Further $\int |k(x)x|^{r+1+\omega}dx < \infty$. The multivariate kernels $k_{J\cdot}(x)$ are defined as $\prod_{j=1}^{J} k(x_j)$, for any $x \in \Re^J$, and can thus be written as $k_{J\cdot}(x) = \int \phi_{Jk}(u)e^{iu^T x}du$, for all x, with $\phi_{Jk}(x) = \prod_{j=1}^{J} \phi_k(x_j)$. Finally, $|K(x) - 1|$ and $|K(-x)|$ are both decreasing at any sufficiently large x, where $K(x) = \int_{-\infty}^{x} k(y)dy$.*

We employ kernel density estimates as $\hat{f}$ and $\hat{f}_{J\cdot}$, and they require both the choice of a kernel and of a bandwidth. The kernels are restricted in Assumption 4.C above; restrictions on the bandwidth are imposed in Assumptions 4.D and 4.F. Unlike Robinson (1991a), we use the same bandwidth in both univariate and multivariate density estimates. The advantage is that there is one less input parameter to choose, the disadvantage that there is less flexibility.

Because the kernel can be chosen by the practitioner, any conditions imposed on it are not as serious as those on the density, for instance. Most of the conditions in Assumption 4.C are fairly standard. An exception is that $|K(x)-1|$ and $|K(-x)|$ both be decreasing at any sufficiently large x. This is a technical (yet mild) condition that is imposed to facilitate the asymptotic validity proof.

It is a well-known fact that kernels of any order can be constructed, but accuracy in practice does not tend to improve much with increasing kernel order in finite samples. The reason is that, although bias is reduced, the variance increases with the kernel order chosen. Further, the greater r, the smoother f and f_J. need be.

Let $\lambda = |K(B)-1| + |K(-B)| + 1 - F(B) + F(-B)$. λ is a technical number that is used in some of the assumptions and proofs, but that has no direct meaning to the practitioner, in the sense that he/she does not have to choose it. λ tends to zero, because $B \to \infty$, F is a distribution function, and Assumption 4.C.

The $a = 1, J$-variate kernel density estimate at x is defined as $\hat{f}_{a\cdot}(x) = \frac{1}{Nh^a}\sum_s k_{ah}(x - X_{as})$, and $\hat{f}_{at} = \frac{1}{Nh^a}\sum_{s \notin D_t} k_{ah}(X_{at} - X_{as})$, where $k_{ah}(x) = k_{a\cdot}(x/h)$, and $D_t = \{t-a+1, \ldots, t, \ldots, t+a-1\}$. The set D_t is introduced to avoid noise due to overlapping terms, e.g. when $J = 2$, X_{2t} and $X_{2,t+1}$ overlap, and certain terms should thus be excluded; $D_t = \{t-1, t, t+1\}$ performs this task.

Assumption 4.D *As* $N \to \infty$,

$$\begin{aligned} Nh^{2J}L^{-2} &\to \infty, \\ h^{r+\omega}L &\to 0, \end{aligned}$$

If f and f_J. are sufficiently smooth, Assumption 4.D can be easily satisfied by choosing r large, provided that L increases at a rate no faster than $N^{\frac{1}{2}-\epsilon}$, for some $\epsilon > 0$. Indeed, choose

$(r+\omega) > J\{(2\epsilon)^{-1}-1\}$, and let $h \sim N^{-\frac{1}{2}(\frac{1-2\epsilon}{2(r+\omega)}+\frac{\epsilon}{J})}$. Because, by the choice of $(r+\omega)$, $\frac{1-2\epsilon}{2(r+\omega)} < \frac{\epsilon}{J}$, $h^{-1} = o(N^{\frac{\epsilon}{J}})$, and hence $N^{-1}h^{-2J}L^2 = o(N^{-1+2\epsilon+1-2\epsilon}) = o(1)$, as $N \to \infty$, which implies $Nh^{2J}L^{-2} \to \infty$, as $N \to \infty$. Similarly, $h = o(N^{\frac{2\epsilon-1}{2(r+\omega)}})$, such that $h^{(r+\omega)}L = o(N^{\frac{1}{2}(2\epsilon-1+1-2\epsilon)}) = o(1)$.

A final issue we should discuss before stating the first theorem, is that of the choice of weights. The structure we impose now is more restrictive than necessary for Theorem 4.1, but to improve readability, we discuss all conditions imposed on the c_{at}'s now. The weights are bounded, centred around one, and take finitely many different values. The number of repetitions of each of these values increases at the same rate as N, the number of observations. Finally, we require $\sum_{t=1}^{2J} \tilde{c}_{l+t} = 0$, for all $l \in \aleph$. An example of such a weight sequence is $c_{1t} = 1$, for all t, $c_{Jt} = 1+\gamma$, $t = 2l+j$, $l = 0, 2J, \ldots$, $j = 1, \ldots, J$, and $c_{Jt} = 1-\gamma$, $t = 2l+j$, $l = J, 3J, \ldots$, for any $\gamma > 0$, such that for any $l \in \aleph$, $\sum_{t=1}^{2J} \tilde{c}_{l+t} = \sum_{t=l+1}^{l+2J}(c_{Jt} - c_{1t}) = 0$.

Theorem 4.1 *Let Assumptions 4.A, 4.B, 4.C, and 4.D hold. Then the test based on $\hat{\tau}$ proposed in the previous section is consistent against all departures from (4.1).*

In the remainder of this section, we investigate the behaviour of our test under the null hypothesis: serial independence. Our first assumption, Assumption 4.E, is an obvious one.

Assumption 4.E *The series $\{X_t\}$ is i.i.d..*

We require stronger conditions on the bandwidth sequence.

Assumption 4.F *As $N \to \infty$,*

$$Nh^{2(r+\omega)}L^2 \to 0,$$

$$Nh^{2J}L^{-4} \to \infty.$$

Assumption 4.F is stronger than Assumption 4.D. If f is very smooth (which now implies that f_J. is very smooth as the X_t's are i.i.d.), r can be chosen large. As long as L does not tend to

∞ faster than $N^{\frac{1}{4}-\epsilon}$, for some $\epsilon > 0$, both an r and a rate for h can be found that satisfy the conditions of Assumption 4.F, provided that the smoothness of f allows for the choice of a large r.

To give an example, let $(r+\omega) > \frac{J(3-4\epsilon)}{8\epsilon}$, and let $h \sim N^{-\frac{1}{2}(\frac{2\epsilon}{J}+\frac{3-4\epsilon}{4(r+\omega)})}$. Then $h = o(N^{-\frac{3-4\epsilon}{4(r+\omega)}})$, such that $Nh^{2(r+\omega)}L^2 = o(N^{1-(\frac{3}{2}-2\epsilon)+\frac{1}{2}-2\epsilon}) = o(1)$, as $N \to \infty$. Also $h^{-1} = o(N^{\frac{2\epsilon}{J}})$, such that $N^{-1}h^{-2J}L^4 = o(N^{-1+4\epsilon+(1-4\epsilon)}) = o(1)$, as $N \to \infty$, which implies $Nh^{2J}L^{-4} \to \infty$.

Theorem 4.2 *Let Assumptions 4.B, 4.C, 4.E, and 4.F hold. Then* $\hat{\tau} \xrightarrow{\mathcal{L}} N(0,1)$.

Theorem 4.2 implies that the test *"Reject serial independence when* $\hat{\tau} > C$*"* will be asymptotically valid, i.e. that asymptotically the probability of rejection of the null hypothesis, when it is correct, will be equal to $\Phi(-C)$, where Φ is the distribution function of the standard normal distribution, irrespective of the value of C.

4.4 Nuisance Parameters

When nuisance parameters are present, the residuals, say $\{Y_{Nt}\}$, are not trigonometric mixing, nor are they i.i.d. under the null. Moreover, they often do not even have a density. The series of interest $\{X_t\}$ does have these properties and, as we shall see, if the two series are sufficiently close and a number of additional regularity conditions are satisfied, we may replace $\{X_t\}$ by $\{Y_{Nt}\}$ in the test described in Section 4.2, where the test results will remain valid with respect to $\{X_t\}$, also.

Let $S_Y = \{t : ||Y_{NJt}||_\infty \leq B\}$, and rename the set S, defined in Section 4.2, to S_X to avoid confusion. $\hat{f}^Y_{a\cdot}(x) = \frac{1}{Nh^a}\sum_s k_h(x - Y_{Nas})$ is, for any $a \leq J$, the a-variate kernel density estimate based on $\{Y_{Nt}\}$ at x. However, when its argument is a random variable with suffix t, its

definition changes, e.g. $\hat{f}^Y_a(Y_{Nat}) = \frac{1}{Nh^a}\sum_{s\notin D_t} k_h(Y_{Nat} - Y_{Nas})$. The reason we do not employ the notation $\hat{f}^Y_{at}$, in analogy to $\hat{f}_{at}$ of the previous sections, is that it would not always be clear whether its argument were Y_{Nat} or X_{at}. We define $\hat{U}^Y_1 = \frac{1}{N}\sum_{t\in S_Y} c_{1t}\sum_{j=0}^{J-1}\log\hat{f}^Y(Y_{N,t+j})$, $\hat{U}^Y_J = \frac{1}{N}\sum_{t\in S_Y} c_{Jt}\log\hat{f}^Y_J(Y_{NJt})$. $\hat{\mathcal{I}}$ defined in (4.3) is replaced by

$$\hat{\mathcal{I}}^Y = \hat{U}^Y_J - \hat{U}^Y_1. \tag{4.6}$$

Let $\hat{\mathcal{V}}_Y = \frac{1}{N}\sum_{t\in S_Y}\log^2\hat{f}^Y(Y_{Nt}) - (\frac{1}{N}\sum_{t\in S_Y}\log\hat{f}^Y(Y_{Nt}))^2$, and $\hat{\mathcal{V}}^*_Y = \sum_{l=0}^{2J-1}\sum_{j=0}^{J-1}\tilde{c}^2_{4J-j-l}\hat{\mathcal{V}}_Y$ replace $\hat{\mathcal{V}}$ and $\hat{\mathcal{V}}^*$, respectively, such that our test statistic $\hat{\tau}$ in (4.5) can be replaced by

$$\hat{\tau}^Y = \sqrt{\frac{N}{2J\hat{\mathcal{V}}^*_Y}}\hat{\mathcal{I}}^Y. \tag{4.7}$$

The proposed test under nuisance parameters thus becomes

"Reject serial independence if and only if $\hat{\tau}^Y > C$",

where C is a critical value based upon the quantiles of the standard normal distribution. The conditions described below ensure that the above test is consistent against all departures from (4.1) for which these conditions hold, and also that under the null, $\hat{\tau}^Y \xrightarrow{\mathcal{L}} N(0,1)$.

In order to achieve the result described in the previous paragraph, we need to describe the relationship between $\{Y_{Nt}\}$ and $\{X_t\}$ in more detail.

Assumption 4.G *Some function m, some vector series $\{Z_t\}$, and two vectors $\hat{\theta}$ and θ_0 exist such that Y_{Nt} and X_t can be written as $Y_{Nt} = m(Z_t;\hat{\theta})$ and $X_t = m(Z_t;\theta_0)$, respectively.*

An example of a model satisfying Assumption 4.G is any parametric (non)linear regression or time series model. The assumptions below will restrict the class of models allowed. The vector θ_0 in Assumption 4.G is the vector of nuisance parameters, and $\hat{\theta}$ is an estimate thereof. Consider the AR(1) (first order autoregressive model) $A_t = \theta_0 A_{t-1} + X_t$, where $\{X_t\}$ is trigonometric mixing and $|\theta_0| < 1$. Then $X_t = A_t - \theta_0 A_{t-1}$, and $Y_{Nt} = A_t - \hat{\theta}A_{t-1}$,

such that $m(Z_t;\theta) = A_t - \theta A_{t-1}$, for all t, θ, where $Z_t = (A_t, A_{t-1})^T$. Similarly, in an ARCH(1) model, $A_t = X_t\sqrt{1+\theta_0 A_{t-1}^2}$, with $\{X_t\}$ trigonometric mixing and θ_0 sufficiently small, $X_t = A_t(1+\theta_0 A_{t-1}^2)^{-\frac{1}{2}} = m(Z_t;\theta_0)$.

Assumption 4.H *The series $\{(X_t, Z_t)\}$ is stationary and ergodic.*

Assumption 4.H disallows linear trends amongst the Z_t. If the practitioner has formulated a model without trends, then Assumption 4.H will usually not be prohibitive.

Assumption 4.I *The parameter vector θ_0 and its estimate $\hat{\theta}$ lie in a parameter space Θ. $\hat{\theta} - \theta_0 = N^{-\frac{1}{2}}\zeta$, where $\zeta = O_p(1)$ is some random vector.*

Assumption 4.I requires that the vector of nuisance parameters can be estimated $\sqrt{N}$-consistently, which is true for most parametric models. The separation of $\hat{\theta} - \theta_0$ is made for notational convenience in the proofs to the theorems below. In the AR(1) model above, if θ_0 were estimated by ordinary least squares,

$$\hat{\theta} - \theta_0 = N^{-\frac{1}{2}} \frac{\frac{1}{\sqrt{N}}\sum_t X_t A_{t-1}}{\frac{1}{N}\sum_t A_{t-1}^2}, \tag{4.8}$$

and ζ is hence the fraction on the right hand side in (4.8). If X_t is uncorrelated with A_{t-1}, $\sum_t X_t A_{t-1} = O_p(N^{\frac{1}{2}})$, and if EA_{t-1}^2 exist, $\sum_t A_{t-1}^2 = O_p(N)$. Under those conditions, evidently $\zeta = O_p(1)$.

Assumption 4.J *m defined in Assumption 4.G is twice differentiable, $E|m'(Z_1;\theta_0)\zeta| < \infty$, and $\sup_{t;\theta\in\Theta} ||m''(Z_t;\theta)||_\infty = O_p(G)$, for some sequence G which may increase with sample size, where $||\cdot||_\infty$ applied to a matrix means the largest element in absolute value, where the derivatives are taken with respect to θ, and where G is a sequence of numbers that may increase with sample size, possibly to ∞.*

The assumption that m be twice differentiable is not overly restrictive. Indeed, most parametric models satisfy this condition, as we will demonstrate in our example below. The condition that

$E|m'(Z_1;\theta_0)\zeta|$ be finite all but implies that $\hat{\theta}$ should have existing first moments, which may not hold for distributions with very thick tails. In our AR(1) example, $m'(Z_t;\theta) = -A_{t-1}$, and $m''(Z_t;\theta) = 0$, such that Assumption 4.J is not overly restrictive in this case. For the ARCH(1) example, it is easily seen that $m''(Z_t;\theta) = \frac{3}{4}A_{t-1}^4 A_t(1+\theta A_{t-1}^2)^{-\frac{5}{2}}$. For non-negative θ, $m''(Z_t;\theta)$ is thus in absolute value bounded by $A_{t-1}^4|A_t|$. Thus, the last restriction in Assumption 4.J requires that $\sup_t |A_{t-1}^4 A_t| = O_p(G)$, where the supremum runs from 1 to N. Later on, we restrict the rate at which G can increase, but it can always be chosen to increase at a rate no less than $N^{\frac{1}{2}}$, which seems adequate for most purposes. In the proofs we use a more compact notation for m and its derivatives than is employed here. We shall write $m_t = m(Z_t;\theta_0)$, $m_t' = m_t'(\theta_0) = m'(Z_t;\theta_0)$, and $m_t''(\theta) = m''(Z_t;\theta)$.

In the previous section, we introduced the sequence λ. We alter its definition somewhat, and now use $\lambda = |K(B) - 1| + |K(-B)| + 1 - F(B) + F(-B) + 1 - F_J.(B_J) + JF(B)$, where $B_J = [B, \ldots, B]^T \in \Re^J$.

Assumption 4.K *As* $N \to \infty$,

$$\lambda \log L \to 0, \tag{4.9}$$

$$h^{r+\omega} L \to 0, \tag{4.10}$$

$$Nh^{2J} L^{-2} \to \infty, \tag{4.11}$$

$$Nh^J L^{-1} G^{-1} \to \infty. \tag{4.12}$$

Assumption 4.K is somewhat stronger than Assumption 4.D, and also restricts the sequence G, introduced in Assumption 4.J, and the sequence λ. Assumption 4.K can nonetheless easily be satisfied. Let $L, G \sim N^{\frac{1}{3}}$, which is not overly restrictive as explained in this and the previous section. Then conditions (4.11) and (4.12) are implied by $N^{\frac{1}{3}}h^{2J} \to \infty$, or by $Nh^{6J} \to \infty$, and (4.10) by $Nh^{3(r+\omega)} \to 0$, as $N \to \infty$. Choosing $r + \omega > 2J$ and letting $h \sim N^{-\frac{2}{3(r+\omega)+6}}$

ensures that Assumption 4.K holds. Condition (4.9) is not very restrictive. Consider e.g. $1 - F_J.(B_J) = P[X_1 \geq B \vee X_2 \geq B \vee \ldots \vee X_J \geq B] \leq JP[X_1 \geq B] = J(1 - F(B))$. Hence, if $(1-F(B))\log L \to 0$, as $N \to \infty$, (4.9) holds. Assumption 4.B requires $L \geq \sup_{||x||_\infty \leq B} f_J^{-1}(x)$. Suppose that some $\eta > 0$ exists, such that for any sufficiently large N, $L \leq f^{-\eta}(B)$. Then $(1 - F(B))\log L = \eta(F(B) - 1)\log f(B)$. Let $g(x) = F(x^{-1})$, such that the afore expression can be written as $\lim_{\epsilon \to 0}\{g(B^{-1}) - g(\epsilon)\}\log f(B) = \lim_{\epsilon \to 0}\{B^{-1} - \epsilon\}g'(B^{-1}; \epsilon)\log f(B) = \lim_{\epsilon \to 0}\{B^{-1} - \epsilon\}\{B^2 f(B^{-1}); \epsilon^{-2} f(\epsilon^{-1})\}\log f(B) = Bf(B^{-1})\log f(B) = o(1)$, if $f(x) = o(x^{-2})$, as $x \to \infty$. This example thus excludes the case of Cauchy-distributed series, but they are, in this section, excluded by other conditions, anyway.

Theorem 4.3 *Let Assumptions 4.A, 4.B, 4.C, 4.G, 4.H, 4.I, 4.J, and 4.K hold. Then the test based on $\hat{\tau}^Y$ proposed earlier in this section is consistent against all departures from (4.1).*

The conditions required for consistency were not overly strong. However, the conditions need to be strengthened considerably for the test to be asymptotically valid in the presence of nuisance parameters. Indeed, many important models such as ARCH are excluded. This is unfortunate, but we found no way to avoid it. It does, however, not mean that one cannot use the entropy based test in these models. In Theorem 4.3, we have shown that the test is still consistent against departures from (4.1), even for the above models. We do lose asymptotic normality and the $\sqrt{N}$-norming, but we could still use simulated critical values to establish whether there is serial independence or not.

Assumption 4.L *For any $s \geq t$, X_s is independent of $m'(Z_t; \theta_0)$.*

Assumption 4.L is an exogeneity condition, and is, in our view, the strongest condition imposed in this section. The first order autoregressive and moving average models satisfy this condition, as they have $m'_t = -A_{t-1}$, and $m'_t = -X_{t-1}$, respectively. For a first order ARCH model,

however, $m_t' = -\frac{1}{2}A_t A_{t-1}^2(1+\theta_0 A_{t-1}^2)^{-\frac{3}{2}}$, for all t, and m_t' can obviously not be independent of X_t, for all t.

Assumption 4.M $E|m'(Z_1;\theta_0)\zeta|^{2(r-1)} = O(1)$.

Assumption 4.M imposes stronger moment conditions on m' and ζ. The implication for the AR(1) model described above is that $E|A_{t-1}N^{\frac{1}{2}}\sum_s X_s A_{s-1}(\sum_s A_{s-1}^2)^{-1}|^{2(n-1)} = O(1)$. For large n, Assumption 4.M is thus restrictive even for a model as convenient as the AR(1) model. The second condition in Assumption 4.M is somewhat unusual. In Assumption 4.J, we already required that $\hat{\theta} - \theta_0 = O_p(N^{-\frac{1}{2}})$, and with the conditions on m imposed in Assumption 4.J, it can be shown (cf. Lemma 4.6) that this implies that $Y_{Nt} - X_t = O_p(N^{-\frac{1}{2}})$, for all t (pointwise). Assumption 4.J did not require the expectation of $Y_{Nt} - X_t$ to exist, not even asymptotically, but Assumption 4.M requires $Y_{Nt} - X_t$'s q-th moment to exist. For the AR(1) example, $E|Y_{Nt} - X_t|^q = E|(\hat{\theta} - \theta_0)A_{t-1}|^q (E|\hat{\theta} - \theta_0|^{2q} E|A_{t-1}|^{2q})^{\frac{1}{2}} = N^{-\frac{q}{2}}(E|\zeta|^{2q}E|A_{t-1}|^{2q})^{\frac{1}{2}}$, and the existence of $E|\zeta|^{2q}$ and of $E|A_{t-1}|^{2q}$ is thus required.

Assumption 4.N $f(B), f(-B) = O(L^{-\frac{1}{J}})$, *as* $N \to \infty$.

In Assumption 4.B, we required that $\sup_{||x||_\infty \leq B} f_J^{-1}(x) \leq L$, which under serial independence equates to $\sup_{||x||_\infty \leq B} \prod_{j=1}^J f^{-1}(x_j) \leq L$. If all marginal densities were strictly decreasing with increasing $|x|$, then the afore condition is equivalent to $\prod_{j=1}^J f^{-1}(B) \leq L$, or $f^{-1}(B) \leq L^{\frac{1}{J}}$. Assumption 4.N thus replaces $\leq$ by $\sim$; a somewhat weaker condition could be imposed if the conditions of Assumption 4.O are strengthened. Note that we only examined a strictly decreasing f for exposatory purposes above; Assumption 4.N in no way implies it.

Assumption 4.O *For some sequence* Υ, *for which* $\Upsilon \to 0$, *as* $N \to \infty$,

$$\lambda \log L \to 0, \tag{4.13}$$

$$NL^{-\frac{2}{J}}\Upsilon^2 \log^2 L \to 0, \tag{4.14}$$

$$N \log^{-\frac{2}{q-1}} L \Upsilon^{\frac{2q}{q-1}} \to \infty, \tag{4.15}$$

$$Nh^{2(r+\omega)} L^2 \to 0, \tag{4.16}$$

$$Nh^{4+4J} L^{-4} \to \infty, \tag{4.17}$$

$$Nh^{\frac{4+4J}{3}} G^{-\frac{4}{3}} L^{-\frac{4}{3}} \to \infty, \tag{4.18}$$

as $N \to \infty$.

Condition (4.13) was copied from Assumption 4.K. Conditions 4.14 through 4.18 are, however, quite restrictive. The sequence Υ is entirely imaginary; existence of such a sequence is sufficient. The greater q, the easier it is to satisfy (4.14) and (4.15) simultaneously, but the stronger the corresponding moment condition of Assumption 4.M is. In the example we now give, J is assumed even. It is not hard to see that the conditions of Assumption 4.O can also be satisfied when J is odd (e.g. by letting $q = 6J$, rather than $q = \frac{11J}{2}$).

Suppose $L \sim N^{\frac{2}{21}}$, $q = \frac{11J}{2}$, and let $\Upsilon = N^{-\frac{231J-43}{462J}}$. Then both (4.14) and (4.15) are satisfied. Condition (4.17) then translates into $N^{\frac{13}{21}} h^{4+4J} \to \infty$, or $Nh^{\frac{84}{13}(1+J)} \to \infty$, for which $h \sim N^{-\frac{1}{7J}}$ will do. If we let $G \sim N^{\frac{1}{3}}$, (4.18) is also satisfied, and if $r + \omega = 3J$, so is (4.16).

The rates chosen in the example above are a bit awkward, but were convenient in their derivation. Obviously, one could adjust the convergence rates of the various sequences. The faster $L \to \infty$, the smaller $r + \omega$ may be, and hence the less smooth f needs to be. On the other hand, when L is large, we have more freedom in selecting q, and the moment restrictions of Assumption 4.M are then less serious.

Theorem 4.4 *Let Assumptions 4.B, 4.C, 4.E, 4.G, 4.H, 4.I, 4.J, 4.L, 4.M, 4.N, and 4.O hold. Then* $\hat{\tau}^Y \xrightarrow{\mathcal{L}} N(0,1)$, *and the proposed test is asymptotically valid in the presence of nuisance parameters.*

4.5 Simulations

In our experiments, we have tried to establish whether the tails of the moderate sample distribution of the test statistic were close to those of its asymptotic distribution, and further against which alternatives the proposed test will have power.

We created 8192 i.i.d. $N(0,1)$ series with 250 elements each for every entry in Table 1. The kernel used is a sixth order polynomial on $[-1,1]$, and it is a fourth order kernel. The weights $\{c_{Jt}, c_{1t}\}$ are chosen according to the rule suggested just before Theorem 4.1 for a variety of values of γ, and three different bandwidths.

For $\gamma = 0$, the asymptotic distribution of our test statistic is not normal, so we expect the approximation to the normal distribution not to be very good for values of γ close to zero. However, even for values of γ as large as 6.4, the normal approximation is poor, for any of the bandwidth choices. For moderate samples, we will therefore need to use simulated critical values, irrespective of the value of γ, and as the proposed test is more powerful for $\gamma = 0$ than for non-zero values, we recommend using $\gamma = 0$.

We should point out that, although the above results do not look promising, it need not necessarily be the case that the approximation is poor for all weight sequences imaginable. There is any number of ways to select the weight sequence, and a definite answer regarding its optimal choice will be hard to find.

In Section 5.5 we make a comparison of a range of serial independence tests, including the entropy based test.

4.6 Conclusions

In this chapter, we extended Robinson's (1991a) entropy based test in various directions. Our test is a test for serial independence against serial dependence of any fixed and finite order, we allow for unbounded support, and we have derived a result for the case with nuisance parameters.

Unfortunately, the experiments of Section 4.5 indicate that the very convenient limiting distribution cannot be used for inference in moderate samples, and in that case, the practitioner will have to revert to simulated critical values.

Appendix

4.A Proofs of Theorems

Proof of Theorem 4.1

We prove the theorem in two steps. We first show that $\hat{\mathcal{I}} \xrightarrow{P} \mathcal{I}_{KL}$, and we then show that $\hat{\mathcal{V}}^* \xrightarrow{P} \mathcal{V}^* > 0$. This suffices, because if $C > 0$ denotes the critical value a practitioner intends to use, then $P[\hat{\tau} > C] = P[(\frac{N}{2J})^{\frac{1}{2}} \frac{\hat{\mathcal{I}}}{\sqrt{\hat{\mathcal{V}}^*}} > C] = P[\frac{\hat{\mathcal{I}}}{\sqrt{\hat{\mathcal{V}}^*}} > (\frac{N}{2J})^{-\frac{1}{2}} C] = P[\frac{\hat{\mathcal{I}}}{\sqrt{\hat{\mathcal{V}}^*}} - \frac{\mathcal{I}_{KL}}{\sqrt{\mathcal{V}^*}} > (\frac{N}{2J})^{-\frac{1}{2}} C - \frac{\mathcal{I}_{KL}}{\sqrt{\mathcal{V}^*}}]$. Under H_1, $\mathcal{I}_{KL} > 0$. Let N be so large that $(\frac{N}{2J})^{-\frac{1}{2}} C < \frac{\mathcal{I}_{KL}}{2\sqrt{\mathcal{V}^*}}$. Then the above probability is bounded by $P[\frac{\hat{\mathcal{I}}}{\sqrt{\hat{\mathcal{V}}^*}} - \frac{\mathcal{I}_{KL}}{\sqrt{\mathcal{V}^*}} > -\frac{\mathcal{I}_{KL}}{2\sqrt{\mathcal{V}^*}}] \geq P[|\frac{\hat{\mathcal{I}}}{\sqrt{\hat{\mathcal{V}}^*}} - \frac{\mathcal{I}_{KL}}{\sqrt{\mathcal{V}^*}}| < \frac{\mathcal{I}_{KL}}{2\sqrt{\mathcal{V}^*}}] \to 1$, by $\hat{\mathcal{I}} \xrightarrow{P} \mathcal{I}_{KL}$, $\hat{\mathcal{V}}^* \xrightarrow{P} \mathcal{V}^*$, and Slutsky's theorem.

We first prove that $\hat{\mathcal{I}} - \mathcal{I}_{KL} = o_p(1)$. Because $\hat{\mathcal{I}} = \hat{U}_J - \hat{U}_1$, and $\mathcal{I}_{KL} = E \log f_{J1} - JE \log f_1$, it suffices to show that $\hat{U}_a - U_a = o_p(1)$, and that $U_a - E \log f_{a1} = o_p(1)$, for $a = 1, J$. That the latter is true follows easily with the ergodic theorem; the presence of the weights is a minor nuisance, but because they may only take a finite number of different and bounded values, the ergodic theorem may be applied to each of the sets of observations with equal weights. For the former we need to show that $\frac{1}{N} \sum_{t \in S} c_{at} \log \hat{f}_{at} - \frac{1}{N} \sum_t c_{at} \log f_{at} = o_p(1)$, for $a = 1, J$, which is implied by $\frac{1}{N} \sum_{t \in S} c_{at} \{\log \hat{f}_{at} - \log f_{at}\} = o_p(1)$, and $P[1 \notin S] = o(1)$. Now, $P[1 \notin S] = P[||X_{J1}||_\infty \geq B] = 1 - P[||X_{J1}||_\infty \leq B] = 1 - P[|X_1| \leq B, |X_2| \leq B, \ldots, |X_J| \leq B] \leq 1 - P[X_1 \leq B, |X_2| \leq B, \ldots, |X_J| \leq B] + P[X_1 \leq -B] \leq 1 - P[X_1 \leq B, X_2 \leq B, |X_3| \leq B, \ldots, |X_J| \leq B] + F(-B) + P[X_2 \leq -B] \leq 1 - P[X_1 \leq B, \ldots, X_J \leq B] + JF(-B) = 1 - F_J.(B, \ldots, B) + JF(-B) = o(1)$, because $F(-x) \to 0$, as $x \to -\infty$, and $F_J.(x) \to 1$, as $\min_j x_j \to \infty$. It now remains to be shown that $\sup_{t \in S} |\log \hat{f}_{at} - \log f_{at}| = o_p(1)$. Note that $\log \hat{f}_{at} - \log f_{at} = \log\{1 + (\hat{f}_{at} - f_{at})/f_{at}\}$, and, $g(x) = \log(1 + x)$ being a continuous function, by Slutsky's theorem it thus suffices to

show that $\sup_{t\in S}|\hat{f}_{at}-f_{at}|/f_{at}=o_p(1)$. Because $t\in S$, $\sup_{t\in S}f_{at}^{-1}\leq L$. Further, Lemmas 4.1 and 4.2 imply that $\sup_{t\in S}|\hat{f}_{at}-f_{at}|=O_p(N^{-\frac{1}{2}}h^{-a}+h^{r+\omega})$; the fact that $\hat{f}$ and f now take random arguments is not relevant. Combining these last two results we have that $\sup_{t\in S}|\hat{f}_{at}-f_{at}|/f_{at}=O_p(N^{-\frac{1}{2}}h^{-a}L+h^{r+\omega}L)=o_p(1)$, by Assumption 4.D. An argument similar to the above leads to $\frac{1}{N}\sum_{t\in S}\log^2\hat{f}_t\xrightarrow{P}E\log^2 f_1$, such that $\hat{\mathcal{V}}\xrightarrow{P}\mathcal{V}$, with $\mathcal{V}=V\log f_1$, and hence also $\hat{\mathcal{V}}^*\xrightarrow{P}\mathcal{V}^*$, with $\mathcal{V}^*=\sum_{l=0}^{2J-1}\sum_{j=0}^{J-1}\tilde{c}_{4J-j-l}^2\mathcal{V}$.

Q.E.D.

Proof of Theorem 4.2

We prove Theorem 4.2 in two stages. In the first stage we show that $\sqrt{N}(\hat{\mathcal{I}}-\tilde{\mathcal{I}})=o_p(1)$, whilst the second step derives an asymptotic distribution for $\sqrt{N}\tilde{\mathcal{I}}$. For the first stage, it suffices to show that $\sqrt{N}(\hat{U}_a-U_a)=o_p(1)$, for $a=1,J$. Let $a=1,J$. Consider

$$N(\hat{U}_a-U_a)=\sum_{t\in S}c_{at}\log\frac{\hat{f}_{at}}{f_{at}}=\sum_{t\in S}c_{at}\frac{\hat{f}_{at}-f_{at}}{f_{at}}-\sum_{t\in S}c_{at}\left(\frac{\hat{f}_{at}-f_{at}}{\sqrt{2}\left(\frac{\hat{f}_{at}}{f_{at}};1\right)}\right)^2, \tag{4.19}$$

where $(x;y)$ denotes a number between x and y. Lemma 4.4 implies that the first term on the right hand side in (4.19) is $O_p(Nh^{r+\omega}L+N^{\frac{1}{2}}\lambda+Lh^{-\frac{a}{2}})$. Similarly, Lemma 4.5 shows that the second term on the right hand side in (4.19) is $O_p(h^{-a}L^2+Nh^{2(r+\omega)}L^2)$. Hence, $N^{\frac{1}{2}}(\hat{U}_a-U_a)=O_p(N^{\frac{1}{2}}h^{r+\omega}L+\lambda+N^{-\frac{1}{2}}h^{-a}L^2)=o_p(1)$, because $Nh^{2(r+\omega)}L^2\to 0$, and $Nh^{2a}L^{-4}\to\infty$, as $N\to\infty$, by Assumption 4.F.

Now, $N\tilde{\mathcal{I}}=\sum_{t\in S}(c_{Jt}-c_{1t})\log f_{Jt}$, where under the present conditions $f_{Jt}=\prod_{j=0}^{J-1}f_{t+j}$. Observe that

$$E\left(\sum_{t\notin S}(c_{Jt}-c_{1t})\log f_{Jt}\right)^2$$

$$= \sum_t \sum_s \tilde{c}_t \tilde{c}_s E[I(t \notin S) I(s \notin S) \log f_{Jt} \log f_{Js}]$$

$$= \sum_t \sum_{|s-t| \geq J} \tilde{c}_t \tilde{c}_s E^2[I(t \notin S) \log f_{Jt}] \tag{4.20}$$

$$+ \sum_t \sum_{|s-t| < J} \tilde{c}_t \tilde{c}_s E[I(t \notin S) I(s \notin S) \log f_{Jt} \log f_{Js}]. \tag{4.21}$$

Now, $E|I(1 \notin S) \log f_{J1}| = o(1)$, by the boundedness restriction on $E \log f_{J1}$. Hence, (4.20) is

$$o(1) \sum_t \sum_{|s-t| \geq J} \tilde{c}_t \tilde{c}_s = o(N) + o(1) \left(\sum_t \tilde{c}_t \right)^2 = o(N),$$

because $\sum_t \tilde{c}_t = \sum_t c_{Jt} - J \sum_t c_{1t} = O(1)$, by construction of c_{1t}, c_{Jt}. (4.21) is also $o(N)$, which is implied by $|E[I(t \notin S) I(s \notin S) \log f_{Jt} \log f_{Js}]| \leq E[I(1 \notin S) \log^2 f_{J1}] = o(1)$, and $\sum_t \sum_{|s-t|<J} 1 = O(N)$. Hence $\sqrt{N}\tilde{\mathcal{I}} = N^{-\frac{1}{2}} \sum_t \tilde{c}_t \log f_{Jt} + o_p(1)$.

Now, $\sum_t \tilde{c}_t \log f_{Jt}$ can be rewritten as $\sum_t \sum_{j=0}^{J-1} \tilde{c}_{t-j} \log f_t$, which in turn can be rewritten as

$$\sum_{t(2J)} \left(\sum_{l=0}^{2J-1} \sum_{j=0}^{J-1} \tilde{c}_{t-j-l} \right) \log f_{t-l}, \tag{4.22}$$

where $\sum_{t(2J)}$ lets t run over all multiples of $2J$. The expression in parentheses in (4.22) is, by construction of the $\tilde{c}_t$'s, equal to $\sum_{l=0}^{2J-1} \sum_{j=0}^{J-1} \tilde{c}_{4J-j-l}$, which does not depend upon t. Let A_t be the summand in (4.22), such that (4.22) can be written as $\sum_{t(2J)} A_t$, where the A_t's are i.i.d.. Because $EA_t = \sum_{l=0}^{2J-1} \sum_{j=0}^{J-1} \tilde{c}_{t-j-l} E \log f_1 = 0$,

$$\sqrt{\frac{N}{2J}} \tilde{\mathcal{I}} \xrightarrow{\mathcal{L}} N(0, VA_1),$$

by the Lindeberg-Levy central limit theorem, where $VA_1 = \sum_{l=0}^{2J-1} \sum_{j=0}^{J-1} \tilde{c}_{4J-j-l}^2 V \log f_1 = \mathcal{V}^*$, where $V \log f_1$ can be consistently estimated by $\hat{\mathcal{V}}$, as we have seen in the proof to Theorem 4.1.

Q.E.D.

Proof of Theorem 4.3

Analogous to the proof to Theorem 4.1, we only need to show that $\hat{\mathcal{I}}^Y \xrightarrow{P} \mathcal{I}_{KL}$, and that $\hat{\mathcal{V}}_Y^* \xrightarrow{P} \mathcal{V}^*$.

For the former, it suffices to show that $\hat{U}_a^Y - U_a^Y = o_p(1)$, and that $U_a^Y - U_a = o_p(1)$, for $a = 1, J$, as $\tilde{\mathcal{I}} \xrightarrow{P} \mathcal{I}_{KL}$ was already established in the proof to Theorem 4.1. Let $a = 1, J$. $\hat{U}_a^Y - U_a^Y = \frac{1}{N}\sum_{t\in S_Y} c_{at} \log \frac{\hat{f}_{a\cdot}^Y(Y_{Nat})}{f_{a\cdot}(Y_{Nat})} = \frac{1}{N}\sum_{t\in S_Y} c_{at} \log\{1 + \frac{\hat{f}_{a\cdot}^Y(Y_{Nat}) - f_{a\cdot}(Y_{Nat})}{f_{a\cdot}(Y_{Nat})}\}$. By Slutsky's theorem and by $\sup_{t\in S_Y} f_{a\cdot}^{-1}(Y_{Nat}) = O_p(L)$, it suffices to show that $\sup_t |\hat{f}_{a\cdot}^Y(Y_{Nat}) - f_{a\cdot}(Y_{Nat})| = o_p(L^{-1})$, which is implied by $\sup_t |\hat{f}_{a\cdot}^Y(Y_{Nat}) - \hat{f}_{a\cdot}(Y_{Nat})| = o_p(L^{-1})$ and $\sup_t |\hat{f}_{a\cdot}(Y_{Nat}) - f_{a\cdot}(Y_{Nat})| = o_p(L^{-1})$. The latter result follows directly from Lemmas 4.1 and 4.2, and Assumption 4.K. The former follows from Lemma 4.7.

That $U_a^Y - U_a = o_p(1)$, for $a = 1, J$, is implied by $\sum_{t\in S_Y} c_{at} \log f_{a\cdot}(Y_{Nat}) - \sum_{t\in S_X} c_{at} \log f_{a\cdot}(X_{at}) = o_p(N)$, which holds if

$$\sum_{t\in S_Y\setminus S_X} c_{at} \log f_{a\cdot}(Y_{Nat}) = o_p(N), \tag{4.23}$$

$$\sum_{t\in S_X\setminus S_Y} c_{at} \log f_{a\cdot}(X_{at}) = o_p(N), \tag{4.24}$$

$$\sum_{t\in S_X\cap S_Y} c_{at} \{\log f_{a\cdot}(Y_{Nat}) - \log f_{a\cdot}(X_{at})\} = o_p(N). \tag{4.25}$$

Conditions (4.23) and (4.24) can be treated in identical fashion: $P[t \in S_Y\setminus S_X] \leq P[t \notin S_X] = P[||X_{J1}||_\infty \geq B] = 1 - F_{J\cdot}(B_J) = O(\lambda)$, and hence $\sum_{t\in S_Y\setminus S_X} c_{at} \log f_{a\cdot}(Y_{Nat}) = O_p(N\lambda \log L) = o_p(N)$, because $\lambda \log L \to 0$, as $N \to \infty$, by Assumption 4.K. Now (4.25). Its absolute value is bounded by $\sum_{t\in S_X\cap S_Y} c_{at} |\log f_{a\cdot}(Y_{Nat}) - \log f_{a\cdot}(X_{at})| = \sum_{t\in S_X\cap S_Y} c_{at} |f'_{at}(Y_{Nat}; X_{at})(Y_{Nat} - X_{at})|$, by the mean value theorem. By Lemma 4.6, we can rewrite the afore expression as $\sum_{t\in S_X\cap S_Y} c_{at} \left\{|N^{-\frac{1}{2}} m'_{at}\zeta| + O_p(N^{-1}G)\right\} = O_p(N^{\frac{1}{2}}) + O_p(G) = o_p(N)$, by the ergodic theorem (and the existence of $E|m'_{a1}|$, and Assumption 4.K).

To establish that $\hat{\mathcal{V}}_Y^* - \mathcal{V}^* = o_p(1)$, it suffices to show that $\frac{1}{N}\sum_{t\in S_Y} \log^2 \hat{f}^Y(Y_{Nt}) - E\log^2 f_1 = o_p(1)$ and $\frac{1}{N}\sum_{t\in S_Y} \log \hat{f}^Y(Y_{Nt}) - E\log f_1 = o_p(1)$. The latter was already established earlier in this proof and in the proof to Theorem 4.1. For the former it suffices to show that $\frac{1}{N}\sum_{t\in S_Y} \log^2 \hat{f}^Y(Y_{Nt}) - \frac{1}{N}\sum_{t\in S_X} \log^2 \hat{f}(X_t) = o_p(1)$, as the proof to Theorem 4.1 already established that $\hat{\mathcal{V}}^* - \mathcal{V}^* = o_p(1)$. The remaining result is easily shown using a procedure analogous to the one used earlier in this proof to show that $\hat{U}_1^Y - \hat{U}_1 = o_p(1)$.

Q.E.D.

Proof of Theorem 4.4

In the proof to Theorem 4.2, we established an asymptotic distribution for $\sqrt{N}(\tilde{\mathcal{I}} - \mathcal{I}_{KL})$. In the proof to Theorem 4.3, we established that $\hat{\mathcal{V}}_Y^* \xrightarrow{P} \mathcal{V}^*$. We now need to show that $\sqrt{N}(\hat{\mathcal{I}}^Y - \tilde{\mathcal{I}}) = o_p(1)$, or that $\hat{U}_a^Y - U_a = o_p(N^{-\frac{1}{2}})$, for $a = 1, J$. Now, to establish the properties of $N(\hat{U}_a^Y - \hat{U}_a)$, we only need to examine

$$\begin{aligned}
&\sum_{t\in S_Y} c_{at} \log \hat{f}_{a\cdot}^Y(Y_{Nat}) - \sum_{t\in S_X} c_{at} \log f_{at} \\
&= \sum_{t\in S_Y \setminus S_X} c_{at} \log \hat{f}_{a\cdot}^Y(Y_{Nat}) && (4.26)\\
&+ \sum_{t\in S_Y \cap S_X} c_{at}\{\log \hat{f}_{a\cdot}^Y(Y_{Nat}) - \log f_{at}\} && (4.27)\\
&+ \sum_{t\in S_X \setminus S_Y} c_{at} \log f_{at}. && (4.28)
\end{aligned}$$

In Lemma 4.8, we establish that (4.26) is $o_p(N^{\frac{1}{2}})$. The mean value theorem can be used to expand (4.27); the first order term is dealt with in Lemma 4.10, and the second in Lemma 4.9. Finally, (4.28) is a direct consequence of the proof to Lemma 4.8.

Q.E.D.

4.B Technical Lemmas

4.B.1 Under the Mixing Condition

Let Assumptions 4.A–4.D hold. The following lemmas are not original, and can be found in a number of other places. Indeed, the first lemma also appears — albeit in a slightly different form — in Robinson (1987), Theorem 3.

Lemma 4.1 *For* $a = 1, J$, $\sup_x |\hat{f}_a.(x) - E\hat{f}_a.(x)| = O_p(N^{-\frac{1}{2}}h^{-a})$.

Proof:

Rewriting the above expression, and applying Assumption 4.C, we obtain

$$\begin{aligned}
&\sup_x \left| \frac{1}{Nh^a} \sum_s \{k_{ah}(x - X_s) - Ek_{ah}(x - X_s)\} \right| \\
&= \frac{1}{Nh^a} \sup_x \left| \int \phi_{ak}(u) \sum_s \{e^{iu^T(\frac{x-X_s}{h})} - Ee^{iu^T(\frac{x-X_s}{h})}\} du \right| \\
&= \frac{1}{N} \sup_x \left| \int \phi_{ak}(hv) \sum_s \{e^{iv^T(x-X_s)} - Ee^{iv^T(x-X_s)}\} dv \right| \\
&\leq \frac{1}{N} \int |\phi_{ak}(hv)| \sup_x |e^{iv^T x}| \left| \sum_s \{e^{-iv^T X_s} - Ee^{-iv^T X_s}\} \right| dv \\
&= \int |\phi_{ak}(hv)| \left| \frac{1}{N} \sum_s \{e^{-iv^T X_s} - Ee^{-iv^T X_s}\} \right| dv, \qquad (4.29)
\end{aligned}$$

where the second equality follows by subtitution of $v = u/h$, and the last equality by $\sup_x |e^{iux}| = 1$. (4.29) is non-negative, and it thus suffices to show that its expectation is $O(N^{-\frac{1}{2}}h^{-a})$. We first deal with the second factor under the integral sign in (4.29). Thus

$$\begin{aligned}
&\sup_v E \left| \frac{1}{N} \sum_s \{e^{-iv^T X_s} - Ee^{-iv^T X_s}\} \right| \\
&\leq \sup_v \sqrt{\frac{1}{N^2} \sum_t \sum_s E[\{e^{-iv^T X_s} - Ee^{-iv^T X_s}\}\{e^{iv^T X_t} - Ee^{iv^T X_t}\}]}
\end{aligned}$$

$$= \sup_v \sqrt{\frac{1}{N^2}\sum_t\sum_s \mathrm{Cov}\{\cos(v^T X_s) - i\sin(v^T X_s), \cos(v^T X_t) + i\sin(v^T X_t)\}}$$

$$= \sup_v \sqrt{\frac{1}{N^2}\sum_t\sum_s [\mathrm{Cov}\{\cos(v^T X_s), \cos(v^T X_t)\} + \mathrm{Cov}\{\sin(v^T X_s), \sin(v^T X_t)\}]}$$

$$\leq \sqrt{\frac{2}{N^2}\sum_s\sum_t \alpha(|s-t|)} = O\left(\sqrt{N^{-1}\sum_s \alpha(s)}\right) = O(N^{-\frac{1}{2}}), \tag{4.30}$$

by Assumption 4.A. Substituting (4.30) into (4.29) yields

$$O(N^{-\frac{1}{2}})\int |\phi_{ak}(hv)|dv = O(N^{-\frac{1}{2}}h^{-a})\int |\phi_{ak}(u)|du = O(N^{-\frac{1}{2}}h^{-a}),$$

by Assumption 4.C.

Q.E.D.

Lemma 4.2 *For* $a = 1, J$, $\sup_x |E\hat{f}_{a\cdot}(x) - f_{a\cdot}(x)| = O(h^{r+\omega})$.

Proof:

The left hand side in the above expression can be rewritten as

$$\sup_x \left|\frac{1}{h^a}Ek_{ah}(x - X_1) - f_{a\cdot}(x)\right| = \sup_x \left|\frac{1}{h^a}\int k_{ah}(x-y)f(y)dy - f(x)\right|$$

$$= \sup_x \left|\int k_{a\cdot}(u)\{f_{a\cdot}(x - hu) - f_{a\cdot}(x)\}du\right|$$

$$= \sup_x \left|\int_{||u||\geq \frac{\rho_{af}}{h}} k_{a\cdot}(u)\{f_{a\cdot}(x + hu) - f_{a\cdot}(x)\}du\right| \tag{4.31}$$

$$+ \sup_x \left|\int_{||u||< \frac{\rho_{af}}{h}} k_{a\cdot}(u)\{f_{a\cdot}(x + hu) - f_{a\cdot}(x)\}du\right|, \tag{4.32}$$

where ρ_{af} was implicitly defined in Assumption 4.B's reference to Definition 4.1. $f_{a\cdot}$ is bounded and hence (4.31) is (for some $C_1 > 0$) bounded by $C_1 \int_{||u||\geq \frac{\rho_{af}}{h}} |k_{a\cdot}(u)|du$, which in turn is bounded by $C_2 \int_{||u||\geq \frac{\rho_{af}}{h}} ||u||^{-r-1-\omega}du$, for some $C_2 > 0$, by Assumption 4.C. This last expression is just $C_2 h^{r+\omega}/\{(r+\omega)\rho_{af}^{r+\omega}\}$, and hence the first term in (4.32) is $O(h^{r+\omega})$. Now the second term in (4.32). Let $R_{fra}(x,y) = f_{a\cdot}(x) - f_{a\cdot}(y) - Q_{fra}(x,y)$, with Q_{fra} defined in Definition 4.1. Then the second term in (4.32) can be bounded by

$$\left|\int k_{a\cdot}(u)Q_{fra}(x - hu, x)du\right| + \left|\int k_{a\cdot}(u)R_{fra}(x - hu, x)du\right|du. \tag{4.33}$$

The first term in (4.33) is 0, by the assumption that $k_{a\cdot}$ is an $(r+1)$-th order kernel. The second term can (because $f_{a\cdot} \in \mathcal{G}_{r+\omega}$) be bounded by $h^{r+\omega} \int |k_{a\cdot}(u)|\, ||u||^{r+\omega} du = O(h^{r+\omega})$, by Assumption 4.C.

Q.E.D.

4.B.2 Under Independence

The following lemmas accompany Theorem 2, and hold under the conditions made therein.

Lemma 4.3 *The following conditions hold for* $a = 1, J$.

$$\sup_{||x||_\infty < B} \left| E\left[\frac{\frac{1}{h^a} k_{ah}(x - X_{a1}) - f_{a\cdot}(x)}{f_{a\cdot}(x)} \right] \right| = O(h^{r+\omega} L) \tag{4.34}$$

$$\begin{aligned} & E\left[\frac{\frac{1}{h^a} k_{ah}(x - X_{a1}) - f_{a\cdot}(X_{a1})}{f(X_{a1})} I(||X_{J1}||_\infty \leq B) \right] \\ &= K_{a\cdot}\left(\frac{B_a - x}{h} \right) - K_{a\cdot}\left(\frac{-B_a - x}{h} \right) - 1, \end{aligned} \tag{4.35}$$

$$\left| E\left[K_{a\cdot}\left(\frac{B_a - X_{a1}}{h} \right) - K_{a\cdot}\left(\frac{-B - X_{a1}}{h} \right) - 1 \right] \right| = O(\lambda). \tag{4.36}$$

Proof:

By Assumption 4.B, $1/f_{a\cdot}(x) \leq L$, whenever $||x||_\infty < B$. Lemma 4.2 then implies (4.34). Let $B_a \in \Re^a$ be a vector of B's. (4.35) follows from

$$\begin{aligned} E\left[\frac{k_{ah}(x - X_{a1})}{h^a f_{a\cdot}(X_{a1})} I(||X_{J1}||_\infty \leq B) \right] &= h^{-a} \int_{-B_a}^{B_a} k_{ah}(x - y) dy \\ &= \int_{\frac{-B_a - x}{h}}^{\frac{B_a - x}{h}} k_{a\cdot}(u) du \\ &= K_{a\cdot}\left(\frac{B_a - x}{h} \right) - K_{a\cdot}\left(\frac{-B_a - x}{h} \right). \end{aligned}$$

Now (4.36). We show that $|EK_{a\cdot}\{(B_a - X_{a1})/h\} - 1| = O(\lambda)$, where $|EK_{a\cdot}\{-(B_a - X_{a1})/h\}| = O(\lambda)$ can be shown in a similar fashion. Because $K_{a\cdot}(x) = \prod_{j=1}^a K(x_j)$, by Assumption 4.C, $K_{a\cdot}(x) - 1 = \sum_{l=1}^a \{K(x_l) - 1\} \prod_{j=l+1}^a K(x_j)$, and because K is bounded, it suffices to show

that $|EK\{(B-X_1)/h\}-1| = O(\lambda)$. Now,

$$\left|EK\left(\frac{B-X_1}{h}\right)-1\right|$$
$$\leq \int_{-\infty}^{(1-h)B}\left|K\left(\frac{B-x}{h}\right)-1\right|f(x)dx+\int_{(1-h)B}^{\infty}\left|K\left(\frac{B-x}{h}\right)-1\right|f(x)dx.$$

Assumption 4.C implies that $|K(B)-1|\to 0$, as $N\to\infty$. Hence, the first term in the last displayed equation can (for sufficiently large N) be bounded by $\int_{-\infty}^{(1-h)B}|K(\{B-(1-h)B\}/h)-1|f(x)dx = |K(B)-1|\int_{-\infty}^{(1-h)B}f(x)dx$, which is $O(\lambda)$, by definition. The second term in the last displayed equation can be bounded by $C\{1-F((1-h)B)\} = O(\lambda)$, by definition, where C is some large positive constant.

Q.E.D.

Lemma 4.4

We need to show that for $a = 1, J$,

$$\sum_{t\in S}\sum_{s\in S}c_tc_s\left(\frac{\hat{f}_{at}-f_{at}}{f_{at}}\right)\left(\frac{\hat{f}_{as}-f_{as}}{f_{as}}\right) = O_p(N^2h^{2(r+\omega)}L^2+N\lambda^2+L^2h^{-a}). \tag{4.37}$$

Proof:

Multiplying the left hand side in (4.37) by N^2, and taking the expectation implies that we only need to show that

$$\sum_t\sum_s c_tc_s\sum_{u\notin D_t}\sum_{v\notin D_s}E\left[\left\{\frac{\frac{1}{h^a}k_{ah}(X_{at}-X_{au})-f_{a\cdot}(X_{at})}{f_{a\cdot}(X_{at})}I(||X_{Jt}||_\infty\leq B)\right\}\right.$$
$$\left.\times\ \left\{\frac{\frac{1}{h^a}k_{ah}(X_{as}-X_{av})-f_{a\cdot}(X_{as})}{f_{a\cdot}(X_{as})}I(||X_{Js}||_\infty\leq B)\right\}\right]$$
$$= O(N^4h^{2(r+\omega)}L^2+N^3\lambda^2+N^2L^2h^{-J}). \tag{4.38}$$

We now examine three cases: when there is no overlap, i.e. when $t,u\notin D_s\cup D_v$, when there is a single overlap, and when there is at least a double overlap. Let $\Omega_\emptyset$, Ω_{ts}, and $\Omega_{ts,uv}$ denote the partial sums in (4.38), for which there is no overlap, overlap between t and s, and for which there is both overlap between t and s and between u and v, respectively, where Ω with other

subscripts are implicitly defined. When there is no overlap, we can condition on (X_{at}, X_{as}), and apply (4.34), such that $\Omega_{\emptyset} = O(N^4 h^{2(r+\omega)} L^2)$, which does not conflict with the right hand side in (4.38). If only $t \in D_s$, the same trick can be applied yielding $\Omega_{ts} = O(N^3 h^{2(r+\omega)} L^2)$. When $u \in D_v$, conditioning on (X_u, X_v) and applying both (4.35) and (4.36) yields $\Omega_{uv} = O(N^3 \lambda^2)$, and when $t \in D_v$ (or $s \in D_u$), a combination of both above techniques results in $\Omega_{tv}, \Omega_{su} = O(N^3 \lambda h^{r+\omega} L)$, which is $O(N^3 \lambda^2 + N^3 h^{2(r+\omega)} L^2)$. This does not conflict with the right hand side in (4.38) either, so we now only need to examine the case of more than one overlap. Notice that the expectation in (4.38) can be rewritten as

$$E\left[\frac{k_{ah}(X_{at} - X_{au}) k_{ah}(X_{as} - X_{av})}{h^{2a} f_{a\cdot}(X_{at}) f_{a\cdot}(X_{as})} I(||X_{Jt}||_\infty \leq B) I(||X_{Js}||_\infty \leq B)\right] + 1 \tag{4.39}$$

$$- \; 2E\left[\frac{k_{ah}(X_{at} - X_{au})}{h^a f_{a\cdot}(X_{at})} I(||X_{Jt}||_\infty \leq B)\right], \tag{4.40}$$

for all t, s, u, v. The expectation in (4.40) is $1 + O(h^{r+\omega} L)$, by (4.34), uniformly in $u \notin D_t$, $v \notin D_s$. So we only have to examine (4.39), for which

$$\sup_{t,s,u \notin D_t, v \notin D_s} \left| E\left[\frac{k_{ah}(X_{at} - X_{au}) k_{ah}(X_{as} - X_{av})}{h^{2a} f_{a\cdot}(X_{at}) f_{a\cdot}(X_{as})} I(||X_{Jt}||_\infty \leq B) I(||X_{Js}||_\infty \leq B)\right]\right|$$

$$\leq \; L^2 h^{-2a} \int k_{ah}^2(x - y) f_{a\cdot}(x) f_{a\cdot}(y) dx dy \leq C L^2 h^{-a} \int k_{a\cdot}^2(x) dx = O(L^2 h^{-a}),$$

where $C = \sup_x f_{a\cdot}(x)$; the first inequality follows from $1/f_{a\cdot}(X_t) < L$, for all $t \in S$, the second by substitution of x for $(x - y)/h$, and the equality by the squared integrability condition on k, imposed in Assumption 4.C. As there are only $O(N^2)$ terms where there is more than one overlap, the $\Omega_{..,..}$'s are $O(N^2 L^2 h^{-a})$.

Q.E.D.

Lemma 4.5 *For $a = 1, J$,*

$$\sum_{t \in S} \frac{(\hat{f}_{at} - f_{at})^2}{(\hat{f}_{at}^2; f_{at}^2)} = O_p(h^{-a} L^2 + N h^{2(r+\omega)} L^2), \tag{4.41}$$

where $(x; y)$ denotes some number between x and y.

Proof:

Note first that

$$\begin{aligned} L^2 \inf_{t\in S}(\hat{f}_{at}^2; f_{at}^2) &= L^2 \inf_{t\in S}\{f_{at}^2 + (\hat{f}_{at}^2 - f_{at}^2; 0)\} \\ &= L^2 \inf_{t\in S}\left\{f_{at}^2 + \left((\hat{f}_{at} - f_{at})^2 + 2(\hat{f}_{at} - f_{at})f_{at}; 0\right)\right\} \\ &\geq L^2 \inf_{t\in S} f_{at}^2 \left[1 - \left\{\sup_{t\in S}\left(\frac{\hat{f}_{at} - f_{at}}{f_{at}}\right)^2 + 2\sup_{t\in S}\frac{|\hat{f}_{at} - f_{at}|}{f_{at}}\right\}\right] \\ &\geq 1 + o_p(1), \end{aligned}$$

where the last inequality follows from the proof to Theorem 4.1, and Assumption 4.B. A consequence of the above result is that $\sup_{t\in S} 1/(\hat{f}_{at}^2; f_{at}^2) = O_p(L^2)$. Thus, the left hand side in (4.41) is $O_p(L^2)\sum_{t\in S}(\hat{f}_{at} - f_{at})^2$. The summation in the afore expression is always non-negative, so it suffices to show that $\sup_t E[(\hat{f}_{at} - f_{at})I(t \in S)]^2 \leq E[\hat{f}_{a1} - f_{a1}]^2 = O(N^{-1}h^{-a} + h^{2(r+\omega)})$, for (4.41) to hold. The inequality in the above expression is a direct consequence from the stationarity condition imposed in Assumption 4.E, and that for any event A and any positive number c, $cI(A) \leq c$. Now

$$\begin{aligned} & N^2 E[\hat{f}_{a1} - f_{a1}]^2 \\ &= \sum_{s\notin D_1}\sum_{u\notin D_1}\frac{1}{h^{2a}}E[\{k_{ah}(X_{a1} - X_{as}) - h^a f_{a\cdot}(X_{a1})\}\{k_{ah}(X_{a1} - X_{au}) - h^a f_{a\cdot}(X_{a1})\}] \\ &= O(N^2 h^{2(r+\omega)} + Nh^{-a}), \end{aligned} \tag{4.42}$$

because if $s \notin D_u$, then X_{as} and X_{au} are independent, and we can condition on X_{a1} and apply Lemma 4.2 which leads to the first convergence rate stated in (4.42); if $s \in D_u$ then the expectation on the right is bounded by $E[k_{ah}^2(X_{a1} - X_{as})] = \int k_{ah}^2(x - y)f_{a\cdot}(x)f_{a\cdot}(y)dxdy = O(h^a)$, by substitution for $(x - y)/h$, and because there are only $O(N)$ combinations of s and u, such that $s \in D_u$, the second convergence rate stated in (4.42) is achieved.

Q.E.D.

4.B.3 Nuisance Parameters

4.B.3.1 Under the Mixing Condition

Lemma 4.6

$$Y_{Nat} - X_{at} = N^{-\frac{1}{2}} m'_{at}\zeta + N^{-1}\psi_{at}, \tag{4.43}$$

uniformly in t, where $\psi_{at} = \frac{1}{2}\left(\sum_{j=1}^{a}\zeta_j D_j\right)^2 m_{at|(\hat{\theta};\theta_0)} = O_p(G)$, *uniformly in t.*

Proof:

We assumed in Assumption 4.G that $Y_{Nat} = m_{at}(\hat{\theta})$ and $X_{at} = m_{at}(\theta_0)$. With the mean value theorem it follows that

$$Y_{Nat} - X_{at} = m'_{at}(\hat{\theta} - \theta_0) + \frac{1}{2}\left(\sum_{j=1}^{a}(\hat{\theta}_j - \theta_{0j})D_j\right)^2 m_{at}(\hat{\theta};\theta_0). \tag{4.44}$$

Because $\hat{\theta} - \theta_0 = N^{-\frac{1}{2}}\zeta$, by Assumption 4.I, the first term on the right hand side in (4.44) is $N^{-\frac{1}{2}}m'_{at}\zeta$, and the the second term is $\frac{1}{2N}\left(\sum_{j=1}^{a}\zeta_j D_j\right)^2 m_{at}(\hat{\theta};\theta_0)$. The proof can thus be concluded by noting that $\sup_{t;\theta\in\Theta}||m''_t(\theta)||_\infty = O_p(G)$, by Assumption 4.J, and that $\zeta = O_p(1)$, by Assumption 4.I.

Q.E.D.

Lemma 4.7 *For* $a = 1, J$,

$$\sup_t |\hat{f}^Y_{a\cdot}(Y_{Nat}) - \hat{f}_{a\cdot}(Y_{Nat})| = o_p(L^{-1}). \tag{4.45}$$

Proof:

The left hand side in (4.45) can be rewritten as

$$\begin{aligned}
&\sup_t \left| \frac{1}{Nh^a} \sum_{s\notin D_t} \left\{ k_{a\cdot}\left(\frac{Y_{Nat} - Y_{Nas}}{h}\right) - k_{a\cdot}\left(\frac{Y_{Nat} - X_{as}}{h}\right)\right\}\right| \\
&= \sup_t \left| \frac{1}{Nh^a} \sum_{s\notin D_t} \int \left\{ \phi_{ak}(u)\left(e^{iu^T\frac{Y_{Nat}-Y_{Nas}}{h}} - e^{iu^T\frac{Y_{Nat}-X_{as}}{h}}\right)\right\}\right| \\
&\leq \int |\phi_{ak}(hv)| \left| \frac{1}{N}\sum_s \left(e^{-iv^T Y_{Nas}} - e^{-iv^T X_{as}}\right)\right| dv,
\end{aligned} \tag{4.46}$$

where the steps that led to the inequality are similar to those in the proof to Lemma 4.1. We have now included the set D_t in the summation for notational convenience; we can safely do this as there are only a finite number of terms in that set, anyway. The expression whose mode is taken in (4.46) can by the mean value theorem be written as

$$\left| v^T \frac{1}{N} \sum_s (Y_{Nas} - X_{as}) e^{-iv^T (Y_{Nas}; X_{as})} \right|. \tag{4.47}$$

From Lemma 4.6, we know that $Y_{Nas} - X_{as} = N^{-\frac{1}{2}} m'_{as}\zeta + N^{-1}\psi_{as}$, uniformly in s, such that (4.47) is

$$N^{-\frac{1}{2}} \left| v^T \frac{1}{N} \sum_s m'_{as}\zeta e^{-iv^T (Y_{Nas}; X_{as})} \right| + |v^T \iota| \frac{1}{N^2} \sum_s \left| e^{-iv^T (Y_{Nas}; X_{as})} \right| |\psi_{as}|, \tag{4.48}$$

where ι is a vector of 1's. (4.48) can be bounded by

$$N^{-\frac{3}{2}} \sum_s |v^T m'_{as}\zeta| + O_p(N^{-1}G)|v^T \iota|, \tag{4.49}$$

because $\sup_s |\psi_{as}| = O_p(G)$, by Lemma 4.6. Note that $\int |\phi_{ak}(hv)v^T| dv = h^{-a} \int |\phi_{ak}(u)u^T| du = O(h^{-a})$, by substitution of $u = hv$, and Assumption 4.C. Also note that $\frac{1}{N} \sum_s |m'_{as}\zeta| \xrightarrow{P} E|m'_{a1}\zeta|$, by the ergodic theorem; the existence of the expectation in the afore convergence result is implied by Assumption 4.J. Substituting (4.49) into (4.46) yields therefore

$$\begin{aligned} & O_p(N^{-\frac{1}{2}}h^{-a}) \int |\phi_{ak}(u)u^T| dv \frac{1}{N} \sum_s |m'_{as}\zeta| + O_p(N^{-1}Gh^{-a}) \int |\phi_{ak}(u)u^T| dv\iota \\ = \; & O_p(N^{-\frac{1}{2}}h^{-a} + N^{-1}Gh^{-a}) = o_p(L^{-1}), \end{aligned}$$

if $N^{-\frac{1}{2}}h^{-J}L \to 0$ and $N^{-1}Gh^{-J}L \to 0$, as $N \to \infty$, which is implied by $Nh^{2J}L^{-2} \to \infty$ and $NG^{-1}h^J L^{-1} \to \infty$, which are implied by Assumption 4.K.

Q.E.D.

4.B.3.2 Under Serial Independence

Lemma 4.8 *For* $a = 1, J$, $\sum_{t \in S_Y \setminus S_X} c_{at} \log \hat{f}^Y_a(Y_{Nat}) = o_p(N^{\frac{1}{2}})$.

Proof:

Let $S_A(\Upsilon) = \{t : \ ||Y_{NJt}||_\infty \leq B,\ B < ||X_{Jt}||_\infty \leq B + \Upsilon\}$, $S_B(\Upsilon) = \{t : \ ||Y_{NJt}||_\infty \leq B,\ ||X_{Jt}||_\infty > B + \Upsilon\}$, such that $S_Y \backslash S_X = S_A(\Upsilon) \cup S_B(\Upsilon)$, for any sequence Υ. Now,

$$\sup_{t \in S_Y} |\log \hat{f}^Y_{a\cdot}(Y_{Nat})| \leq \sup_{t \in S_Y} |\log \hat{f}^Y_{a\cdot}(Y_{Nat}) - \log f_{a\cdot}(Y_{Nat})| + \sup_{t \in S_Y} |\log f_{a\cdot}(Y_{Nat})|. \quad (4.50)$$

The second term on the right hand side in (4.50) is (for sufficiently large N) bounded by $\log L$, which is a consequence of $t \in S_Y$ and Assumption 4.B. The first term on the right hand side in (4.50) is $o_p(1)$, which is implied by $\sup_{t \in S_Y} |\log\{1 + \frac{\hat{f}^Y_{a\cdot}(Y_{Nat}) - f_{a\cdot}(Y_{Nat})}{f_{a\cdot}(Y_{Nat})}\}| = o_p(1)$ or by $\sup_{t \in S_Y} |\hat{f}^Y_{a\cdot}(Y_{Nat}) - f_{a\cdot}(Y_{Nat})| = o_p(L^{-1})$. In Lemma 4.7, we established that $\sup_{t \in S_Y} |\hat{f}^Y_{a\cdot}(Y_{Nat}) - \hat{f}_{a\cdot}(Y_{Nat})| = o_p(L^{-1})$, and Lemmas 4.1 and 4.2 ensure that $\sup_{t \in S_Y} |\hat{f}_{a\cdot}(Y_{Nat}) - f_{a\cdot}(Y_{Nat})| = o_p(L^{-1})$. Hence, for sufficiently large N, the left hand side in (4.50) is bounded by $2 \log L$, and hence $|\sum_{t \in S_Y \backslash S_X} c_{at} \log \hat{f}^Y_{a\cdot}(Y_{Nat})|$ is bounded by $\log L \sum_{t \in S_Y \backslash S_X} C$, for some large $C > 0$, because the c_{at}'s are bounded. We thus only need to show that $\sum_{t \in S_Y \backslash S_X} 1 = o_p(N^{\frac{1}{2}} \log^{-1} L)$, or equivalently that $\sum_{t \in S_A(\Upsilon)} 1 = o_p(N^{\frac{1}{2}} \log^{-1} L)$ and that $\sum_{t \in S_B(\Upsilon)} 1 = o_p(N^{\frac{1}{2}} \log^{-1} L)$. Consider first the latter. Choose some $C > 0$. $P[\sum_t I(t \in S_B(\Upsilon)) > C^{-1} N^{\frac{1}{2}} \log^{-1} L]$ is by Markov's inequality bounded by $CN^{\frac{1}{2}} \log L\ EI(1 \in S_B(\Upsilon)) = CN^{\frac{1}{2}} \log L\ P[1 \in S_B(\Upsilon)] \leq CN^{\frac{1}{2}} \log L\ P[||Y_{NJ1} - X_{J1}||_\infty > \Upsilon] = CN^{\frac{1}{2}} \log L\ P[\min\{||Y_{NJ1} - X_{J1}||, 1\} > \Upsilon] \leq CN^{\frac{1}{2}} \log L \Upsilon^{-q} E[\min\{||Y_{NJ1} - X_{J1}||_\infty, 1\}]^q = O(N^{-\frac{1}{2}(q-1)} \log L \Upsilon^{-q}) = o(1)$, by again Markov's inequality and by Assumptions 4.I and 4.O. Finally, consider $P[\sum_t I(t \in S_A(\Upsilon)) > C^{-1} N^{\frac{1}{2}} \log^{-1} L]$, which, by an argument analogous to that above, is bounded by $CN^{\frac{1}{2}} \log L\ P[1 \in S_A(\Upsilon)]$. So, we need to show that $P[1 \in S_A(\Upsilon)] = o(N^{-\frac{1}{2}} \log^{-1} L)$. Now, $P[1 \in S_A(\Upsilon)] \leq P[||X_{J1}||_\infty \in [B, B + \Upsilon)] \leq P[|X_1| \in [B, B + \Upsilon) \vee \cdots \vee |X_J| \in [B, B + \Upsilon)] = JP[|X_1| \in [B, B + \Upsilon)] = J\{F(B + \Upsilon) - F(B) + F(-B) - F(-B - \Upsilon)\} = J\Upsilon\{f(B; B + \Upsilon) + f(-B - \Upsilon; -B)\} = O(\Upsilon L^{-\frac{1}{\Upsilon}}) = o(N^{-\frac{1}{2}} \log^{-1} L)$, where the second but last equality follows with the mean value theorem, the penultimate equality from Assumption 4.N, and the final one from Assumption 4.O.

Q.E.D.

Lemma 4.9 *For* $a = 1, J$,

$$\sum_{t \in S_X \cap S_Y} c_{at} \frac{\{\hat{f}^Y_{a\cdot}(Y_{Nat}) - f_{at}\}^2}{\{\hat{f}^Y_{a\cdot}(Y_{Nat}); f_{at}\}^2} = o_p(N^{\frac{1}{2}}). \tag{4.51}$$

Proof:

The denominator term in (4.51) can be eliminated in a similar manner as in the proof to Lemma 4.5. Indeed, $\sup_{t \in S_X \cap S_Y} |\{\hat{f}^Y_{a\cdot}(Y_{Nat}); f_{at}\}^{-2}| = O_p(L^2)$. Hence, for (4.51) to hold, it suffices to show that $\sum_t \{\hat{f}^Y_{a\cdot}(Y_{Nat}) - f_{at}\}^2 = o_p(N^{\frac{1}{2}}L^{-2})$, because $\sup_t |c_{at}| < \infty$. Now, $\sum_t \{\hat{f}^Y_{a\cdot}(Y_{Nat}) - f_{at}\}^2 \leq \sum_t \{\hat{f}^Y_{a\cdot}(Y_{Nat}) - \hat{f}_{at}\}^2 + \sum_t \{\hat{f}_{at} - f_{at}\}^2$. In Lemma 4.5, we dealt with the second term in the afore expansion. Now the first term. Let $\sum_{tsu}$ denote $\sum_t \sum_{s \notin D_t} \sum_{u \notin D_t}$.

$$\begin{aligned}
& N^2 h^{2a+2} \sum_t \{\hat{f}^Y_{a\cdot}(Y_{Nat}) - \hat{f}_{at}\}^2 \\
&= h^2 \sum_{tsu} \left\{ k_{a\cdot}\left(\frac{Y_{Nat} - Y_{Nas}}{h}\right) - k_{a\cdot}\left(\frac{X_{at} - X_{as}}{h}\right) \right\} \left\{ k_{a\cdot}\left(\frac{Y_{Nat} - Y_{Nau}}{h}\right) - k_{a\cdot}\left(\frac{X_{at} - X_{au}}{h}\right) \right\} \\
&\leq h^2 \sum_{tsu} \left| k_{a\cdot}\left(\frac{Y_{Nat} - Y_{Nas}}{h}\right) - k_{a\cdot}\left(\frac{X_{at} - X_{as}}{h}\right) \right| \left| k_{a\cdot}\left(\frac{Y_{Nat} - Y_{Nau}}{h}\right) - k_{a\cdot}\left(\frac{X_{at} - X_{au}}{h}\right) \right| \\
&= \sum_{tsu} |k'_{a\cdot}(\cdot)| \, |(Y_{Nat} - X_{at} + X_{as} - Y_{Nas})| \, |k'_{a\cdot}(\cdot)| \, |(Y_{Nat} - X_{at} + X_{au} - Y_{Nau})|,
\end{aligned}$$

by the mean value theorem. Because $|k'_{a\cdot}|$ is bounded, it can further be ignored. Let $\sum_{tsu}$ now denote $\sum_t \sum_s \sum_u$. Then the last displayed expression can be bounded by a large positive constant times

$$N^{-1} \sum_{tsu} \{|\zeta^T(m'_{at} - m'_{as})|\iota + N^{-\frac{1}{2}}(\psi_{at} - \psi_{as})\} \{|\zeta^T(m'_{at} - m'_{au})|\iota + N^{-\frac{1}{2}}(\psi_{at} - \psi_{au})\}, \tag{4.52}$$

which follows from Lemma 4.6, where ι is a vector of 1's. $E[N^{-3}\sum_{tsu} |\zeta^T(m'_{at} - m'_{as})|\iota|\zeta^T(m'_{at} - m'_{au})|\iota] \leq N^{-3}\sum_{tsu}\{E[|\zeta^T(m'_{at} - m'_{as})|\iota]^2 E[|\zeta^T(m'_{at} - m'_{au})|\iota]^2\}^{\frac{1}{2}} = O(1)$, by Assumption 4.M. Because $\sup_t |\psi_{at}| = O_p(G)$, by Lemma 4.6, (4.52) is $O_p(N^2 + NG^2)$, such that $N^{-1}\sum_t \{\hat{f}^Y_{a\cdot}(Y_{Nat}) - \hat{f}_{at}\}^2 = O_p(N^{-3}h^{-2-2a}(N^{-1}N^3 + N^3N^{-2}G^2)) = O_p(N^{-1}h^{-2-2J} + N^{-2}h^{-2-2J}G^2) = o_p(N^{-\frac{1}{2}}L^{-2})$, because $Nh^{4+4J}L^{-4} \to \infty$ and $Nh^{\frac{4}{3}+\frac{4}{3}J}G^{-\frac{4}{3}}L^{-\frac{4}{3}} \to \infty$, by

Assumption 4.O.

Q.E.D.

Lemma 4.10 *For* $a = 1, J$,

$$\frac{1}{N^2 h^a} \sum_t \sum_{s \notin D_t} \frac{c_{at}\Omega_t}{f_{at}} \left\{ k_{a\cdot} \left(\frac{Y_{Nat} - Y_{Nas}}{h} \right) - k_{a\cdot} \left(\frac{X_{at} - X_{as}}{h} \right) \right\} = o_p(N^{-\frac{1}{2}}), \tag{4.53}$$

where $\Omega_t = I(||X_{Jt}||_\infty < B)$.

Proof:

By Lemma 4.6,

$$Y_{Nat} - X_{at} = N^{-\frac{1}{2}} \begin{bmatrix} m_t'\zeta \\ \vdots \\ m_{t+a-1}'\zeta \end{bmatrix} + N^{-1}\psi_{at},$$

and hence the left hand side in (4.53) mutliplied by $N^2 h^a$ can, by the mean value theorem and Lemma 4.6, be written as

$$\begin{aligned}
& h^{-1} \sum_{ts} \frac{c_{at}\Omega_t}{f_{at}} \\
& \times \sum_{j=1}^{a} \{N^{-\frac{1}{2}}(m_{t+j-1}' - m_{s+j-1}')\zeta + N^{-1}(\psi_{at} - \psi_{as})\} k'_{t+j-1,s+j-1} \\
& \times \prod_{l \neq j} k_{t+l-1,s+l-1} \qquad (4.54) \\
& + \sum_{i=2}^{r-1} \sum_{ts} \frac{c_{at}\Omega_t}{h^i i! f_{at}} \\
& \times \left(\sum_{j=1}^{a} \{N^{-\frac{1}{2}}(m_{t+j-1}' - m_{s+j-1}')\zeta + N^{-1}(\psi_{at} - \psi_{as})\} D_j \right)^i k_{a\cdot|(\frac{X_{at} - X_{ss}}{h})} \qquad (4.55) \\
& + \sum_{ts} \frac{c_{at}\Omega_t}{h^r r! f_{at}} \\
& \times \left(\sum_{j=1}^{a} \{N^{-\frac{1}{2}}(m_{t+j-1}' - m_{s+j-1}')\zeta + N^{-1}(\psi_{t+j-1} - \psi_{s+j-1})\} D_j \right)^r k_{a\cdot|\cdot}, \qquad (4.56)
\end{aligned}$$

where $\sum_{ts} = \sum_t \sum_{s \notin D_t}$. Consider (4.54). Select some $j = 1, \ldots, a$. We have to establish that

$$\sum_{ts} c_{at}\Omega_t(m'_{t+j-1} - m'_{s+j-1})\zeta \frac{k'_{t+j-1,s+j-1}}{f_{t+j-1}} \prod_{l \neq j} \frac{k_{t+l-1,s+l-1}}{f_{t+l-1}} = o_p(N^2h^{a+1}), \quad (4.57)$$

$$\sum_{ts} c_{at}\Omega_t \frac{k'_{t+j-1,s+j-1}}{f_{t+j-1}} \prod_{l \neq j} \frac{k_{t+l-1,s+l-1}}{f_{t+l-1}}(\psi_{t+j-1} - \psi_{s+j-1}) = o_p(N^{\frac{5}{2}}h^{a+1}). \quad (4.58)$$

The left hand side in (4.58) is $O_p(N^2GL)$, because $\sup_t |\psi_t| = O_p(G)$, and because $\Omega_t f_{at}^{-1} \leq L$. This implies that (4.58) holds, because $N^2GLN^{-\frac{5}{2}}h^{-a-1} = N^{-\frac{1}{2}}GLh^{-a-1} \to 0$, as $N \to \infty$, because $Nh^{2J+2}G^{-2}L^{-2} \to \infty$, as $N \to \infty$. Now (4.57). Note that m'_t may depend upon past X_s's, although it was assumed in Assumption 4.L that m'_t is independent of X_s, for $s \geq t$. Let $\pi_t = I(|X_t| \leq B)$, such that $\Omega_t = \prod_{j=1}^{J} \pi_{t+j-1}$. expectation of the left hand side in (4.57) is

$$\sum_{ts} c_{at} E\left[\frac{k'_{t+j-1,s+j-1}}{f_{t+j-1,s+j-1}}\pi_{t+j-1}\right] E\left[\{m'_{t+j-1} - m'_{s+j-1}\}\zeta \prod_{l \neq j} \frac{\pi_{t+l-1}k_{t+l-1,s+l-1}}{f_{t+l-1}}\right]$$

which is zero because $E[f_t^{-1}k'_{ts}\pi_t] = E[f_t^{-1}\pi_t E[k'_{ts}|X_t]] = 0$, as established earlier. Hence the expectation of the left hand side in (4.57) is zero. Its squared expectation has N^3 cross product terms, each bounded by the expectation of its squared summand, which is $O(L^2h^a)$, because $f_t^{-2}\pi_t \leq L^2$, and $E[\{(m'_{t+j-1} - m'_{s+j-1})\zeta\}^2(k'_{t+j-1,s+j-1})^2 \prod_{l \neq j} k^2_{t+l-1,s+l-1}] = O(h^a)$, by repeated substitution if the above expectation is written as an integral, where the expectation exists because of Assumptions 4.C and 4.M. Hence the squared expectation of the left hand side of (4.57) is $O(N^3L^2h^a)$, and therefore the left hand side in (4.57) is itself $O_p(N^{\frac{3}{2}}Lh^{\frac{a}{2}})$. By Assumption 4.I, $N^{\frac{3}{2}}Lh^{\frac{a}{2}}(N^2h^{a+1})^{-1} = N^{-\frac{1}{2}}h^{-\frac{a}{2}-1}L \to 0$, as $N \to \infty$, because $Nh^{J+2}L^{-2} \to \infty$, as $N \to \infty$, by Assumption 4.O. Now (4.56). Because k_a. was assumed to be r times boundedly differentiable in every direction, it suffices to show that

$$\sum_{ts} \frac{c_{at}\Omega_t}{f_{at}} \sum_{l=0}^{n} \frac{r!}{l!(r-l)!} |N^{-\frac{1}{2}} \sum_{j=1}^{a}(m'_{t+j-1} - m'_{s+j-1})\zeta|^l |N^{-1} \sum_{j=1}^{a}(\psi_{t+j-1} - \psi_{s+j-1}))|^{r-l}$$
$$= o_p(N^2h^{a+r}),$$

or equivalently that for any $l = 0, \ldots, r$,

$$\sum_{ts} \left| \sum_{j=1}^{a} (m'_{t+j-1}\zeta - m'_{s+j-1}\zeta) \right|^r = o_p(N^{r+2-\frac{l}{2}} G^{l-r} L^{-1} h^{a+r}), \tag{4.59}$$

because $\Omega_t / f_{at} \leq L$, by Assumption 4.B, and because $\sup_t |\psi_t| = O_p(G)$, by Lemma 4.6. The left hand side in (4.59) is bounded by $\sum_{ts} (|\sum_{j=1}^{a} m'_{t+j-1}\zeta| + |\sum_{j=1}^{a} m'_{s+j-1}\zeta|)^r \leq 2C \sum_{ts} \sum_{j=1}^{a} |m'_{t+j-1}\zeta|^r \leq 2NC \sum_t \sum_{j=1}^{a} ||m'_{t+j-1}||^r ||\zeta||^r$, which is $O_p(N^2)$, because $\zeta = O_p(1)$, by Assumption 4.I, and because of the ergodic theorem and the fact that $E||m'_1||^r < \infty$, by Assumption 4.M. Hence, the left hand side in (4.59) is $O_p(N^2)$, which is indeed $o_p(N^{2+r-\frac{l}{2}} G^{r-l} L^{-1} h^{a+r})$, for all $l = 0, \ldots, r$, because $N^{-\frac{n}{2}} h^{-J-n} L \to 0$ and $N^{-n} G^n h^{-J-n} L \to 0$, as $N \to \infty$, because $N h^{2+\frac{2J}{r}} L^{-\frac{2}{r}} \to \infty$ and $N G^{-1} h^{1+\frac{J}{r}} L^{-\frac{1}{r}} \to \infty$, as $N \to \infty$, by Assumption 4.O. Finally, examine (4.55). Select some $i = 2, \ldots, r-1$. We can expand the expression that is raised to the power i in (4.55) into a finite summation indexed by $l = 0, \ldots, i$, with as summands a finite constant times

$$\begin{aligned} & O_p(N^{\frac{1}{2}-i}) \left(\sum_{j=1}^{a} (m'_{t+j-1} - m'_{s+j-1}) \zeta D_j \right)^l \left(\sum_{j=1}^{a} (\psi_{t+j-1} - \psi_{s+j-1}) D_j \right)^{i-l} \\ \leq \; & O_p(N^{\frac{1}{2}-i} G^{i-l}) \left| \sum_{j=1}^{a} (m'_{t+j-1} - m'_{s+j-1}) \zeta D_j \right|^l \left(\sum_{j=1}^{a} D_j \right)^{i-l}, \end{aligned} \tag{4.60}$$

because $\sup_t |\psi_t| = O_p(G)$, by Lemma 4.6. Let $\nu, \varrho \in \Re^a$. Then, using the notation of Appendix A.2, we can write (4.60) as

$$O_p(N^{\frac{1}{2}-i} G^{i-l}) \sum_{\nu \in V_{al}} \sum_{\varrho \in V_{a,i-l}} \Gamma_\nu \Gamma_\varrho |(m'_{at} - m'_{as})\zeta|^\nu D^{\nu+\varrho}, \tag{4.61}$$

where the Γ's are finite constants and the summations run over finite sets. Note that $|(m'_{at} - m'_{as})\zeta|^\nu = \prod_{j=1}^{a} |(m'_{t+j-1} - m'_{s+j-1})\zeta|^{\nu_j} \leq ||\zeta||^l \prod_{j=1}^{a} ||m'_{t+j} - m'_{s+j}||^{\nu_j}$, that $||\zeta|| = O_p(1)$, and that $|D^{\nu+\varrho} k_{a \cdot |(\frac{X_{at} - X_{as}}{h})}| = |\prod_{j=1}^{a} k^{(\nu_j+\varrho_j)}(\frac{X_{t+j-1} - X_{s+j-1}}{h})|$. Combining (4.55), (4.60) and (4.61) implies that we need to establish that

$$\frac{N^{\frac{1}{2}-i-2} G^{i-l}}{h^{i+a}} \sum_{ts} \frac{c_{at}\Omega_t}{f_{at}} |(m'_{at} - m'_{as})\zeta|^\nu \left| \prod_{j=1}^{a} k^{(\nu_j+\varrho_j)} \left(\frac{X_{t+j-1} - X_{s+j-1}}{h} \right) \right| = o_p(N^{-\frac{1}{2}}), \tag{4.62}$$

for every combination of i, l, ν, ϱ. We now intend to take the expectation of the left hand side in the last displayed equation. Note first that $c_{at}\Omega_t f_{at}^{-1} \leq CL$, for some large C. By the inequality of Cauchy-Schwarz, the remainder of the summand can be bounded by the square root of $E|(m'_{at} - m'_{as})\zeta|^{2\nu} \prod_{j=1}^{a} E|k^{(\nu_j+\varrho_j)}(\frac{X_{t+j-1}-X_{s+j-1}}{h})| = O(h^a)$, because $E|m'_t\zeta|^{2(r-1)} < \infty$, by Assumption 4.M, such that (4.62) holds if $N^{\frac{l+1}{2}-i}G^{i-l}h^{-i-\frac{a}{2}} \to 0$, as $N \to \infty$, for all $i = 2, \ldots, r-1$, and all $l = 0, \ldots, i$. We need to verify that $N^{-\frac{3}{2}}G^2h^{-2-\frac{J}{2}}L \to 0$, and $N^{-\frac{1}{2}}h^{-2-\frac{J}{2}}L \to 0$, as $N \to \infty$; both hold because $NG^{-\frac{4}{3}}h^{\frac{4+J}{3}}L^{-\frac{2}{3}} \to \infty$, $Nh^{4+J}L^{-2} \to \infty$, as $N \to \infty$, by Assumption 4.O.

Q.E.D.

Chapter 5

A General Characteristic Function Based Measure Applied to Serial Independence Testing

5.1 Introduction

In this chapter we suggest a measure for the distance between two distributions, we indicate how it can be estimated and we apply it to testing for serial independence against a serial dependence of order one alternative. We also suggest a statistic for serial independence against serial dependence of order $J-1$ ($2 \leq J < \infty$). Some other possible applications of the measure are in testing for structural breaks, for time series reversibility, for Gaussianity, and for the equivalence of the distributions of elements in separate stationary series. We discuss such applications in Section 5.8.

Many measures, that can be used for similar purposes, exist and, as we have seen in Chapter 3, quite a few of them have been applied to serial independence testing. There are however a number of characteristics that set the proposed test apart. An advantage is that it can be used for continuous, discrete and mixed distributions alike. As kernel density estimate based statistics, a kernel needs to be chosen, but not a sample size dependent bandwidth sequence. Although the fact that no bandwidth sequence needs to be chosen is fortunate (bandwidths have been found to strongly influence estimation results), the choice of a kernel does imply a degree of arbitrariness. In Section 5.6, we discuss its choice. Under the alternative hypothesis, as is usually the case in nonparametric independence testing, a weak dependence condition is required in addition to serial dependence of order one. We say a bit more about this condition in Subsection 5.3.1. Under the null hypothesis, the proposed test is shown to have a limiting χ_1^2 distribution. In Section 5.5, Monte Carlo simulations suggest that the small sample distribution of the statistic is quite close to the asymptotic distribution, even for samples as small as 100 observations. As we shall see in Section 5.4, the proposed test is generally more powerful than the correlation dimension test of Brock, Dechert and Scheinkman (1987). It is also fairly easy to compute.

The measure we propose bears some similarities to Csörgö's (1985) test for cross-sectional independence, but there are, as we shall see, some major differences, also.

In Section 5.2 we propose our characteristic function based measure. In Section 5.3 we explain how, along the lines set out above, a test for serial independence against serial dependence of order one can be constructed. In Subsection 5.3.1 this is done for the standard case and in Subsection 5.3.2 for the case involving nuisance parameters . In Section 5.4 we use local alternatives to make theoretical efficiency comparisons, in Section 5.5 we present the results of Monte Carlo simulations, Section 5.6 discusses the impact of the user-chosen parameters in

the model, in Section 5.7 we carry out a modest empirical study, and in Section 5.8 we give a short description of some other uses for our measure, including the analogous test statistic for serial independence against a serial dependence of order $J-1$ alternative. Our conclusions are summarised in Section 5.9.

5.2 Characteristic Function Based Measure

Suppose we wish to compare the distribution functions F_1 and F_2 of the J-variate random variables X_1 and X_2, respectively. It would be convenient if we would have a measure B, for which $B = 0$ if $F_1(x) = F_2(x)$, almost everywhere, and $B > 0$, otherwise.

It is a well-known fact that two distribution functions are equal almost everywhere, if and only if their respective characteristic functions are equal almost everywhere [cf. Lukacs (1970), Theorem 3.1.1]. It thus suffices to compare $E\exp(iu^TX_1)$ and $E\exp(iu^TX_2)$, where T denotes transposition, for all u. Define

$$\psi(u) = Ee^{iu^TX_1} - Ee^{iu^TX_2},\ u \in \Re^J,$$

the difference of the characteristic functions at u. Obviously, $\psi(u) = 0$, almost everywhere if and only if $F_1(x) = F_2(x)$, almost everywhere. Define

$$B = \int g(u)|\psi(u)|^2 du, \tag{5.1}$$

with g a density. Because a characteristic function is bounded, B is now bounded, also. There is a certain degree of arbitrariness involved in the choice of g, but this equally holds for other statistics. Indeed, one could easily allow for more generality in the way distribution functions are estimated in the empirical distribution function based serial independence tests of Skaug and Tjøstheim (1992) and Delgado (1993), and the very choice of the estimation method they made — although in their setting indeed the most obvious — is arbitrary, in the sense that it is

only one out of a number of possible distribution function estimation methods. It goes without saying that different choices lead to different results. We say a bit more about the arbitrariness regarding the choice of g in Section 5.6.

Obviously, if $\psi(u)$ is zero almost everywhere, then $B = 0$ and otherwise $B > 0$. Let $dF_\Delta(x) = dF_1(x) - dF_2(x)$, for all x. Then we can rewrite (5.1) as

$$\begin{aligned}
B &= \int g(u) \left| \{E\cos(u^T X_1) - E\cos(u^T X_2)\} + i\{E\sin(u^T X_1) - E\sin(u^T X_2)\} \right|^2 du \\
&= \int g(u) \left\{ E^2[\cos(u^T X_1) - \cos(u^T X_2)] + E^2[\sin(u^T X_1) - \sin(u^T X_2)] \right\}^2 du \\
&= \int g(u) \left\{ \left(\int \cos(u^T x) dF_\Delta(x) \right)^2 + \left(\int \sin(u^T x) dF_\Delta(x) \right)^2 \right\} du \\
&= \int g(u) \left\{ \cos(u^T x)\cos(u^T y) + \sin(u^T x)\sin(u^T y) \right\} dF_\Delta(x) dF_\Delta(y) du \\
&= \int g(u) \cos\{u^T(x-y)\} dF_\Delta(x) dF_\Delta(y) du.
\end{aligned}$$

Let us now have a closer look at g. Suppose now that g is symmetric. Then it has a real characteristic function, say a. Because $a(v) = \int g(u)\cos(u^T v)du$, for all v, we rewrite the last displayed equation as

$$B = \int a(x-y) dF_\Delta(x) dF_\Delta(y).$$

As B is just a sum of expectations relating to an observable series, it is very easy to estimate B once the applied researcher has selected a. Sometimes it may, for performance reasons, be desirable not to look at B, but rather at $\mathcal{I}$, where

$$\mathcal{I} = c_1(\mathcal{I}_1 - \mathcal{I}_2)^2 + c_2(\mathcal{I}_2 - \mathcal{I}_3)^2,$$

with c_1 and c_2 positive constants and $\mathcal{I}_1 = \int a(x-y)dF_1(x)dF_1(y)$, $\mathcal{I}_2 = \int a(x-y)dF_1(x)dF_2(y)$, and $\mathcal{I}_3 = \int a(x-y)dF_2(x)dF_2(y)$. We can do this, because $\mathcal{I}$ is bounded from below by $c_3 B^2$, for $c_3 = \min(c_1, c_2)/2 > 0$, as we now show. Indeed,

$$\frac{\mathcal{I}}{c_3} \geq 2(\mathcal{I}_1 - \mathcal{I}_2)^2 + 2(\mathcal{I}_2 - \mathcal{I}_3)^2 \quad \geq \quad (\mathcal{I}_1 - \mathcal{I}_2)^2 + 2(\mathcal{I}_1 - \mathcal{I}_2)(\mathcal{I}_3 - \mathcal{I}_2) + (\mathcal{I}_3 - \mathcal{I}_2)^2$$

$$= (\mathcal{I}_1 - 2\mathcal{I}_2 + \mathcal{I}_3)^2 = B^2, \tag{5.2}$$

and hence we can use $\mathcal{I}$ to replace B, if we desire.

An obvious use of the above principle to test for the equivalence of the distributions of elements of two mutually independent stationary series. Another possibility is to test for independence between elements in different series. We shall however look at the case, where we want to establish whether $F_{12}(x,y) = F(x)F(y)$, for almost all (x,y), with F_{12} the distribution function of (X_1, X_2) and F that of X_1 (and hence also of X_2). This case is more complicated than the previous one, as we shall see, because we will be faced with certain dependencies under the null hypothesis, due to overlapping terms.

The problem Csörgö (1985) investigated was that of testing for independence between elements in two i.i.d. series. He proposed to test for independence using a statistic of the form $N \sup_u |\psi(u)|^2$, and he showed that, under certain regularity conditions, this statistic converges to an estimable constant times a χ_1^2-distributed random variable. His statistic could, with minor alterations, obviously also be used to test for serial independence, but the proofs would be still more complicated than under the case he considered. Numerical optimisation routines need to be applied to compute it, and as ψ is unlikely to be unimodal this may well be expensive in computer-time. We have not found any literature on the performance details of Csörgö's (1985) statistic, neither in the setting of cross-sectional independence nor for serial independence. It is none the less an interesting alternative.

5.3 Testing for Serial Independence

We first consider at length the case of testing for serial independence against serial dependence of order one, and we suggest a test against serial dependence of any finite order in Section 5.8. We shall derive a result both when $\{X_t\}$ is observed and in the case of nuisance parameters.

As noted in Chapter 3, one is typically confronted with nuisance parameters in testing the specification of a model, in our setting usually a time series model. The disturbances in such a model are not observed, but our test may, under conditions explained in Section 5.3.2, be applied to the corresponding residuals. A few examples of models for which this is the case in a finance context are the AR (Autoregressive) model, the MA (Moving Average) model, and the NLMA (Non-Linear Moving Average) model. Unfortunately, the proposed test applied to residuals of an ARCH (AutoRegressive Conditional Heteroskedasticity) model does not asymptotically lead to the same value of the test statistic as that applied to the (unobserved) disturbances. However, the proposed test will still be consistent; only the asymptotic distribution under the null is no longer χ_1^2, and one will have to put up with simulated critical values. This would take away the advantage of the well-approximated and convenient asymptotic distribution of the proposed test statistic.

If we are to test for serial independence against a serial dependence of order one alternative, we are to test whether

$$F_{12}(x,y) = F(x)F(y), \text{ for almost all } x, y, \tag{5.3}$$

where F_{12} and F are the distribution functions of (X_1, X_2) and X_1, respectively. As we now wish to compare two bivariate distributions, we will use a slightly different notation from that used in the previous section. Still using the same framework as in the introduction, we are to test whether

$$\psi(u,v) = Ee^{i(uX_1+vX_2)} - Ee^{iuX_1}Ee^{ivX_1},$$

for almost all u, v, our statistic is based on $B = \int g(u)g(v)|\psi(u,v)|^2 dudv$. As suggested in Section 5.2, we do not use B itself, but $\mathcal{I}$. Setting $c_1 = c_2 = 1$, which is perhaps again

somewhat arbitrary, we have

$$\mathcal{I} = (\mathcal{I}_1 - \mathcal{I}_2)^2 + (\mathcal{I}_2 - \mathcal{I}_3)^2,$$

where (analogous to Section 5.2)

$$\begin{aligned}
\mathcal{I}_1 &= \int a(x-w)a(y-z)dF_{12}(x,y)dF_{12}(w,z), \\
\mathcal{I}_2 &= \int a(x-w)a(y-z)dF_{12}(x,y)dF(w)dF(z), \\
\mathcal{I}_3 &= \int a(x-w)a(y-z)dF(x)dF(y)dF(w)dF(z),
\end{aligned}$$

where $a = \int g(u)\cos(ux)du$ satisfies certain conditions, stated in Assumption 5.B. It is easily seen that $\mathcal{I}$ is indeed zero, if $F_{12}(x,y) = F(x)F(y)$, for all x, y. We shall estimate $\mathcal{I}_1$ to $\mathcal{I}_3$ by $\hat{\mathcal{I}}_1$ to $\hat{\mathcal{I}}_3$, which are defined further below.

5.3.1 Standard Case

In this section we set out the assumptions required for our main results for the standard statistic to go through. We actually just need two assumptions, both of which are fairly modest. The first is the trigonometric mixing weak dependence condition which is needed for some standard law of large numbers results for dependent processes. As noted in Chapter 3, it is implied by strong mixing, which was also explained in Chapter 3. The condition is complementary to the serial dependence of order one restriction that applies under the alternative hypothesis. It should be pointed out, though, that these are sufficient conditions, and that there is a large class of alternatives not satisfying the conditions imposed against which the test will none the less have power. The second applies to the choice of the function a.

Assumption 5.A *The stationary ergodic series $\{X_t\}$ is trigonometric mixing with mixing numbers $\alpha(t)$, in the sense of Definition 3.10.*

Definition 5.1 *We define $\mathcal{G}$ as the class of functions that are positive everywhere on $\Re$, and that have a Fourier transform that is bounded.*

The above definition obviously includes all densities that are everywhere positive, including the Gaussian density.

Assumption 5.B *Let a be a Fourier transform of any even function g that belongs to the class $\mathcal{G}$ and for which a, with $a(x) = \int g(u)\cos(ux)dx$, is not a degenerate U-statistic kernel with respect to F, in the sense of Definition 3.12.*

The functions a in Assumption 5.B are not kernels in the sense of traditional nonparametric theory. They need not integrate to one and need not be positive (almost) everywhere. Because they are based on an even function, however, they are even (and real) themselves, also. If a is degenerate, the asymptotic distribution will no longer be normal (cf. also Definition 3.12, Theorem 3.1, and the ensuing discussion). a is degenerate if and only if $Q_1 = \int a(x - X_1)dF(x)$ has a degenerate distribution. Choosing $a(x) = \exp(-\frac{1}{2}x^2)$, for all x, the characteristic function of the standard normal distribution, this would imply that $\int \exp\{-\frac{1}{2}(x-y)^2\}dF(x)$ does not depend on y, which for continuous distributions amounts to saying that $Ef(Y)$, with $Y \sim N(y,1)$ and $f(x) = F'(x)$, does not depend on y.

Let $a(X_t - X_s)$ be denoted by a_{ts}. We define

$$\hat{\tau} = \frac{1}{2}\left(\frac{\hat{\mathcal{I}}}{(\hat{\gamma} - \hat{\mathcal{I}}_3)^2}\right), \tag{5.4}$$

where

$$\begin{aligned}
\hat{\mathcal{I}} &= (\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2)^2 + (\hat{\mathcal{I}}_2 - \hat{\mathcal{I}}_3)^2, \\
\hat{\mathcal{I}}_1 &= \frac{1}{N^2}\sum_t\sum_s a_{ts}a_{t+1,s+1}, \\
\hat{\mathcal{I}}_2 &= \frac{1}{N^3}\sum_t\sum_s\sum_u a_{ts}a_{t+1,u},
\end{aligned}$$

$$\hat{I}_3 = \frac{1}{N^4}\sum_t\sum_s\sum_u\sum_v a_{ts}a_{uv},$$

$$\hat{\gamma} = \frac{1}{N}\sum_t\left(\frac{1}{N}\sum_s a_{ts}\right)^2.$$

The denominator in the definition of $\hat{\tau}$ is a variance estimate. It ensures that the asymptotic distribution of $N\hat{\tau}$ under the null is always χ^2_1. It should be noted that it is, in practice, much better to exclude the cross terms in the above summations to reduce unnecessary noise due to overlapping terms, but for notational clarity, we shall use the above V-statistic representation, here. In Section 5.5 we shall give the form in which it is best implemented.

Theorem 5.1 below establishes consistency against all departures from (5.3). This means that whenever (5.3) does not hold, our test statistic $N\hat{\tau}$ tends to infinity, in probability; that is, when the assumptions made in the theorem are satisfied. This is a very valuable property, because the test will, sample size permitting, reject *any* alternative to (5.3).

Theorem 5.1 (Consistency) *Let Assumptions 5.A and 5.B hold. The test* 'Reject H_0 if $N\hat{\tau} > C$', *for some* $0 < C < \infty$, *is consistent against all departures from (5.3).*

In Theorem 5.2 we establish the asymptotic distribution of $N\hat{\tau}$ under the null hypothesis.

Theorem 5.2 (Asymptotic Validity) *Let Assumption 5.B hold. Let the series* $\{X_t\}$ *be i.i.d.. Then*

$$N\hat{\tau} \xrightarrow{\mathcal{L}} \chi^2_1.$$

5.3.2 Nuisance Parameters

Often, the series of interest is not observed. In the case we have observed a proxy series that satisfies certain conditions, we can, as we will show below, apply our test to the proxy series, where the obtained results apply to the series of interest as well. An example of a situation in

which this problem may arise, is the standard regression model, where we are interested in the disturbances, but have only observed the residuals.

Assumption 5.C (Nuisance Parameters 1) *Instead of the series of interest, $\{X_t\}$, we have observed a proxy series, $\{Y_{Nt}\}$, for which a, possibly unknown, function m, a 'true' parameter vector, θ_0, its estimate, $\hat{\theta}$, and a vector series $\{Z_t\}$, possibly of infinite length, exist such that $X_t = m(Z_t;\theta_0)$ and $Y_{Nt} = m(Z_t;\hat{\theta})$, and for which a sequence $\{\delta_N\}$ exists such that $\hat{\theta} - \theta_0 = O_p(\delta_N^{-1})$. Finally, we assume that $\sup_x |a'(x)| < \infty$ and*

$$||m'(Z_t;\theta)||_\infty = o_p(\delta_N), \tag{5.5}$$

where we use the notation $m'(Z_t;\theta) = (\partial m/\partial\theta)|_{(Z_t;\theta)}$. Existence of derivatives is assumed implicitly.

Often, the convergence rate of the parameter estimate to the true parameter value will be of order $N^{-\frac{1}{2}}$. This would imply $\delta_N = N^{\frac{1}{2}}$. In the standard linear regression model with regressand Z_{t1} and regressors Z_{t2}, we have $m(Z_t;\theta) = Z_{t1} - \theta^T Z_{t2}$. Thus, we would need $\sup_t ||Z_{t2}||_\infty = o_p(N^{\frac{1}{2}})$. For i.i.d. regressors this would be a very weak condition, but it does for instance exclude linear trends, because $\sup_{t=1,\ldots,N} t = N$. In the case of linear trends one might consider taking first differences, but $Z_{t+1,1} - Z_{t1} = \theta^T(Z_{t+1,2} - Z_{t2}) + (\varepsilon_{t+1} - \varepsilon_t)$ and the disturbances are 1-dependent, even when the model is correctly specified. One could test if the series $\{(\varepsilon_1,\varepsilon_2),(\varepsilon_3,\varepsilon_4),\ldots\}$ is serially independent, but its serial independence would not necessarily imply serial independence of $\{\varepsilon_t\}$.

In effect, a special case of the above set up is the (stationary) AR(1)-model, where $A_t = \theta_0 A_{t-1} + X_t$, $|\theta_0| < 1$, and hence $Z_t = (A_{t-1}, A_t)^T$, for all t. Therefore, $m(Z_t;\theta) = A_t - \theta A_{t-1}$, and $m'(Z_t;\theta) = -A_{t-1}$, for all t,θ. Condition (5.5) thus implies that $\sup_t |A_t| = o_p(N^{\frac{1}{2}})$.

The above framework is, however, not restricted to linear models. In the case of the nonlinear moving average (NLMA) model, $A_t = X_t + \theta_0 X_{t-1} X_{t-2}$, with $\{X_t\}$ white noise, we have $Z_t = (X_{t-1}, X_{t-2}, A_t)^T$, $m'(Z_t;\theta) = -X_{t-1}X_{t-2}$, such that condition (5.5) is implied by $\sup_t |X_t| = o_p(N^{\frac{1}{4}})$, because θ_0 can, in the present setting, be estimated $\sqrt{N}$-consistently.

For the asymptotic validity result we require the following condition.

Assumption 5.D (Nuisance Parameters 2) *We require the series $\{W_t\}$, with $W_t = (X_t, Z_t, X_{t+1}, Z_{t+1})$, to be absolutely regular with mixing numbers $\beta(t)$. We assume that $\{X_t\}$ is i.i.d.. We also assume that Z_s does not depend on X_t for $s < t$ and that $E[m'(Z_1;\theta_0)|X_1]$ does not depend on X_1. Let $||M||_\infty$ denote the greatest element (in absolute value) in the matrix M. For the function m and the sequence $\{\delta_N\}$, described in Assumption 5.C, we need*

$$\delta_N^{-1} = o(N^{-\frac{1}{4}}), \tag{5.6}$$

$$\sup_x |a''(x)| < \infty, \tag{5.7}$$

$$\sup_{t,\theta\in(\theta_0,\hat{\theta})} ||m''(Z_t;\theta)||_\infty = o_p(N^{-\frac{1}{2}}\delta_N^{-2}), \tag{5.8}$$

$$\sup_t ||m'(Z_t;\theta_0)||_\infty = o_p(N^{-\frac{1}{4}}\delta_N), \tag{5.9}$$

and for some $d > 0$,

$$E||m'(Z_1;\theta_0)||_\infty^{2(1+d)} < \infty, \tag{5.10}$$

$$\sum_t \beta^{\frac{d}{1+d}}(t) < \infty, \tag{5.11}$$

where $m''(Z_t;\theta) = \partial^2 m/\partial\theta\partial\theta^T|_{(Z_t;\theta)}$. Existence of derivatives is again implicitly assumed.

The most restrictive condition in Assumption 5.D is the condition that Z_s does not depend on X_t for $s < t$ and that $E[m'(Z_1;\theta_0)|X_1]$ does not depend on X_1. This excludes ARCH, for instance, as $A_t = X_t\sqrt{1+\theta_0 A_{t-1}^2}$, such that $m(Z_t;\theta_0) = A_t(1+\theta_0 A_{t-1}^2)^{-\frac{1}{2}}$, $m'(Z_t;\theta_0) = -\frac{1}{2}A_t A_{t-1}^2(1+\theta_0 A_{t-1}^2)^{-\frac{3}{2}}$, and $E[m'(Z_t;\theta_0)|X_t] = -\frac{1}{2}E[X_t A_{t-1}^2(1+$

$\theta_0 A_{t-1}^2)^{-1}|X_t] = -\frac{1}{2}X_t E[A_{t-1}^2(1+\theta_0 A_{t-1}^2)^{-1}]$, which is generally dependent on X_t. As explained in Section 5.3, the consistency result is not affected, and one may use simulated critical values instead of those implied by the asymptotic distribution. Condition (5.11) restricts the speed at which the mixing numbers may tend to zero. The smaller d, the faster the mixing numbers $\beta(t)$ must converge to zero (as $t \to \infty$), and hence the less dependence is allowed for. The value for d is limited by the number of moments $m'(Z_t;\theta_0)$ has [cf. condition (5.10)]. So, the more moments m' has, the more dependence is allowed and vice versa. Examples of $m'(Z_t;\theta_0)$ are, as we have seen above, $-Z_{t2}$, in a linear regression framework, $-A_{t-1}$, for an AR(1) model and $-X_{t-1}X_{t-2}$ for an NLMA model of the form $A_t = X_t + \theta_0 X_{t-1}X_{t-2}$.

The sequence $\{\delta_N\}$ is restricted by conditions (5.6), (5.8) and (5.9). Condition (5.6) implies that parameter estimates should converge to the associated parameter values at a rate faster than $N^{-\frac{1}{4}}$. This is true for most parameter estimates used in practice. Conditions (5.8) and (5.9) are more restrictive. They are very similar, but more restrictive than condition (5.5) of Assumption 5.C. Sufficient conditions are for the linear regression model that $\sup_t |Z_{t2}| = o_p(N^{\frac{1}{4}})$, as $\delta_N^{-1} = N^{-\frac{1}{2}}$, for the AR(1) model that $\sup_t |A_t| = o_p(N^{\frac{1}{4}})$, and for the NLMA-model that $\sup_t |X_t| = o_p(N^{\frac{1}{8}})$.

Below we state the consistency and asymptotic validity theorems for the case with nuisance parameters.

Theorem 5.3 (Nuisance Parameters - Consistency) *Let $\hat{\tau}^Y$ be defined as $\hat{\tau}$, but using $\{Y_{Nt}\}$, rather than $\{X_t\}$. Under Assumptions 5.A, 5.B and 5.C, the test 'Reject the null when $N\hat{\tau}^Y > C$', is consistent against all departures from (5.3).*

Theorem 5.4 (Nuisance Parameters - Asymptotic Validity) *Let Assumptions 5.B, 5.C and 5.D hold. Then*

$$N\hat{\tau}^Y \xrightarrow{\mathcal{L}} \chi_1^2.$$

5.4 Local Alternatives

In this section we study the behaviour of the proposed test against local alternatives [cf. Pitman (1948)], and demonstrate that it has more power than the correlation dimension test of Brock, Dechert and Scheinkman (1987) against all departures satisfying certain regularity conditions.

For any two tests for a hypothesis for which under the null hypothesis $\theta_0 = 0$, say test 1 and test 2, denote the local alternative parameters by $\theta_{1N} = c_1 N^{-\eta_1}$ and $\theta_{2N} = c_2 N^{-\eta_2}$, where c_1, c_2, η_1, η_2 are chosen such that both tests have the same power for a sample of size N. We stress the dependence of θ on N here, to improve clarity. Obviously, if $\eta_1 > \eta_2$, test 1 is more powerful than test 2. Indeed, the asymptotic relative efficiency of test 2 with respect to test 1 is then equal to zero. If $\eta_1 = \eta_2$, then if $c_1 < c_2$, test 1 is again more powerful. For asymptotic χ^2_1-statistics (under the null), the asymptotic relative efficiency is defined as the ratio of the respective non-centrality parameters of the asymptotic non-central χ^2_1-distribution under the local alternative. As we shall see, however, it will be hard to compare the proposed statistic to other statistics, as the former does not generally have an asymptotic non-central χ^2_1-distribution under the local alternative.

We shall first give a general result, which we shall thereafter use for specific alternatives. We shall only look at the simple case with one parameter.

Assumption 5.E (Local Alternatives) *Let $\{X_t\}$ denote an i.i.d. series and let $\{Y_{Nt}\}$, which depends on N, denote a series corresponding to the local alternative, such that $Y_{Nt} = m(X_t, X_{t-1}, \ldots; \theta_N)$ and $X_t = m(X_t, X_{t-1}, \ldots; 0)$, for all t. Let $m'_t = \partial m/\partial\theta_N|_{(X_t, X_{t-1}, \ldots; 0)}$ and $m''_t = \partial^2 m/\partial\theta_N^2|_{(X_t, X_{t-1}, \ldots; 0)}$. Let $W_t = (m'_t, m''_t, X_t)$ be absolutely regular with mixing numbers $\beta(t)$. Let furthermore, for some $d > 0$,*

$$\sum_t \beta^{\frac{d}{1+d}}(t) \quad < \quad \infty, \tag{5.12}$$

$$\sup_t E|m_t'|^{4(1+d)} < \infty, \tag{5.13}$$

$$\sup_t E|m_t''|^{2(1+d)} < \infty, \tag{5.14}$$

$$\sup_t |m_t'| = o_p(\theta_N^{-\frac{1}{3}}), \tag{5.15}$$

$$\sup_t |m_t'| \sup_t |m_t''| = o_p(\theta_N^{-1}), \tag{5.16}$$

$$\sup_{t;\tilde{\theta}_N \in (0,\theta_N)} |m'''(X_t, X_{t-1}, \ldots; \tilde{\theta}_N)| = o_p(\theta_N^{-1}). \tag{5.17}$$

We also assume that a is three times boundedly differentiable.

Many of the conditions in Assumption 5.E are similar to those in Assumption 5.D. Notice though, that in the case of nuisance parameters, we were interested in the behaviour of the disturbance terms of a certain model, whereas here we are interested in the behaviour of the series itself, under the (local) alternative. As a consequence, m has a somewhat different interpretation from that in the case of nuisance parameters, which will become clear in the example below. As we shall see in Theorem 5.5, usually either $\theta_N = cN^{-\frac{1}{4}}$, or $\theta_N = cN^{-\frac{1}{2}}$. The conditions are strongest for the slowest convergence rate, and we shall therefore examine their implications when that convergence rate applies. If the local alternative is an AR(1)-process, then $Y_{Nt} = \theta_N Y_{N,t-1} + X_t$, or $Y_{Nt} = \sum_{j=0}^{\infty} \theta_N^j X_{t-j} = m(X_t, X_{t-1}, \ldots; \theta_N)$. Therefore $m_t' = X_{t-1}$, $m_t'' = 2X_{t-2}$, $m'''(X_t, X_{t-1}, \ldots; \tilde{\theta}_N) = \sum_{j=0}^{\infty}(j+3)(j+2)(j+1)\tilde{\theta}_N^j X_{t-j}$, which is bounded from above by $6|X_{t-3}| + \sum_{j=1}^{\infty}(j+1)(j+2)(j+3)(c^{-4}N)^{-\frac{j}{4}}|X_{t-j}|$. Conditions (5.15) to (5.17) are thus implied by $\sup_t |X_t| = o_p(N^{\frac{1}{12}})$.

Theorem 5.5 below discusses conditions under which the proposed test statistic has local alternative parameter tending to zero at rate $N^{-\frac{1}{2}}$ or $N^{-\frac{1}{4}}$. The conditions are fairly complicated, but their structure is, hopefully adequately, explained in the ensuing discussion.

Theorem 5.5 (Local Alternatives) *Let Assumptions 5.B and 5.E hold. Define*

$$\mathcal{E}_{01}(t,s,u,v) = a_{ts}(m_u' - m_v')a_{uv}' + a_{uv}(m_t' - m_s')a_{ts}',$$

$$\begin{aligned}\mathcal{E}_{02}(t,s,u,v) &= \frac{1}{2}\{2(m'_t - m'_s)a'_{ts}(m'_u - m'_v)a'_{uv} \\ &+ a_{ts}(m''_u - m''_v)a'_{uv} + a_{uv}(m''_t - m''_s)a'_{ts} \\ &+ a_{ts}(m'_u - m'_v)^2 a''_{uv} + a_{uv}(m'_t - m'_s)^2 a''_{ts}\}.\end{aligned}$$

If $E_I[\mathcal{E}_{01}(t,s,t+1,s+1)] = E_I[\mathcal{E}_{01}(t,s,t+1,u)] = E_I[\mathcal{E}_{01}(t,s,u,v)]$, *then let* $\theta_N = cN^{-\frac{1}{4}}$, *else* $\theta_N = cN^{-\frac{1}{2}}$. *Define*

$$\mathcal{E}_1(\theta_N) = \theta_N E_I[\mathcal{E}_{01}(t,s,t+1,s+1)] + \theta_N^2 E_I[\mathcal{E}_{02}(t,s,t+1,s+1)], \tag{5.18}$$

$$\mathcal{E}_2(\theta_N) = \theta_N E_I[\mathcal{E}_{01}(t,s,t+1,u)] + \theta_N^2 E_I[\mathcal{E}_{02}(t,s,t+1,u)], \tag{5.19}$$

$$\mathcal{E}_3(\theta_N) = \theta_N E_I[\mathcal{E}_{01}(t,s,u,v)] + \theta_N^2 E_I[\mathcal{E}_{02}(t,s,u,v)], \tag{5.20}$$

$$\tilde{\mathcal{E}}_{ij} = \lim_{N\to\infty} \sqrt{N}\{\mathcal{E}_i(\theta_N) - \mathcal{E}_j(\theta_N)\},\ i,j = 1,2,3. \tag{5.21}$$

Then

$$N\hat{\tau}^Y \xrightarrow{\mathcal{L}} \left(\lambda_0 + \frac{\tilde{\mathcal{E}}_{13}}{2(\gamma - \mathcal{I}_3)}\right)^2 + \left(\frac{\tilde{\mathcal{E}}_{12} + \tilde{\mathcal{E}}_{32}}{2(\gamma - \mathcal{I}_3)}\right)^2 \tag{5.22}$$

where $\lambda_0 \sim N(0,1)$.

As said before, we wish to establish how $\hat{\tau}^Y$ behaves for any local alternative parameter sequence θ_N tending to zero. We use a Taylor series expansion, with respect to θ_N around 0, to examine the behaviour of $\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1$, $\hat{\mathcal{I}}_2^Y - \hat{\mathcal{I}}_2$ and $\hat{\mathcal{I}}_3^Y - \hat{\mathcal{I}}_3$ for different θ_N's. The $\mathcal{E}_{01}$'s in Theorem 5.5 are used in the first order terms of the afore-mentioned expansions, and the $\mathcal{E}_{02}$'s in the second order terms. Indeed, we show that $\theta_N^{-1}(\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1) \xrightarrow{P} E_I[\mathcal{E}_{01}(t,s,t+1,s+1)]$, and we obtain similar expressions for $\theta_N^{-1}(\hat{\mathcal{I}}_2^Y - \hat{\mathcal{I}}_2)$ and $\theta_N^{-1}(\hat{\mathcal{I}}_3^Y - \hat{\mathcal{I}}_3)$, where the second order terms are $O_p(\theta_N^2)$. However, if $E_I[\mathcal{E}_{01}(t,s,t+1,s+1)] = E_I[\mathcal{E}_{01}(t,s,t+1,u)]$ and $E_I[\mathcal{E}_{01}(t,s,t+1,u)] = E_I[\mathcal{E}_{01}(t,s,u,v)]$, then both $\theta_N^{-1}(\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2^Y + \hat{\mathcal{I}}_2)$ and $\theta_N^{-1}(\hat{\mathcal{I}}_2^Y - \hat{\mathcal{I}}_2 - \hat{\mathcal{I}}_3^Y + \hat{\mathcal{I}}_3)$ converge to zero, and the second order term is then of lowest order. Under these circumstance $\hat{\tau}^Y - \hat{\tau} = O_p((\theta_N^2)^2) = O_p(\theta_N^4)$, such that $\theta_N \sim N^{-\frac{1}{4}}$, in order to ensure that $N(\hat{\tau}^Y - \hat{\tau}) = O_p(1)$.

Consider for instance an MA(1) process, $Y_{Nt} = \theta_N X_{t-1} + X_t$, with $\{X_t\}$ white noise. Then Assumption 5.E is satisfied, and $m_t' = X_{t-1}$, $m_t'' = 0$. Thus, $\mathcal{E}_{01}(t,s,u,v) = a_{ts}(X_{u-1} - X_{v-1})a_{uv}' + a_{uv}(X_{t-1} - X_{s-1})a_{ts}'$, for all t,s,u,v, and therefore $E_I\mathcal{E}_{01}(t,s,t+1,s+1) = E_I[a_{ts}(X_t - X_s)a_{t+1,s+1}' + a_{t+1,s+1}(X_{t-1} - X_{s-1})a_{ts}'] = 0 = E_I\mathcal{E}_{01}(t,s,t+1,u) = E_I\mathcal{E}_{01}(t,s,u,v)$, because $E_I a_{ts}' = E_I[a_{ts}(X_t - X_s)] = 0$, and hence we can not let $\theta_N \sim N^{-\frac{1}{2}}$, but need to let $\theta_N \sim N^{-\frac{1}{4}}$, and the proposed test therefore has zero (asymptotic relative) efficiency, against an MA(1)-alternative, in comparison with a parametric test, for which $\theta_N \sim N^{-\frac{1}{2}}$. This suggests one may want to examine the possibilities of using asymmetric a; it is probably feasible to extend Theorems 5.1 and 5.2 to cover certain asymmetric a. However, it would generally not be possible to provide an adequate nuisance parameter result, as such results depend on the 'inefficiency' of the test statistic, in the sense that replacing the nuisance parameters by their estimates does not (asymptotically) affect the test result.

Theorem 5.5 basically states that the proposed statistic is asymptotically non-central χ_1^2 plus a positive constant, under the local alternative. We can therefore not compare the non-centrality parameters, or rather: if the non-centrality parameter of the asymptotic distribution of the proposed test is greater than or equal to that of some other test (given that both θ's converge at the same *rate*), the proposed test is more powerful; if it is less, on the other hand, the results are inconclusive, as they will depend on the values of c_1 and c_2. In the latter case, if $c_1 = c_2$, the power of the proposed test is one for large c_1 (provided that $\tilde{\mathcal{E}}_{12} + \tilde{\mathcal{E}}_{32} \neq 0$), whereas it is less than that for the other test, for small c_1; this can be most easily seen by examining the definition of $\tilde{\mathcal{E}}_{ij}$ in expression (5.21), above, substituting for the $\mathcal{E}_i(\theta)$, from (5.18) through (5.20).

The BDS test statistic is equal to $\sqrt{N}(\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_3)/\{2(\hat{\gamma} - \hat{\mathcal{I}}_3)\}$, for a kernel $a(x) = I(|x| < \eta)$, for some $\eta > 0$ (a kernel that does not satisfy the conditions of Assumption 5.B, nor those of

Assumption 5.E). Squaring their statistic yields $N[(\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_3)^2/\{4(\hat{\gamma} - \hat{\mathcal{I}}_3)^2\}$. Obviously, their a is not differentiable, but if their a is replaced by a differentiable kernel, it is easy to establish (from the proof to Theorem 5.5), that the asymptotic distribution of their statistic (squared) under the local alternative is $[\lambda_0 + \tilde{\mathcal{E}}_{13}/\{2(\gamma - \mathcal{I}_3)\}]^2$, or the same distribution as the proposed test, albeit with a zero additional constant; that is if the same kernel is used for both test statistics. Thus, the proposed test is always at least as efficient as an adapted BDS-test using the same kernel, and only equally efficient if $\tilde{\mathcal{E}}_{12} + \tilde{\mathcal{E}}_{32} = 0$.

5.5 Simulations

It may appear from the definition of $\hat{\tau}$, that our statistic is very expensive in terms of computer time as it involves a four-fold sum. This is not the case, however, as we can write

$$\begin{aligned}
\hat{\mathcal{I}}_2 &= \frac{1}{N^3}\sum_t\sum_s\sum_u a_{ts}a_{t+1,u} = \frac{1}{N}\sum_t \mathcal{H}_t\mathcal{H}_{t+1}, \\
\hat{\mathcal{I}}_3 &= \frac{1}{N^4}\sum_t\sum_s\sum_u\sum_v a_{ts}a_{uv} = \left(\frac{1}{N}\sum_t \mathcal{H}_t\right)^2,
\end{aligned}$$

where $\mathcal{H}_t = \frac{1}{N}\sum_s a_{ts}$. The number of operations required to compute $\hat{\tau}$ is therefore $O(N^2)$. Computing the test statistic for a single data set will generally not be prohibitive in terms of computer time, and as the actual size is quite close to the nominal size under the null, there will generally be no need to use simulated critical values.

Additional time savings may be had by using the symmetry in a (such that $a_{ts} = a_{st}$, for all s, t), or, if one if willing to discretise the data, to use the Fast Fourier Transform [cf. Cooley and Tukey (1965)]. The approach is very similar to that used for the computation of kernel estimates using the Fast Fourier Transform [cf. Silverman (1982) and Härdle (1987)]. Indeed, for instance $\int \sum_t a(x - X_t)e^{iux}dx = \sum_t \int a(y)e^{iu(y+X_t)}dy = \tilde{a}(u)\sum_t e^{iuX_t}$, where $\tilde{a}(u)$ is the Fourier transform of a. Using the Fast Fourier Transform to compute the afore Fourier

transform, and then using a similar procedure to obtain $\sum_t a(x - X_t)$.

To simplify notation we have let all sums run over all observations. Particularly in smaller samples, it is advisable to exclude terms for which the difference between the summation indices is less than or equal to one, as these may generate noise due to overlaps. The definitions we use in our experiments are as follows.

$$\begin{aligned}
\hat{I}_1 &= \frac{1}{(N-1)(N-4)} \sum_{t<N} \sum_{s \neq t-1,t,t+1,N} a_{ts} a_{t+1,s+1}, \\
\hat{I}_2 &= \frac{1}{(N-1)(N-2)(N-3)} \sum_{t<N} \sum_{s \neq t,t+1} \sum_{u \neq t,t+1,s} a_{ts} a_{t+1,u}, \\
\hat{I}_3 &= \frac{1}{N(N-1)(N-2)(N-3)} \sum_{t} \sum_{s \neq t} \sum_{u \neq t,s} \sum_{v \neq t,s,u} a_{ts} a_{uv}, \\
\hat{\gamma} &= \frac{1}{N(N-1)(N-2)} \sum_{t} \sum_{s \neq t} \sum_{u \neq s,t} a_{ts} a_{tu}.
\end{aligned}$$

We have carried out experiments to establish the size of our test and also to compare the power of our test with that of some other statistics. It is impossible to make performance comparisons with all other statistics, in view of the vast number of such alternatives. We shall therefore limit ourselves to some of the better performing ones [cf. Skaug and Tjøstheim (1992b)].

We have simulated the 5%, 2.5%, and 1% critical values for simulated Gaussian i.i.d. time series with 25, 100, and 250 observations, using 8192 replications in each of these 9 cases. The results, represented in table 2, are very encouraging. Even for a sample size of 100 observations, the critical values are close to the asymptotic ones, and for 250 observations, they are closer still. As one would expect, the critical values for 25 observations deviate substantially from the asymptotic ones, rendering the rejection rate, under the null hypothesis for a test based on the 5% asymptotic critical values, to be almost twice as high as it should be. It seems that for sample sizes of over 100 observations, one may use the asymptotical critical values, and still get reliable results, at least for Gaussian series.

Some tests are not consistent against all departures from the null and there are therefore

departures from the null against which they have no power, even asymptotically. The only such statistics included are the BDS-test, the proposed test with asymmetric a (Gaussian when its argument is positive and Cauchy, otherwise), and the correlation test. The rationale for including the first is that it is widely used in finance. Used as a one-sided test, it is not consistent against alternatives for which $I(|X_t - X_s| \leq \lambda)$ and $I(|X_{t+1} - X_{s+1}| \leq \lambda)$, for some $\lambda > 0$ chosen in advance, are negatively correlated, which is true for a large class of models, including the *'inverted ARCH'* model $X_{t+1} = \varepsilon_{t+1}\sqrt{1 + \frac{\rho}{1+X_t^2}}$, where $\{\varepsilon_t\}$ is an i.i.d. Gaussian process. We shall, therefore only consider the two-sided variant, which is still not consistent against *all* departures, but covers a much larger set of alternatives than the one-sided version. For reasons set out in Section 5.4, we also try an asymmetric a, although it does not satisfy the conditions made earlier. The correlation test is included as it is the oldest test and is still used very frequently. The format in which it is used here is $N(\hat{r}_{DW} - 2)^2$, where $\hat{r}_{DW}$ denotes the Durbin-Watson (1950) test statistic.

The other tests included are the proposed test with Gaussian and Double Exponential g (such that a is $\exp(-\frac{1}{2}x^2)$ and $1/(1+x^2)$, respectively), the empirical distribution function test of Skaug and Tjøstheim (1992b) and Delgado (1993), and the entropy based test of Robinson (1991a).

We try nine different models, each with a few parameter values. Let $\{\varepsilon_t\}$ be i.i.d., and in our case $N(0,1)$. Then our models are the Autoregressive model of order one model [**AR**, $X_{t+1} = \rho X_t + \varepsilon_{t+1}$], the Moving Average model of order one model [**MA**, $X_{t+1} = \rho\varepsilon_t + \varepsilon_{t+1}$], the Autoregressive Conditional Heteroskedasticity or order one model [**ARCH**, $X_{t+1} = \varepsilon_{t+1}\sqrt{1 + \rho X_t^2}$], the Threshold Autoregressive model [**TAR**, $X_{t+1} = \rho X_t + \varepsilon_{t+1}$, when $X_t > 1$, and $X_{t+1} = \zeta X_t + \varepsilon_{t+1}$, when $X_t \leq 1$], the Bilinear model [**BLM**, $X_{t+1} = (\rho + \zeta\varepsilon_t)X_t + \varepsilon_{t+1}$], the Nonlinear Moving Average model [**NLMA**, $X_{t+1} = \varepsilon_{t+1} + \rho\varepsilon_t\varepsilon_{t-1}$], the Inverted ARCH model

[**INVARCH**, $X_{t+1} = \varepsilon_{t+1}\sqrt{1 + \frac{\rho}{1+X_t^2}}$], the Exponential Model [**EM**, $X_{t+1} = \rho X_t e^{-\zeta X_t^2} + \varepsilon_{t+1}$], and the Cosine Model [**CM**, $X_{t+1} = \rho\cos(X_t) + \varepsilon_{t+1}$].

The power comparison results are presented in table 3. The results apply to data sets of 100 observations. The number of replications is 8192. To create a level playing field we have used simulated critical values for all tests.

As expected, the correlation test performed much better than all other tests against the linear AR and MA alternatives, and also against EM. It also did fairly well against the TAR alternatives. There are, however, a significant number of alternatives against which it has no or little power, such as ARCH, INVARCH, and NLMA. One may find it surprising that the correlation test often rejects significantly more or less than in 5% of the cases. The reason, as explained in Chapter 3, is that $E[X_1^2 X_2^2] \neq E^2[X_1^2]$, such that $(1/N)\sum_t X_t^2$ does not consistently estimate the standard deviation of $(1/\sqrt{N})\sum_t X_t X_{t+1}$. The results suggest that the standard deviation is underestimated in the ARCH and NLMA cases and overestimated in the INVARCH case. One should note that the rejection rates (for the correlation test) will, for all alternatives under which the tested series is uncorrelated, remain more or less the same, regardless of sample size.

The empirical distribution function test also has very limited power against a number of alternatives, such as ARCH, INVARCH and NLMA, but because it is consistent against all departures from (5.3), its power will tend to one, when sample size tends to infinity. In small and moderate samples its usefulness is limited, as it basically has power against the same range of alternatives as the correlation test, albeit slightly less in general. An exception is the not extremely relevant cosine model.

The entropy based test generally has less power than both the correlation test and the empirical distribution function test against the linear alternatives, but more power against the

nonlinear alternatives. Its performance against the latter is, however, not quite as good as that of the BDS test, or of the proposed test for any of the tried kernels.

The asymmetric kernel did not, in contrast to what we had expected, lead to a higher rejection rate. Perhaps the kernel tried was not asymmetric enough, or the enhanced power will only become visible when sample size is increased. The performance of the proposed test with either symmetric kernel seems to perform a bit better than the correlation dimension test against all alternatives. Undoubtedly, there is a range of alternatives against which the BDS test or the proposed test performs much better than the other, in small samples, aside from the alternatives against which the BDS test is not consistent. In view of Section 5.4, it is to be expected, however, that the proposed test will perform better than the BDS test, against any alternative satisfying the assumptions made there, as sample size increases.

5.6 How to Choose g

There are many ways to choose the kernel a, or g for that matter, which implies a degree of arbitrariness in the proposed test. Often authors make implicit arbitrary choices, as we have done setting $c_1 = c_2 = 1$ in Section 5.2 where any combination of positive weights could have been chosen, or as Brock, Dechert and Scheinkman (1987) have done in selecting the kernel $I(|X_t - X_s| \leq \eta)$, for some positive η, rather than any other possible choice. Indeed, the practitioner will be faced with an even more difficult problem of arbitrariness, namely which test statistic to choose.

It is hard to give general guide lines, and we have not succeeded, indeed we think it is impossible, to find *the* optimal kernel. Which kernel is optimal depends on the unknown (possible) dependence structure in the data. If the dependence structure were known, one would have no need to use a nonparametric independence test.

In the simulations of Section 5.5, we found that there was not much difference in power between the two symmetric kernels selected. It is probably possible to select awkwardly shaped kernels that do not perform well, however. Generally, it seems that the Gaussian kernel gives quite acceptable results, so, until significant cases are found in which it does not have reasonable power, one may well use it. However, if the practitioner has some indication of what the most likely class of alternatives would be, he may wish to see if a more efficient kernel can be chosen for his specific alternative, nevertheless maintaining the valuable consistency property of the test. A possible approach would be to obtain an expression for the power against his alternative as a function of the local alternative parameter [cf. Section 5.4], and optimise it with respect to a. It will not generally be possible to obtain the optimal a, but one may, by a trial and error method, improve power.

A very important issue is to set the scale of the argument of a. Indeed, both $\exp(-\frac{1}{2}x^2)$ and $\exp(-10000x^2)$ are allowed choices for a, but they will not lead to the same results. An intuitive strategy is to first normalise the data by dividing all the elements by an estimate of $\sqrt{VX_1}$ or $\text{Med}(|X_1|)$, and adjusting the scale of the argument of the kernel until the nominal (asymptotic) size is close to the real (small sample) size. We did not select the kernels in Section 5.5 according to this rule, so that the size results therein contained are still valid. Indeed, the above rule of thumb was conceived as a result of the simulations.

5.7 Testing the Random Walk Hypothesis

In this section we present a very limited empirical example using daily, weekly and monthly exchange rate data. Our aim is to test the random walk hypothesis or, in other words, to test whether $(\log X_{t+1} - \log X_t)$ is serially independent, where X_t is the exchange rate at time t.

Table 4 contains the results. A few remarks are in place here. All exchange rates are

against the U.S. dollar. The data were derived from the Bank of England Quarterly Bulletin. The weekly and monthly series run over a period starting January 1974 and ending July 1985, whereas the daily series start on October 1, 1981, also ending July 1985. The reason for this is that as of October 1, 1981, transaction settlement procedures were changed [cf. Whistler (1990)], which affected the behaviour of the series, as Whistler found. A consequence is that the number of observations in the daily series is not much greater than that in the weekly series. The weekly data were collected every Friday, and the monthly data on the last Friday of every month. We have not attempted to eliminate calendar effects. We have, however, divided all observations (the first differences of the logarithms of the original observations, that is) through by the sample standard deviation, to avoid evaluating a in its tails too much. The a chosen here, is the standard-normal density.

As the critical value of the χ^2_1-distribution for a 5 % significance level is 3.84, the random walk hypothesis is rejected for all series. A point of interest is that the random walk hypothesis seems to be rejected much more strongly for the weekly series than for the daily. However, as noted above the number of daily observations is not much greater than the number of weekly observations. Moreover, calendar effects are likely to be far more pronounced in the daily series than in the weekly series. A final observation one could make is that there may still be some day of the week effects, although these should be fairly small under the new settlement procedure.

Whistler (1990) found that testing the random walk hypothesis parametrically, did not lead to a rejection for the weekly or monthly series, except for the weekly \$/JY rates. At the 5 % level, ARCH effects could be found for all, but the BP/\$ and \$/JY monthly rates. The proposed test also rejected the null for the BP/\$ and \$/JY rates, suggesting other factors than (parametric) ARCH may also play a role.

Robinson (1991a) found that his entropy based test also rejected serial independence for

all above series, but as we have seen in Chapter 4, and as has independently been discovered by Drost and Werker (1993), its asymptotic distribution under the null is generally not well approximated in samples of moderate size, which may well cause rejection rates to be overstated.

The above example underlines the proposed test as a general test, picking up characteristics parametric tests, or tests testing a specific alternative hypothesis, may not.

5.8 Extensions

An obvious extension is to examine the impact of using an asymmetric a. There is plenty of scope for extensions outside the serial independence against serial dependence of order one test setting, as the number of problems our measure can be applied to is virtually unlimited. Any hypothesis requiring two distributions to be compared can be treated in a similar way. Indeed, Section 5.2 provides the foundation for any such test. Examples are serial independence tests with higher order alternatives, tests for time series reversibility, structural breaks, and normality.

A test for $J-1$-th order dependence would involve comparing the characteristic functions of $F_J(x_1,\ldots,x_J)$ and $\prod_{j=1}^{J} F(x_j)$, for almost all $x \in \Re^J$. A statistic in our setup would then estimate an upperbound to $\int g(u)|\psi(u)|^2 du$, where $\psi(u) = E\exp(i\sum_{j=1}^{J} u_j X_j) - \prod_{j=1}^{J} E\exp(iu_j X_1)$, for all $u \in \Re^J$. The corresponding test statistic is

$$N\hat{T}^{(J)} = \frac{N}{2}\left(\frac{(\hat{\mathcal{I}}_1^{(J)} - \hat{\mathcal{I}}_2^{(J)})^2 + (\hat{\mathcal{I}}_2^{(J)} - \hat{\mathcal{I}}_3^{(J)})^2}{\hat{V}}\right), \tag{5.23}$$

where

$$\hat{\mathcal{I}}_1^{(J)} = \frac{1}{N^2}\sum_{t}\sum_{s}\prod_{j=1}^{J} a_{t+j-1,s+j-1}, \tag{5.24}$$

$$\hat{\mathcal{I}}_2^{(J)} = \frac{1}{N^{J+1}}\sum_{t}\sum_{s_1}\cdots\sum_{s_J}\prod_{j=1}^{J} a_{t+j-1,s_j}, \tag{5.25}$$

$$\hat{\mathcal{I}}_3^{(J)} = \frac{1}{N^{2J}}\sum_{t_1}\cdots\sum_{t_J}\sum_{s_1}\cdots\sum_{s_J}\prod_{j=1}^{J} a_{t_j,s_j}, \tag{5.26}$$

$$\hat{V} = \frac{\hat{\gamma}^{J+1} + \hat{\mu}^2\hat{\gamma}^J + J(J-4)\hat{\mu}^{2J-2}\hat{\gamma}^2 + (-2J^2+8J-5)\hat{\mu}^{2J}\hat{\gamma} + (J^2-4J+3)\hat{\mu}^{2J+2}}{\hat{\gamma} - \hat{\mu}^2} \quad (5.27)$$

with $\hat{\mu} = \frac{1}{N^2}\sum_t \sum_s a_{ts}$. The denominator in the last displayed equation can be factorised out to obtain

$$\hat{V} = \hat{\gamma}^J + 2\hat{\mu}^2\hat{\gamma}^{J-1} + \cdots + 2\hat{\mu}^{2J-4}\hat{\gamma}^2 + (J^2-4J+2)\hat{\mu}^{2J-2}\hat{\gamma} + (-J^2+4J-3)\hat{\mu}^{2J},$$

if one so desires. A derivation of the above result is given in Appendix 5.C. Extending the proofs of Theorems 5.1 and 5.2 is only cumbersome notationally, as we also show in Appendix 5.C, and under serial independence we have

$$N\hat{\tau}^{(J)} \xrightarrow{\mathcal{L}} \chi^2_1, \quad (5.28)$$

for $2 \leq J < \infty$.

If a time series is *reversible*, if the process generating $\{X_t\}$ is the same as that of the same time series in reversed order. Testing time series reversibility of order $J-1$ would involve testing whether $F_J(x_1,\ldots,x_J) = F_J(x_J,\ldots,x_1)$, for almost all $x \in \Re^J$ and thus whether $\psi(u) = E\exp(i\sum_{j=1}^J u_j X_j) - E\exp(i\sum_{j=1}^J u_j X_{J+1-j})$ is zero for almost all u. This problem is harder to tackle than the others suggested here, because we can not assume independence under the null.

If one wishes to know whether there is a structural break at a certain point, one could test for the equivalence of the distribution function of an element in the series, before and after this time period. This specific time period needs to be known in advance, though, and such a test would be harder to construct than a serial independence test, because there may again not be serial independence under the null.

Normality could be tested by comparing F to the Gaussian distribution function with the sample mean and variance. A complicating factor here is that we have not observed the true

mean and variance and it is not certain that the sample mean and variance converge fast enough for them not to have an impact asymptotically.

Finally, as suggested before, we could test the equivalence of the distributions of elements in two separate stationary series or their independence.

5.9 Conclusions

In this chapter we have suggested a measure, that can be used to test a wide variety of hypotheses. We have given an example for the hypothesis of serial independence against a serial dependence of order one alternative. The conditions are weak, but performance (both size and power) is quite good and the asymptotic distribution is tractable. We have also shown that the proposed test will generally be at least as efficient as the correlation dimension test of Brock, Dechert and Scheinkman (1987). Finally, in the modest empirical example, it turned out that the proposed test can be very useful in practice, also.

Appendix

5.A Proofs of Main Results

In this appendix we shall prove the main results of Chapter 5. In some of the proofs technical lemmas are used that can be found in Appendix 5.B. We shall use the following quantities and estimates fairly frequently: $\mu = \int a(x-w)dF(x)dF(w)$, $\hat{\mu} = \frac{1}{N^2}\sum_t \sum_s a_{ts}$, $\gamma = \int a(x-w)a(x-z)dF(x)dF(w)dF(z)$, $Q_t = \int a(X_t - x)dF(x)$, and $R_t = \int a(X_t - x)a(X_{t+1} - y)dF_{12}(x,y)$.

Proof of Theorem 5.1

Suppose $\hat{\tau} \xrightarrow{P} \tau$, where $\tau > 0$ under the alternative hypothesis. Suppose the null does not hold and hence $\tau > 0$. Now choose N so large as to ensure $C/N < \tau$ and define $\varepsilon = \tau - C/N > 0$. Then

$$P[N\hat{\tau} > C] = P[N(\hat{\tau} - \tau) > C - N\tau] = P[\tau - \hat{\tau} < \tau - C/N] \geq P[\tau - \hat{\tau} < \varepsilon] \to 1,$$

because $\hat{\tau} \xrightarrow{P} \tau$. So we only need to show that $\hat{\tau}$ converges in probability to a positive number, whenever the null is violated.

We assume in the rest of the proof that the null hypothesis does not hold. We will prove $\hat{\tau} \xrightarrow{P} \tau > 0$ by verifying that the nine sufficient conditions stated further below are satisfied. For τ to be positive, it suffices that the numerator of both terms in the definition of $\hat{\tau}$ [expression (5.4)], converges to a positive number and that the sum of the (squared) numerators does, also.

Sufficient for $\hat{\mathcal{I}}_1 - \mathcal{I}_1 = \frac{1}{N^2}\sum_{t,s} a_{ts}a_{t+1,s+1} - \mathcal{I}_1 = o_p(1)$, $\hat{\mathcal{I}}_2 - \mathcal{I}_2 = \frac{1}{N^3}\sum_{t,s,u} a_{ts}a_{t+1,u} - \mathcal{I}_2 = o_p(1)$ and $\hat{\mathcal{I}}_3 - \mathcal{I}_3 = (\frac{1}{N^2}\sum_{t,s} a_{ts})^2 - \mu^2 = o_p(1)$ are (using $\mathcal{I}_3 = \mu^2$) (5.30) and (5.32), (5.29) and (5.33), and (5.29) and (5.31), respectively. $\hat{\gamma} - \hat{\mu}^2 \xrightarrow{P} c > 0$, is implied by (5.35) to (5.37).

Thus,

$$\begin{aligned}\hat{\tau} &\xrightarrow{P} \frac{1}{2(\gamma-\mu^2)^2}\left\{(\mathcal{I}_1-\mathcal{I}_2)^2+(\mathcal{I}_2-\mathcal{I}_3)^2\right\} \\ &= \frac{1}{2(\gamma-\mu^2)^2}\mathcal{I} > 0,\end{aligned}$$

by (5.34). Hence, (5.29) to (5.37) are sufficient for this Theorem to hold.

$$\sup_t \left|\frac{1}{N}\sum_s a_{ts} - Q_t\right| = o_p(1), \tag{5.29}$$

$$\sup_t \left|\frac{1}{N}\sum_s (a_{ts}a_{t+1,s+1} - R_t)\right| = o_p(1), \tag{5.30}$$

$$\frac{1}{N}\sum_t (Q_t - \mu) = o_p(1), \tag{5.31}$$

$$\frac{1}{N}\sum_t (R_t - \mathcal{I}_1) = o_p(1), \tag{5.32}$$

$$\frac{1}{N}\sum_t (Q_tQ_{t+1} - \mathcal{I}_2) = o_p(1), \tag{5.33}$$

$$\mathcal{I} > 0, \tag{5.34}$$

$$\hat{\gamma} - \gamma = o_p(1), \tag{5.35}$$

$$\hat{\mu} - \mu = o_p(1), \tag{5.36}$$

$$\gamma - \mu^2 > 0. \tag{5.37}$$

Expressions (5.29) and (5.30) are proved in Lemmas 5.1 and 5.2. Conditions (5.31) to (5.33) follow directly with the ergodic theorem. Because (5.29) holds, $\hat{\gamma}$ can be written as $\frac{1}{N}\sum_t Q_t^2 + o_p(1)$. The ergodic theorem gives that $\frac{1}{N}\sum_t Q_t^2 \xrightarrow{P} EQ_1^2 = \gamma$ and thus (5.35) holds. Also by (5.29), $\hat{\mu}$ can be written as $\frac{1}{N}\sum_t Q_t + o_p(1)$ and (5.31) then implies that (5.36) holds. Because $\gamma - \mu^2$ is the variance of Q_1, it is only zero when the distribution of Q_1 would be degenerate, which we excluded in Assumption 5.B. So (5.37) is satisfied, also.

We now only need to establish (5.34). The argument is basically the same as that used in Section 5.2. From (5.2) it follows that $\mathcal{I} \geq B^2/2$. It thus suffices to show that $B > 0$, if H_0 is

incorrect. Let $dF_\Delta(x,y) = dF_{12}(x,y) - dF(x)dF(y)$, for all x, y. Now

$$\begin{aligned} B &= \int a(x-w)a(y-z)dF_\Delta(x,y)dF_\Delta(w,z) \\ &= \int g(u)g(v)e^{i\{u(x-w)+v(y-z)\}}dF_\Delta(x,y)dF_\Delta(w,z)dudv \\ &= \int g(u)g(v)\cos(u(x-w)+v(y-z))dF_\Delta(x,y)dF_\Delta(w,z)dudv, \end{aligned}$$

where the second equality follows from Assumption 5.B and the last from the fact that g is even and symmetric. Because g is everywhere positive, we only have to show that $\int \cos(u(x-w)+v(y-z))dF_\Delta dF_\Delta$ is non-negative for all u, v and that it is not almost everywhere zero. So we look at

$$\begin{aligned} &\int \{\cos(ux+vy)\cos(uw+vz)+\sin(ux+vy)\sin(uw+vz)\}dF_\Delta(x,y)dF_\Delta(w,z) \\ &= \left(\int \cos(ux+vy)dF_\Delta(x,y)\right)^2 + \left(\int \sin(ux+vy)dF_\Delta(x,y)\right)^2 \\ &= \left|Ee^{i(uX_1+vX_2)} - Ee^{iuX_1}Ee^{ivX_1}\right|^2, \end{aligned}$$

which — being the squared norm of the difference of the characteristic functions of the joint and the product of the marginal distributions respectively — is only zero almost everywhere, if the difference of the joint and marginal distribution functions is zero, almost everywhere [cf. Lukacs (1970), Theorem 3.1.1], which would exclude serial dependence of order one. Thus, the theorem holds.

Q.E.D.

Proof of Theorem 5.2

In this proof we will rely heavily on standard U-statistic theory. Although we are dealing with $\{W_t\}$, where $W_t = (X_t, X_{t+1})$, which is a 1-dependent process rather than an i.i.d. process,

the results that are of interest to us still hold. For formal justification, we refer to Denker and Keller (1983), who prove the result used below for absolutely regular processes, a much weaker condition than k-dependence.

Applying Denker and Keller's result to $(\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2)$ and $(\hat{\mathcal{I}}_2 - \hat{\mathcal{I}}_3)$ yields (noting that — under the null — $Q_t = E[a_{ts}|X_t] = E[a_{st}|X_t]$)

$$\begin{aligned}
\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2 &= \frac{1}{N^3}\sum_t\sum_s\sum_u a_{ts}(a_{t+1,s+1} - a_{t+1,u}) \\
&= \frac{1}{N}\sum_t (Q_tQ_{t+1} - 2\mu Q_t + \mu^2) + O_p(N^{-1}), \\
\hat{\mathcal{I}}_2 - \hat{\mathcal{I}}_3 &= \frac{1}{N^4}\sum_t\sum_s\sum_u\sum_v a_{ts}(a_{t+1,u} - a_{u,v}) \\
&= \frac{1}{N}\sum_t (Q_tQ_{t+1} - 2\mu Q_t + \mu^2) + O_p(N^{-1}).
\end{aligned}$$

So under serial independence, $(\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2)$ and $(\hat{\mathcal{I}}_2 - \hat{\mathcal{I}}_3)$ differ by only $O_p(N^{-1})$ and $\sqrt{N}$ times each of these quantities therefore converges (in distribution) to the same random variable, λ_0, which is introduced further below. But, if $L_t = Q_tQ_{t+1} - 2\mu Q_t + \mu^2$, for all t, then $\{L_t\}$ is a series of 1-dependent strictly stationary random variables. The variance of $\frac{1}{\sqrt{N}}\sum_t L_t$ is

$$\begin{aligned}
&\frac{1}{N}\sum_t (EL_t^2 + 2E[L_tL_{t+1}]) \\
&= E[Q_1Q_2 - 2\mu Q_1 + \mu^2]^2 + 2E[(Q_1Q_2 - 2\mu Q_1 + \mu^2)(Q_2Q_3 - 2\mu Q_2 + \mu^2)] \\
&= (\gamma^2 - \mu^4) + 2(\mu^4 - \mu^2\gamma) = (\gamma - \mu^2)^2,
\end{aligned}$$

which can be — as we have seen in the proof to Theorem 5.1 — consistently estimated by $(\hat{\gamma} - \hat{\mathcal{I}}_3)^2$. The standard deviation of $\frac{1}{\sqrt{N}}\sum_t L_t$ is therefore $\gamma - \mu^2$, which is itself a variance and hence non-negative; it actually is positive, by Assumption 5.B. Thus, $\sqrt{N}\frac{\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2}{\hat{\gamma} - \hat{\mathcal{I}}_3} \xrightarrow{\mathcal{L}} \lambda_0$, and $\sqrt{N}\frac{\hat{\mathcal{I}}_2 - \hat{\mathcal{I}}_3}{\hat{\gamma} - \hat{\mathcal{I}}_3} \xrightarrow{\mathcal{L}} \lambda_0$, where $\lambda_0 \sim N(0,1)$. This implies that $N\left(\frac{\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2}{\hat{\gamma} - \hat{\mathcal{I}}_3}\right)^2 \xrightarrow{\mathcal{L}} \lambda_1$, and $N\left(\frac{\hat{\mathcal{I}}_2 - \hat{\mathcal{I}}_3}{\hat{\gamma} - \hat{\mathcal{I}}_3}\right)^2 \xrightarrow{\mathcal{L}} \lambda_1$, where $\lambda_1 \sim \chi_1^2$. Hence $N\hat{\tau} \xrightarrow{\mathcal{L}} \frac{1}{2}(\lambda_1 + \lambda_1) = \lambda_1 \sim \chi_1^2$.

Q.E.D.

Proof of Theorem 5.3

Let us add a superscript Y to the quantities defined in Section 5.3.1, when they are based on the proxy series $\{Y_{Nt}\}$ rather than on $\{X_t\}$. Because of the results we obtained in the proof to Theorem 5.1, it suffices to prove that

$$\hat{\mu}^Y - \hat{\mu} = o_p(1), \tag{5.38}$$

$$\hat{\gamma}^Y - \hat{\gamma} = o_p(1), \tag{5.39}$$

$$\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1 = o_p(1), \tag{5.40}$$

$$\hat{\mathcal{I}}_2^Y - \hat{\mathcal{I}}_2 = o_p(1), \tag{5.41}$$

$$\hat{\mathcal{I}}_3^Y - \hat{\mathcal{I}}_3 = o_p(1). \tag{5.42}$$

Before dealing with conditions (5.38) to (5.42) consider the following. We know that $Y_{Nt} - X_t = (\hat{\theta} - \theta_0)^T m'(Z_t; \theta_t) = o_p(1)$, uniformly in t, where θ_t lies between θ_0 and $\hat{\theta}$, by Assumption 5.C and the Mean Value Theorem. By again the Mean Value Theorem we get

$$a(Y_{Nt} - Y_{Ns}) - a(X_t - X_s) = \{(Y_{Nt} - X_t) - (Y_{Ns} - X_s)\}\, a'(\cdot) = o_p(1), \tag{5.43}$$

uniformly in t and s, because $Y_{Nt} - X_t = o_p(1)$, uniformly in t, by Assumption 5.C and because a' is bounded, by the same Assumption. Because

$$\hat{\mu}^Y - \hat{\mu} = \frac{1}{N^2} \sum_t \sum_s \{a(Y_{Nt} - Y_{Ns}) - a(X_t - X_s)\} = o_p(1),$$

conditions (5.38) and (5.42) are satisfied and by a nearly identical argument, so is (5.39). But

$$\begin{aligned} & a(Y_{Nt} - Y_{Ns}) a(Y_{N,t+1} - Y_{N,s+1}) - a(X_t - X_s) a(X_{t+1} - X_{s+1}) \\ &= \{a(Y_{Nt} - Y_{Ns}) - a(X_t - X_s)\} \{a(Y_{N,t+1} - Y_{N,s+1}) - a(X_{t+1} - X_{s+1})\} \\ &+ a(X_t - X_s) \{a(Y_{N,t+1} - Y_{N,s+1}) - a(X_{t+1} - X_{s+1})\} \end{aligned}$$

$$+ \quad a(X_{t+1} - X_{s+1}) \{a(Y_{Nt} - Y_{Ns}) - a(X_t - X_s)\} = o_p(1),$$

uniformly in t, s by the boundedness of a and by (5.43). This deals with (5.40); (5.41) can be dealt with in the very same way.

Q.E.D.

Proof of Theorem 5.4

We need to prove that under the null $\hat{\tau}^Y - \hat{\tau} = o_p(N^{-1})$ and sufficient conditions for this condition — aside from the results already obtained in (the proofs to) Theorems 5.2 and 5.3 — are

$$(\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1) = o_p(N^{-\frac{1}{2}}), \tag{5.44}$$

$$(\hat{\mathcal{I}}_2^Y - \hat{\mathcal{I}}_2) = o_p(N^{-\frac{1}{2}}), \tag{5.45}$$

$$(\hat{\mathcal{I}}_3^Y - \hat{\mathcal{I}}_3) = o_p(N^{-\frac{1}{2}}). \tag{5.46}$$

First we shall obtain some useful results, by means of the Mean Value Theorem. By Assumptions 5.C and 5.D, we have [writing $m_t' = m'(Z_t; \theta_0)$]

$$\begin{aligned} Y_{Nt} - X_t &= (\hat{\theta} - \theta_0)^T m_t' + \frac{1}{2}(\hat{\theta} - \theta_0)^T m''(Z_t; \theta_t)(\hat{\theta} - \theta_0) \\ &= (\hat{\theta} - \theta_0)^T m_t' + o_p(N^{-\frac{1}{2}}), \end{aligned}$$

where θ_t lies between θ_0 and $\hat{\theta}$. Then [using $a_{ts}' = a'(X_t - X_s)$]

$$\begin{aligned} & a(Y_{Nt} - Y_{Ns}) - a(X_t - X_s) \\ &= \{(Y_{Nt} - X_t) - (Y_{Ns} - X_s)\} a_{ts}' + \frac{1}{2}\{(Y_{Nt} - X_t) - (Y_{Ns} - X_s)\}^2 a''(\cdot) \\ &= (\hat{\theta} - \theta_0)^T (m_t' - m_s') a_{ts}' + o_p(N^{-\frac{1}{2}}), \end{aligned} \tag{5.47}$$

again by Assumptions 5.C and 5.D. Now

$$\begin{aligned}\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1 &= \frac{1}{N^2}\sum_t\sum_s (a_{ts}^Y a_{t+1,s+1}^Y - a_{ts}a_{t+1,s+1}) \\ &= \frac{1}{N^2}\sum_t\sum_s \{(a_{ts}^Y - a_{ts})(a_{t+1,s+1}^Y - a_{t+1,s+1}) \\ &+ a_{ts}(a_{t+1,s+1}^Y - a_{t+1,s+1}) + a_{t+1,s+1}(a_{ts}^Y - a_{ts})\}. \qquad (5.48)\end{aligned}$$

Only the last two terms on the right hand side in (5.48) are relevant, because

$$\begin{aligned}&(a_{ts}^Y - a_{ts})(a_{t+1,s+1}^Y - a_{t+1,s+1}) \\ &= (\hat{\theta}-\theta_0)^T(m_t' - m_s')a_{ts}'(\hat{\theta}-\theta_0)^T(m_{t+1}' - m_{s+1}')a_{t+1,s+1}' + o_p(N^{-\frac{1}{2}}) = o_p(N^{-\frac{1}{2}}),\end{aligned}$$

uniformly in t, s by (5.47) and Assumption 5.D. We have

$$a_{t+1,s+1}(a_{ts}^Y - a_{ts}) = (\hat{\theta}-\theta_0)^T a_{t+1,s+1}(m_t' - m_s')a_{ts}' + o_p(N^{-\frac{1}{2}}),$$

by (5.47). Combining these last two results with (5.48) and carrying out similar operations on $\hat{\mathcal{I}}_2^Y - \hat{\mathcal{I}}_2$ and $\hat{\mathcal{I}}_3^Y - \hat{\mathcal{I}}_3$ leads to

$$\begin{aligned}\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1 &\approx (\hat{\theta}-\theta_0)^T\frac{1}{N^2}\sum_t\sum_s\{a_{t+1,s+1}(m_t' - m_s')a_{ts}' + a_{ts}(m_{t+1}' - m_{s+1}')a_{t+1,s+1}'\}, \\ \hat{\mathcal{I}}_2^Y - \hat{\mathcal{I}}_2 &\approx (\hat{\theta}-\theta_0)^T\frac{1}{N^3}\sum_t\sum_s\sum_u\{a_{t+1,u}(m_t' - m_s')a_{ts}' + a_{ts}(m_{t+1}' - m_u')a_{t+1,u}'\}, \\ \hat{\mathcal{I}}_3^Y - \hat{\mathcal{I}}_3 &\approx (\hat{\theta}-\theta_0)^T\frac{1}{N^4}\sum_t\sum_s\sum_u\sum_v\{a_{uv}(m_t' - m_s')a_{ts}' + a_{ts}(m_u' - m_v')a_{uv}'\},\end{aligned}$$

where $\approx$ means that the omitted terms are $o_p(N^{-\frac{1}{2}})$. We now proceed to proving (5.44), where the proofs of (5.45) and (5.46) follow in identical fashion. On the basis of the last displayed equation we write

$$\begin{aligned}&(\hat{\mathcal{I}}_1^Y - \hat{\mathcal{I}}_1) \\ &= (\hat{\theta}-\theta_0)^T\frac{1}{N^2}\sum_t\sum_s\{a_{ts}(m_{t+1}' - m_{s+1}')a_{t+1,s+1}' + a_{t+1,s+1}(m_t' - m_s')a_{ts}'\}. \qquad (5.49)\end{aligned}$$

The expression on the right hand side, excluding $(\hat{\theta}-\theta_0)^T$, is a symmetric V-statistic. All conditions of Denker and Keller's (1983) Theorem 1 are satisfied, and hence the V-statistic

described above converges to the expectation of its V-statistic kernel; that is, to the expectation of its V-statistic kernel when W_t, W_s, W_u are mutually independent. Observe that X_t is also independent of X_{t+1}, by Assumption 5.C, but that Z_t need not be independent of Z_{t+1}, nor is Z_{t+1} necessarily independent of X_t. Note further that $Ea'_{ts} = -Ea'_{st} = 0$, by symmetry of a and because $\{X_t\}$ is i.i.d.. Thus, for the expression inside the curly brackets on the right hand side in (5.49) we have

$$\begin{aligned}
&E_I[a_{ts}(m'_{t+1} - m'_{s+1})a'_{t+1,s+1} + a_{t+1,s+1}(m'_t - m'_s)a'_{ts}] \\
&= E_I E_I[a_{ts}(m'_{t+1} - m'_{s+1})a'_{t+1,s+1} + a_{t+1,s+1}(m'_t - m'_s)a'_{ts}|X_{t+1}, X_{s+1}, X_t, X_s] \\
&= E_I[a_{ts}(E_I[m'_{t+1}|X_{t+1}] - E_I[m'_{s+1}|X_{s+1}])a'_{t+1,s+1}] \\
&+ E_I[a_{t+1,s+1}(E_I[m'_t|X_t] - E_I[m'_s|X_s])a'_{ts}] \\
&= 0,
\end{aligned}$$

by $Ea'_{ts} = 0$ and the fact that $E[m'_t|X_t]$ does not depend on X_t by Assumption 5.D, where E_I denotes the expectation where W_t, W_s, W_u are assumed independent. The above argument can be repeated for (5.45) and (5.46).

Q.E.D.

Proof of Theorem 5.5

Using the Mean Value Theorem we obtain

$$Y_{Nt} - X_t = \theta m'_t + \frac{1}{2}\theta^2 m''_t + \frac{1}{6}\theta^3 m'''(Z_t; \tilde{\theta}) \approx \theta m'_t + \frac{1}{2}\theta^2 m''_t,$$

uniformly in t, where $\tilde{\theta} \in [0, \theta]$, and where $\approx$ means that the lower order terms are $o_p(\theta^2)$, which

follows from condition (5.17). Using the above result we get by the Mean Value Theorem

$$\begin{aligned}
a^Y_{ts} - a_{ts} &= \{(Y_{Nt} - X_t) - (Y_{Ns} - X_s)\}a'_{ts} + \frac{1}{2}\{(Y_{Nt} - X_t) - (Y_{Ns} - X_s)\}^2 a''_{ts} \\
&+ \frac{1}{6}\{(Y_{Nt} - X_t) - (Y_{Ns} - X_s)\}^3 a'''(\cdot) \\
&\approx \theta(m'_t - m'_s)a'_{ts} + \frac{1}{2}\theta^2(m''_t - m''_s)a'_{ts} \\
&+ \frac{1}{2}\{\theta(m'_t - m'_s) + \frac{1}{2}\theta^2(m''_t - m''_s)\}^2 a''_{ts} \\
&+ \frac{1}{6}\{\theta(m'_t - m'_s) + \frac{1}{2}\theta^2(m''_t - m''_s)\}^3 a'''(\cdot) \\
&\approx \theta(m'_t - m'_s)a'_{ts} + \frac{\theta^2}{2}\{(m''_t - m''_s)a'_{ts} + (m'_t - m'_s)^2 a''_{ts}\},
\end{aligned}$$

uniformly in t, s, where the last $\approx$ follows with (5.15) and (5.16) and the fact that a'' and a''' are assumed bounded. Thus, we write

$$\begin{aligned}
&(a^Y_{ts} - a_{ts})(a^Y_{t+1,s+1} - a_{t+1,s+1}) + a_{ts}(a^Y_{t+1,s+1} - a_{t+1,s+1}) + a_{t+1,s+1}(a^Y_{ts} - a_{ts}) \\
&\approx \theta\{a_{ts}(m'_{t+1} - m'_{s+1})a'_{t+1,s+1} + a_{t+1,s+1}(m'_t - m'_s)a'_{ts}\} \\
&+ \frac{\theta^2}{2}\{2(m'_t - m'_s)a'_{ts}(m'_{t+1} - m'_{s+1})a'_{t+1,s+1} \\
&+ a_{ts}(m''_{t+1} - m''_{s+1})a'_{t+1,s+1} + a_{t+1,s+1}(m''_t - m''_s)a'_{ts} \\
&+ a_{ts}(m'_{t+1} - m'_{s+1})^2 a''_{t+1,s+1} + a_{t+1,s+1}(m'_t - m'_s)^2 a''_{ts}\} \\
&= \theta\mathcal{E}_{01}(t, s, t+1, s+1) + \theta^2\mathcal{E}_{02}(t, s, t+1, s+1),
\end{aligned}$$

uniformly in t, s, where the $\approx$ again follows with (5.15) and (5.16). Conditions (5.12) to (5.14), in conjunction with the absolute regularity condition, ensure that

$$\begin{aligned}
\frac{1}{N^2}\sum_t\sum_s \mathcal{E}_{01}(t, s, t+1, s+1) &- E_I[\mathcal{E}_{01}(t, s, t+1, s+1)] = O_p(N^{-\frac{1}{2}}), \\
\frac{1}{N^2}\sum_t\sum_s \mathcal{E}_{02}(t, s, t+1, s+1) &- E_I[\mathcal{E}_{02}(t, s, t+1, s+1)] = O_p(N^{-\frac{1}{2}}),
\end{aligned}$$

which follows from Theorem 1 of Denker and Keller (1983). Thus,

$$\hat{I}^Y_1 - \hat{I}_1 = \frac{1}{N^2}\sum_t\sum_s\{(a^Y_{ts} - a_{ts})(a^Y_{t+1,s+1} - a_{t+1,s+1})$$

$$+ \quad a_{ts}(a^Y_{t+1,s+1} - a_{t+1,s+1}) + a_{t+1,s+1}(a^Y_{ts} - a_{ts})\}$$

$$= \quad \mathcal{E}_1(\theta) + o_p(\theta^2) + O_p(\theta N^{-\frac{1}{2}}) = \mathcal{E}_1(\theta) + o_p(N^{-\frac{1}{2}}).$$

A similar procedure can be carried out for $(\hat{\mathcal{I}}^Y_2 - \hat{\mathcal{I}}_2)$, and $(\hat{\mathcal{I}}^Y_3 - \hat{\mathcal{I}}_3)$, such that (5.18) through (5.20) hold. One can in a similar fashion also show that $\hat{\gamma}^Y - \hat{\gamma} = o_p(1)$. By the conditions on θ, and by Assumption 5.E, $\tilde{\mathcal{E}}_{ij}$ exists for all $i, j = 1, 2, 3$. Now,

$$\begin{aligned} N(\hat{\mathcal{I}}^Y_1 - \hat{\mathcal{I}}^Y_2)^2 &= N\{(\hat{\mathcal{I}}_1 - \hat{\mathcal{I}}_2) + (\hat{\mathcal{I}}^Y_1 - \hat{\mathcal{I}}_1) + (\hat{\mathcal{I}}^Y_2 - \hat{\mathcal{I}}_2)\}^2 \\ &\xrightarrow{\mathcal{L}} \{(\gamma - \mathcal{I}_3)\lambda_0 + \tilde{\mathcal{E}}_{12}\}^2, \end{aligned}$$

where λ_0 is the $N(0,1)$ random variate of the proof to Theorem 5.2, which follows from that theorem and the above discussion. A similar argument can be applied to $N(\hat{\mathcal{I}}^Y_2 - \hat{\mathcal{I}}^Y_3)^2$, such that

$$\begin{aligned} N\hat{\tau}^Y &\xrightarrow{\mathcal{L}} \frac{\{(\gamma - \mathcal{I}_3)\lambda_0 + \tilde{\mathcal{E}}_{12}\}^2 + \{(\gamma - \mathcal{I}_3)\lambda_0 + \tilde{\mathcal{E}}_{23}\}^2}{2(\gamma - \mathcal{I}_3)^2} \\ &= \left(\lambda_0 + \frac{\tilde{\mathcal{E}}_{13}}{2(\gamma - \mathcal{I}_3)}\right)^2 + \left(\frac{\tilde{\mathcal{E}}_{12} + \tilde{\mathcal{E}}_{32}}{2(\gamma - \mathcal{I}_3)}\right)^2, \end{aligned}$$

or (5.22).

Q.E.D.

5.B Technical Lemmas

The first few lemmas are used in the proof to Theorem 5.1 and the assumptions made with respect to that theorem are assumed to hold here also.

Lemma 5.1

$$\sup_t \left| \frac{1}{N} \sum_s (a_{ts} a_{t+1,s+1} - R_t) \right| = o_p(1). \tag{5.50}$$

Proof:

We have by Assumption 5.B that an even and bounded everywhere non-negative function g exists such that $a(x) = \int g(u)\exp(iux)du$ and therefore

$$\begin{aligned}
&\sup_t \left| \frac{1}{N} \sum_s (a_{ts} a_{t+1,s+1} - R_t) \right| \\
&= \sup_t \left| \frac{1}{N} \sum_s \int g(u)g(v) \left(e^{i\{u(X_t - X_s) + v(X_{t+1} - X_{s+1})\}} - e^{i(uX_t + vX_{t+1})} E e^{-i(uX_s + vX_{s+1})} \right) \right| \\
&\leq \sup_t \int g(u)g(v) \left| e^{i(uX_t + vX_{t+1})} \right| \left| \frac{1}{N} \sum_s \left(e^{-i(uX_s + vX_{s+1})} - E e^{-i(uX_s + vX_{s+1})} \right) \right| . \quad (5.51)
\end{aligned}$$

Obviously, $\sup_x |\exp(ix)| = 1$. Using Jensen's inequality, we take the expectation of the last factor under the integral in (5.51) to obtain (with $w = [u, v]^T$)

$$\begin{aligned}
&E \left| \frac{1}{N} \sum_s \left(e^{-i(uX_s + vX_{s+1})} - E e^{-i(uX_s + vX_{s+1})} \right) \right| \\
&= \sqrt{\frac{1}{N^2} E \left[\sum_s \left(e^{-iw^T W_s} - E e^{-iw^T W_1} \right) \sum_r \left(e^{iw^T W_r} - E e^{iw^T W_1} \right) \right]} \\
&\leq \frac{1}{N} \sqrt{\sum_s \sum_r \left(\mathrm{Cov}[\cos(w^T W_s), \cos(w^T W_r)] + \mathrm{Cov}[\sin(w^T W_s), \sin(w^T W_r)] \right)} \\
&\leq \sqrt{\frac{2}{N^2} \sum_s \sum_r \alpha(|s - r|)} = O\left(\sqrt{N^{-1} \sum_s \alpha(s)} \right),
\end{aligned}$$

by Assumption 5.A. Also by Assumption 5.A, $\alpha(t) \to 0$, as $t \to \infty$ and therefore, by Kronecker's Lemma, (5.50) holds.

Q.E.D.

Lemma 5.2

$$\sup_t \left| \frac{1}{N} \sum_s (a_{ts} - Q_t) \right| = o_p(1).$$

Proof:

Follows with an argument similar to that of the proof of Lemma 5.1.

Q.E.D.

5.C Higher Order Alternatives

In this part of the Appendix, we indicate how to extend the proposed test to higher order alternatives. As most steps are similar to those taken earlier on in the paper, we will be very brief here.

If we are to test for serial independence against a serial dependence of order $J-1$ alternative, we wish to test whether $\psi(u) = E[\exp(i\sum_{j=1}^{J} u_j X_j)] - E_I[\exp(i\sum_{j=1}^{J} u_j X_j)] = 0$, for all $u \in \Re^J$, where E_I is the expectation under independence of $(X_1, \ldots, X_J)$. Let $dF_\Delta(x) = dF_J(x) - \prod_{j=1}^{J} dF(x_j)$, where F_J is the joint distribution function of $(X_1, \ldots, X_J)$. Then

$$\begin{aligned} |\psi(u)|^2 &= \left(\int \cos(u^T x) dF_\Delta(x)\right)^2 + \left(\int \sin(u^T x) dF_\Delta(x)\right)^2, \\ &= \int \cos(u^T(x-y)) dF_\Delta(x) dF_\Delta(y). \end{aligned}$$

Let $\tilde{g}(u) = \prod_{j=1}^{J} g(u_j)$, $\tilde{a}(x) = \prod_{j=1}^{J} a(x_j)$, with $a(x_1) = \int g(u_1)\cos(x_1 u_1) du_1$. Then

$$B^{(J)} = \int \tilde{g}(u)|\psi(u)|^2 du = \int \tilde{a}(x-y) dF_\Delta(x) dF_\Delta(y).$$

Let $\mathcal{I}_1^{(J)} = \int \tilde{a}(x-y) dF_J(x) dF_J(y)$, $\mathcal{I}_2^{(J)} = \int \tilde{a}(x-y) dF_J(x) \prod_{j=1}^{J} dF(y_j)$, $\mathcal{I}_3^{(J)} = \int \tilde{a}(x-y) \prod_{j=1}^{J} dF(x_j) dF(y_j)$, and $\mathcal{I}^{(J)} = \{(\mathcal{I}_1^{(J)} - \mathcal{I}_2^{(J)})^2 + (\mathcal{I}_2^{(J)} - \mathcal{I}_3^{(J)})^2\}/2$. Then, once again $\mathcal{I}^{(J)}$ is greater than zero if and only if $B^{(J)}$ is. The $\mathcal{I}_i^{(J)}$'s can be consistently estimated by the $\hat{\mathcal{I}}_i^{(J)}$'s defined in (5.24) to (5.26).

Suppose now that the null hypothesis holds Again using the projection method for U-statistics, we obtain (where again $Q_t = E[a_{ts}|X_t]$, $t \neq s$)

$$\begin{aligned} \hat{\mathcal{I}}_1^{(J)} - \hat{\mathcal{I}}_2^{(J)} &= \frac{1}{N}\sum_t \left\{ \prod_{j=1}^{J} Q_{t+j-1} - J\mu^{J-1} Q_t + (J-1)\mu^J \right\} + O_p(N^{-1}), \\ \hat{\mathcal{I}}_2^{(J)} - \hat{\mathcal{I}}_3^{(J)} &= \frac{1}{N}\sum_t \left\{ \prod_{j=1}^{J} Q_{t+j-1} - J\mu^{J-1} Q_t + (J-1)\mu^J \right\} + O_p(N^{-1}), \end{aligned}$$

and hence [noting that the summations on the right hand side in the last two displayed equations

are both $O_p(N^{-\frac{1}{2}})$]

$$
\begin{aligned}
&(\hat{\mathcal{I}}_1^{(J)} - \hat{\mathcal{I}}_2^{(J)})^2 + (\hat{\mathcal{I}}_2^{(J)} - \hat{\mathcal{I}}_3^{(J)})^2 \\
&= 2\left[\frac{1}{N}\sum_t \left\{\prod_{j=1}^{J} Q_{t+j-1} - J\mu^{J-1}Q_t + (J-1)\mu^J\right\}\right]^2 + O_p(N^{-\frac{3}{2}}).
\end{aligned}
$$

Obviously

$$
\begin{aligned}
&\frac{1}{N}\sum_t \left\{\prod_{j=1}^{J} Q_{t+j-1} - J\mu^{J-1}Q_t + (J-1)\mu^J\right\} \\
&\xrightarrow{\mathcal{L}} N\left(0, E[V_{11}+V_{21}+V_{31}]^2 + 2\sum_{s=1}^{J-1} E[(V_{11}+V_{21}+V_{31})(V_{1,1+s}+V_{2,1+s}, V_{3,1+s})]\right),
\end{aligned}
$$

where V_{is} denotes the i-th term in curly brackets on the left hand side in the last displayed equation with t replaced by s. Note that for $l = 0, \ldots, J-1$, $E[V_{11}V_{1,1+l}] = \mu^{2l}\gamma^{J-l}$, $E[V_{11}V_{2,1+l}] = -J\mu^{2J-2}\gamma$, $E[V_{11}V_{3,1+l}] = (J-1)\mu^{2J}$, $E[V_{21}V_{1,1+l}] = -J\mu^{2J}I(l \neq 0) - J\mu^{2J-2}\gamma I(l = 0)$, $E[V_{21}V_{2,1+l}] = J^2\mu^{2J}I(l \neq 0) + J^2\mu^{2J-2}\gamma I(l = 0)$, $E[V_{21}V_{3,1+l}] = -J(J-1)\mu^{2J}$, $E[V_{31}V_{1,1+l}] = (J-1)\mu^{2J}$, $E[V_{31}V_{2,1+l}] = -J(J-1)\mu^{2J}$, $V_{31}V_{3,1+l} = (J-1)^2\mu^{2J}$. Therefore, the variance parameter of the normal distribution in the last displayed equation is

$$
\begin{aligned}
&\gamma^J + J^2\mu^{2J-2}\gamma + (J-1)^2\mu^{2J} - 2J\mu^{2J-2}\gamma + 2(J-1)\mu^{2J} - 2J(J-1)\mu^{2J} \\
&+ 2\sum_{l=1}^{J-1}\{\mu^{2l}\gamma^{J-l} + J^2\mu^{2J} + (J-1)^2\mu^{2J} - J\mu^{2J-2}\gamma + (J-1)\mu^{2J} - J\mu^{2J} \\
&- J(J-1)\mu^{2J} + (J-1)\mu^{2J} - J(J-1)\mu^{2J}\} \\
&= \gamma^J + (J^2-4J)\mu^{2J-2}\gamma + (-J^2+4J-3)\mu^{2J} + 2\frac{\gamma^J\mu^2 - \gamma\mu^{2J}}{\gamma-\mu^2} \\
&= \frac{\gamma^{J+1} + \mu^2\gamma^J + J(J-4)\mu^{2J-2}\gamma^2 + (-2J^2+8J-5)\mu^{2J}\gamma + (J^2-4J+3)\mu^{2J+2}}{\gamma-\mu^2},
\end{aligned}
$$

which can be estimated consistently by (5.27). Thus (5.28) holds.

Chapter 6

Summary and Conclusions

In this thesis, we have discussed two fairly different issues. In Part I, we discussed ways of pooling nonparametric estimates for regression functions with a similar shape. Part II discussed nonparametric tests for serial independence against serial dependence of fixed and finite order.

In Chapter 2, we found that it is possible to improve the accuracy, measured in terms of the asymptotic mean squared error, of nonparametric kernel estimates for regression functions with a similar shape. Numerical simulations suggest that success or failure of the procedure is context-sensitive. Pooling is applied at each individual point. It would be of interest to see if a global pooling rule would yield qualitatively different results in practice.

In Chapter 4, we extended Robinson's (1991a) entropy based test for serial independence to the case where the observations have infinite support. We also proved that the results still hold when nuisance parameters are present. Although the test has a convenient limiting distribution, the limiting distribution should not be used in practice, because there is a sizeable disparity between the actual and nominal size.

Chapter 5 discusses a new test for serial independence based on characteristic functions. The

test has convenient theoretical properties that held up well in Monte Carlo experiments. The only point of concern is the required choice of an input parameter, i.e. the kernel. This test should be a useful tool in time series analysis, particularly in nonlinear time series analysis.

Appendix A

Miscellaneous

In this appendix, we review some basic results used elsewhere.

A.1 Mean Value Theorem

If m is n times differentiable on an open convex set including $x+d$ and x, then the mean value theorem implies that

$$m(x+d) - m(x) = \sum_{i=0}^{n-1} \frac{1}{i!} \left(\sum_{j=0}^{J} d_j D_j \right)^i m_{|x} + \frac{1}{n!} \left(\sum_{j=0}^{J} d_j D_j \right)^n m_{|(x;x+d)}, \tag{A.1}$$

where $(x; x+d)$ denotes some point between x and $x+d$.

A.2 Products

It is easy to see that for any finite positive i, l, and any $z \in \Re^i$,

$$\left(\sum_{j_0=1}^{l} z_{j_0} \right)^i = \sum_{j_1} \cdots \sum_{j_{l-1}} \Gamma_{j_1,\ldots,j_l} \prod_{g=1}^{l} z_j^{j_g}, \tag{A.2}$$

where the sum over j_g runs from 0 to $i-j_1-\cdots-j_{g-1}$, for $1 \leq g < l$, and $j_l = i-j_1-\cdots-j_{l-1}$, and where the Γ's are positive constants. For instance, $(z_1+z_2)^3 = \sum_{j_1=0}^{3} \frac{3!}{j_1!(3-j_1!)} z_1^{j_1} z_2^{3-j_1}$.

Appendix B

Tables and Figures

B.1 Figures of Chapter 2

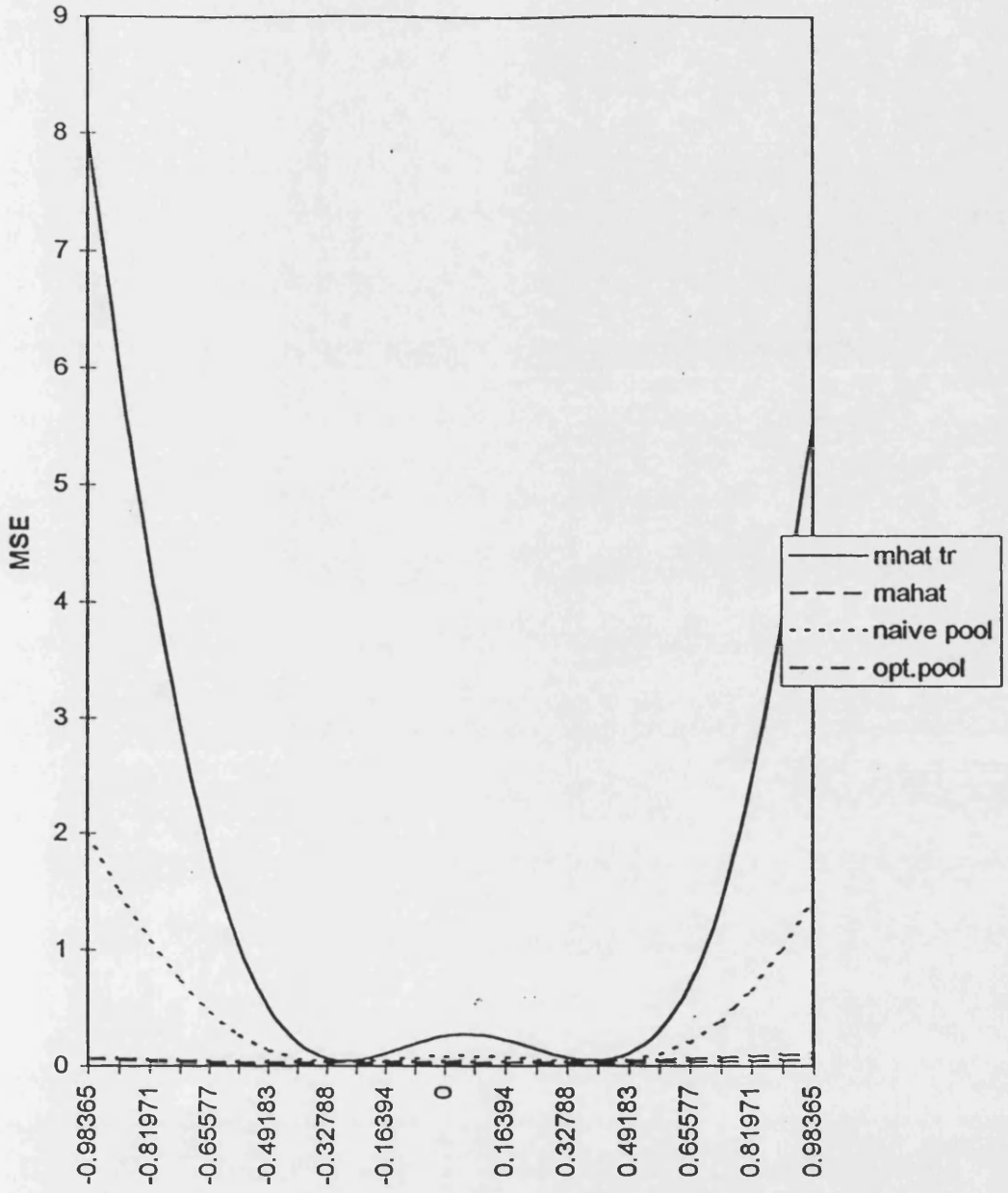

Figure 1: MSE Exponential Form, std.dev. (0.5,2.0)

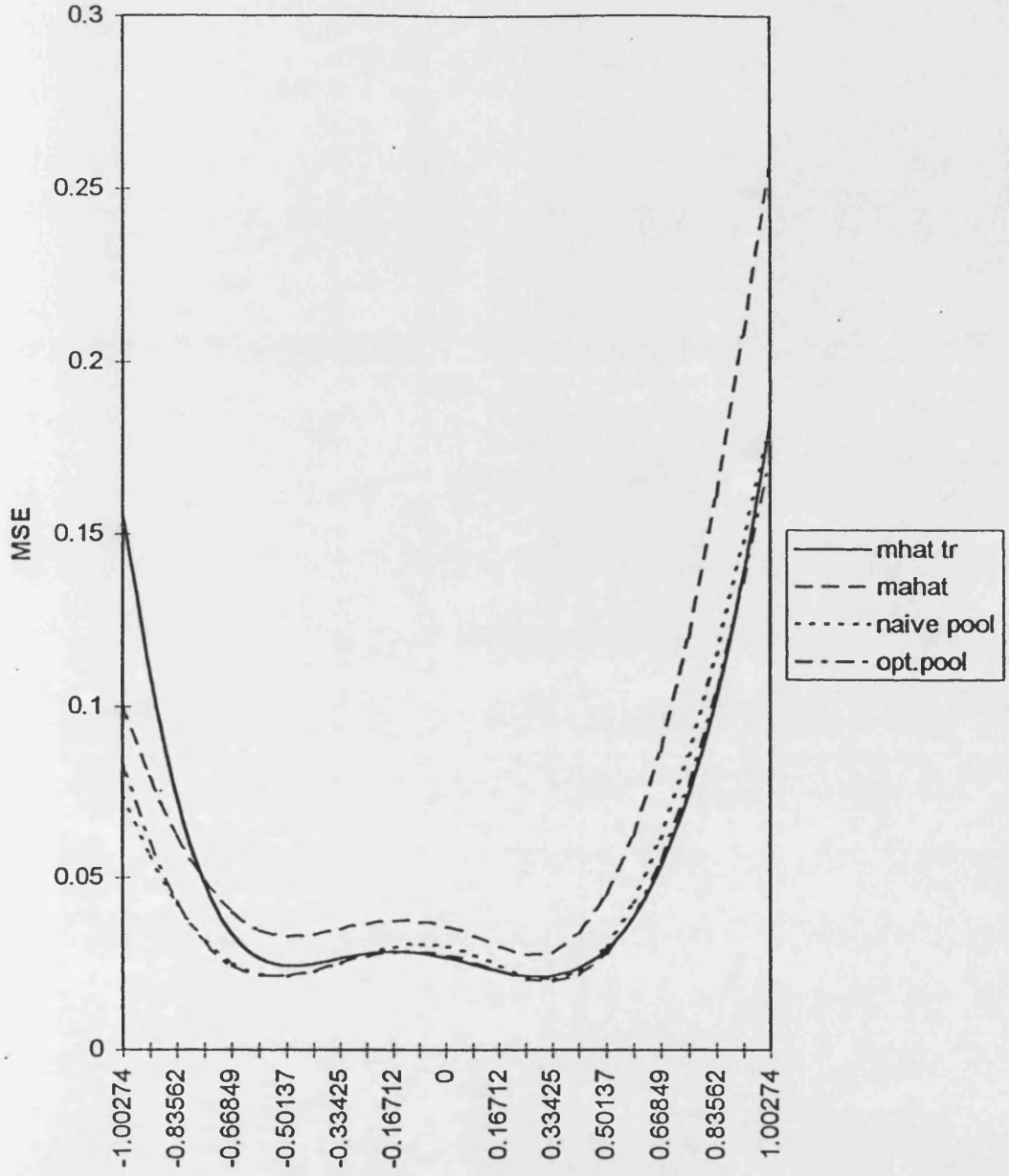

Figure 2: MSE Cosine Form, std.dev. (0.5,2.0)

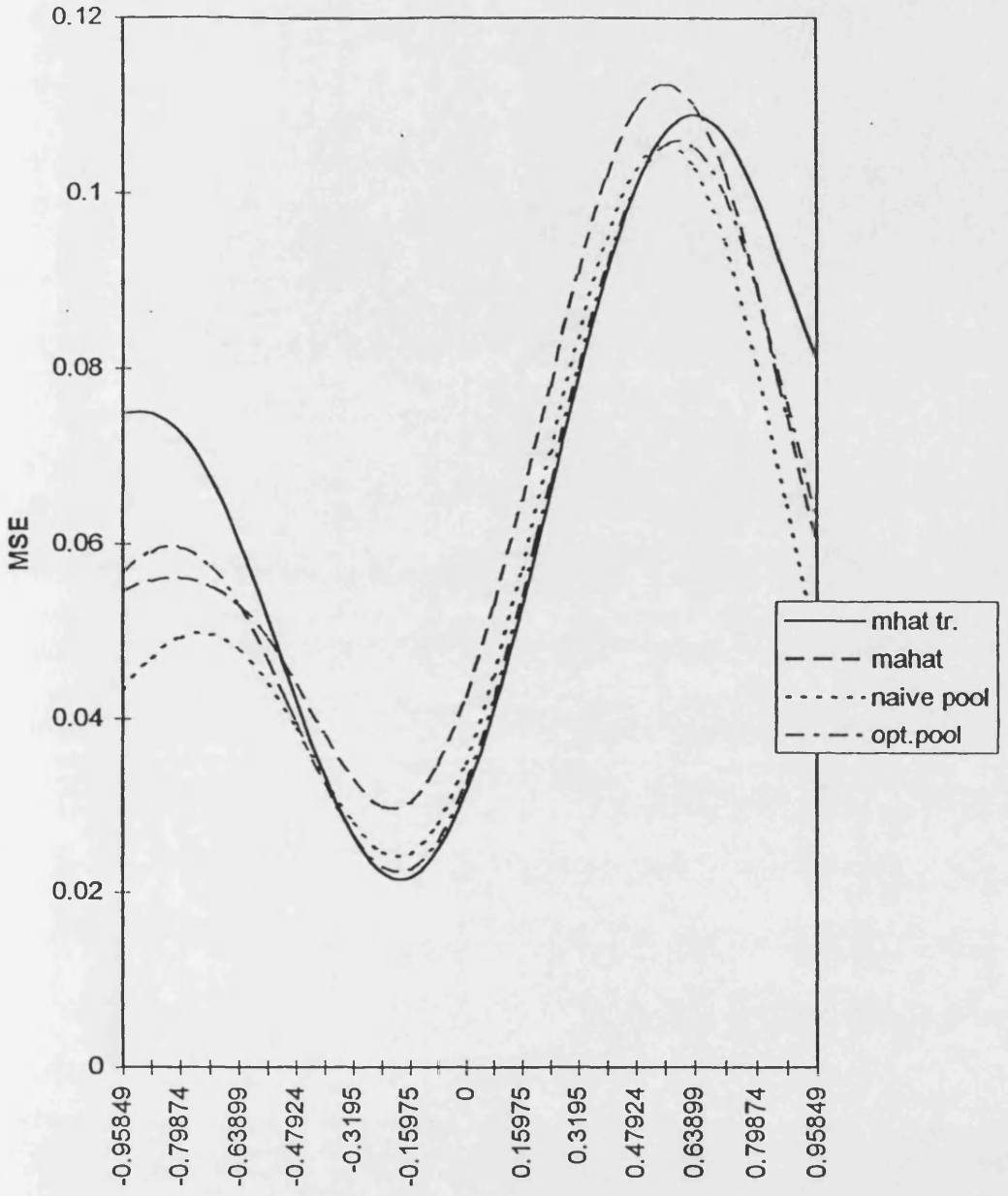

Figure 3: MSE Sine Form, std.dev. (0.5,2.0)

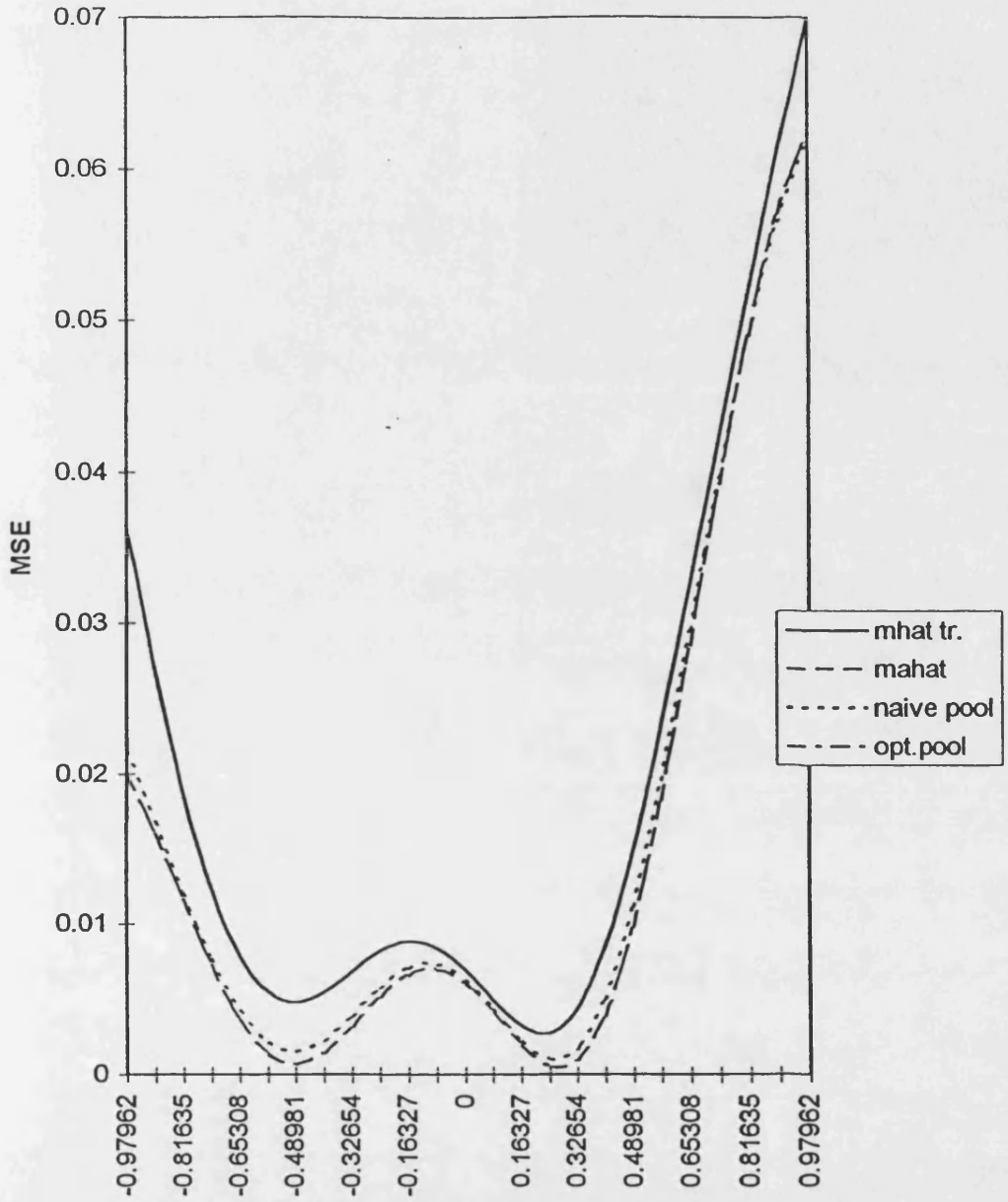

Figure 4: MSE Exponential Form, std.dev. (0.5,0.2)

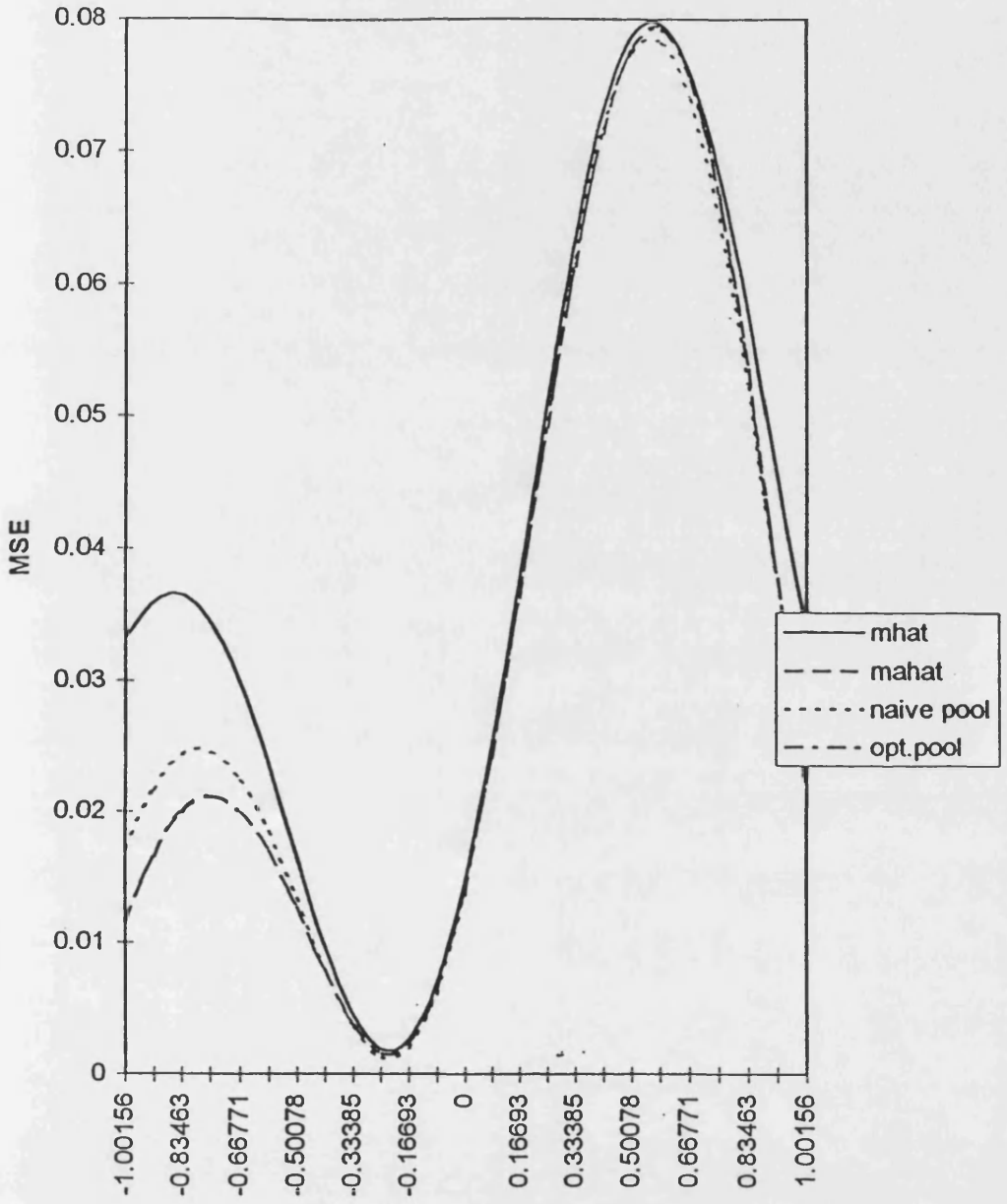

Figure 5: MSE Cosine Form, std.dev. (0.5,0.2)

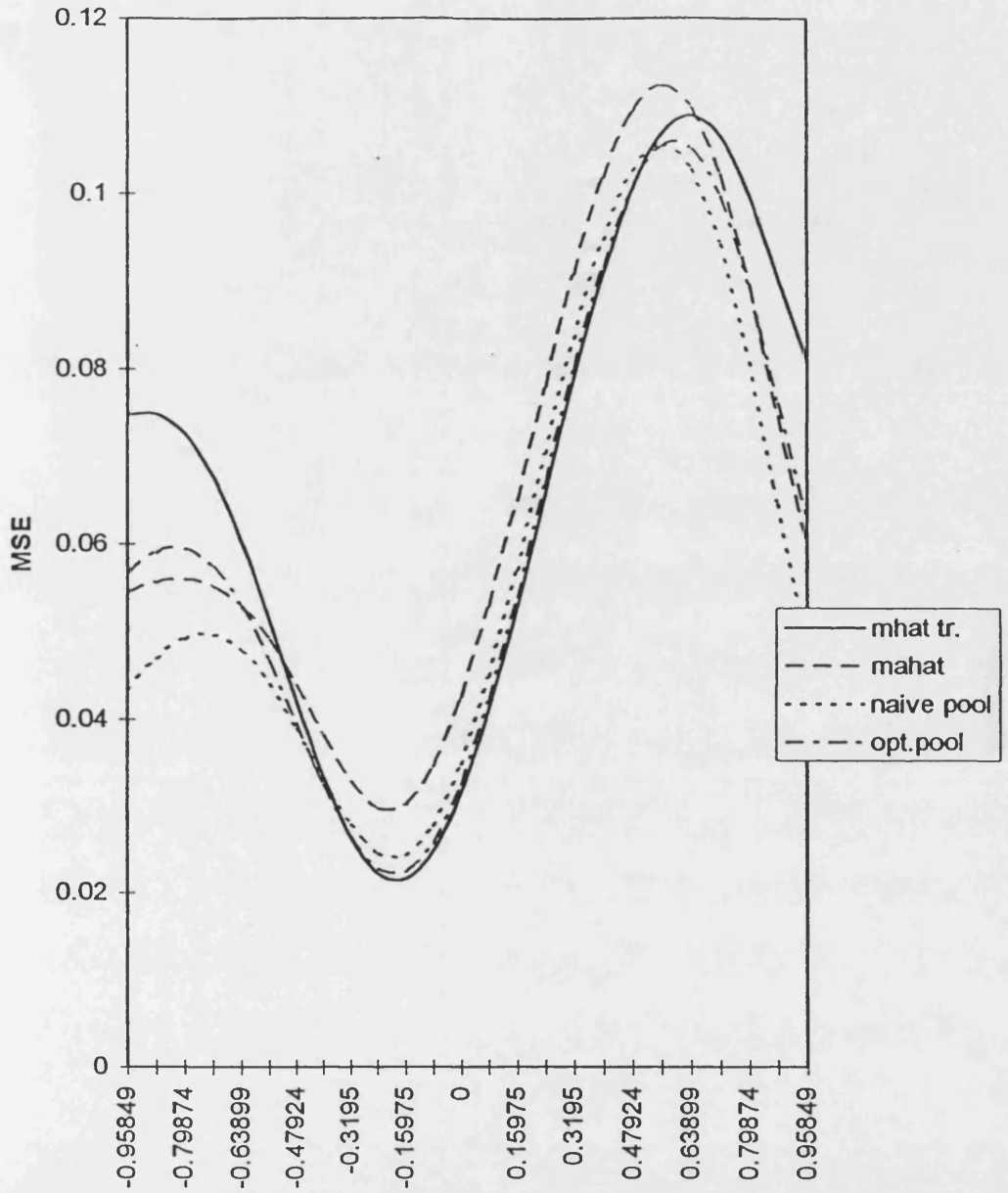

Figure 6: MSE Sine Form, std.dev. (0.5,0.2)

B.2 Tables of Chapter 4

$h = N^{-\frac{1}{3}}$			
γ	90%	95%	99%
0.100	60.56	63.42	66.52
0.200	30.30	31.76	33.64
0.400	15.49	16.26	17.15
0.800	8.26	8.73	9.23
1.600	4.88	5.27	5.66
3.200	3.37	3.73	4.21
6.400	2.61	2.97	3.44
$h = N^{-\frac{1}{5}}$			
γ	90%	95%	99%
0.100	83.58	85.98	89.24
0.200	41.80	43.03	44.46
0.400	21.20	21.98	22.71
0.800	11.13	11.58	12.06
1.600	6.29	6.64	7.05
3.200	3.99	4.36	4.76
6.400	2.83	3.15	3.56
$h = 2N^{-\frac{1}{5}}$			
γ	90%	95%	99%
0.100	117.18	119.29	121.97
0.200	58.64	59.78	61.25
0.400	29.56	30.18	30.96
0.800	15.29	15.70	16.17
1.600	8.30	8.63	9.07
3.200	4.90	5.22	5.58
6.400	3.21	3.54	3.95

Table 1: Entropy Based Test Statistic: Quantiles under the Null

B.3 Tables of Chapter 5

	Fraction Rejected	95 % Quantile	Fraction Rejected	97.5 % Quantile	Fraction Rejected	99 % Quantile
N=25	0.093	(5.71)	0.062	(8.77)	0.041	(14.1)
N=100	0.054	(3.98)	0.032	(5.57)	0.017	(7.94)
N=250	0.052	(3.90)	0.028	(5.22)	0.013	(7.01)
Asymptotic	0.050	(3.84)	0.025	(5.02)	0.010	(6.63)

Table 2: Size, nominal versus actual for 100 and 250 observations based on 8192 replications, with the corresponding quantiles between brackets.

Model	Statistics						
	CFG	CFE	CFA	CDT	EDF	ENT	CT
AR 0.2	0.119	0.136	0.085	0.110	0.372	0.147	0.507
AR 0.5	0.891	0.891	0.847	0.824	0.989	0.923	0.998
MA 0.2	0.117	0.129	0.082	0.105	0.332	0.125	0.460
MA 0.5	0.662	0.666	0.570	0.551	0.946	0.792	0.987
ARCH 0.2	0.310	0.295	0.244	0.295	0.065	0.244	0.083
ARCH 0.5	0.784	0.755	0.720	0.771	0.101	0.325	0.172
INVARCH 2.0	0.207	0.178	0.115	0.186	0.037	0.042	0.016
INVARCH 5.0	0.488	0.429	0.327	0.423	0.037	0.066	0.009
NLMA 0.5	0.265	0.254	0.185	0.244	0.070	0.140	0.088
NLMA 0.9	0.509	0.489	0.401	0.474	0.085	0.259	0.124
TAR -0.5 0.5	0.643	0.664	0.604	0.597	0.744	0.616	0.526
TAR 0 0.5	0.736	0.750	0.666	0.667	0.896	0.796	0.879
BLM 0 1.0	0.998	0.997	0.998	0.998	0.657	0.614	0.532
BLM 0.2 0.3	0.706	0.656	0.602	0.645	0.633	0.644	0.837
BLM 0 0.5	0.927	0.895	0.876	0.906	0.407	0.590	0.591
CM 1.0	0.883	0.904	0.834	0.849	0.945	0.857	0.620
EM 0.5 -0.11	0.524	0.531	0.417	0.416	0.902	0.614	0.902

The acronyms used for the tests mean:
CFG=Characteristic Function Test, Gaussian g
CFE=Idem, Double Exponential g
CFA=Idem, Asymmetric h
ENT=Entropy Based Test
EDF=Empirical Distribution Function Test
CT=Correlation Test
CDT=Correlation Dimension Test
AR=AutoRegressive Model
MA=Moving Average Model
ARCH=AutoRegressive Conditional Heteroskedasticity Model
INVARCH=Inverted ARCH Model
NLMA=NonLinear Moving Average Model
TAR=Treshold AutoRegressive Model
BLM=Bilinear Model
CM=Cosine Model
EM=Exponential Model

Table 3: Power Comparison; 100 observations, 8192 replications, 5% significance.

Series	Test Statistic Value
daily \$/DM	7.07
weekly \$/DM	60.14
monthly \$/DM	4.69
daily \$/JY	11.47
weekly \$/JY	35.08
monthly \$/JY	9.43
daily \$/SF	9.39
weekly \$/SF	38.37
monthly \$/SF	8.47
daily BP/\$	28.44
weekly BP/\$	33.43
monthly BP/\$	16.17
Daily: 946 (945) observations	
Weekly: 600 (599) observations	
Monthly: 138 (137) observations	

Table 4: Testing the Random Walk Hypothesis

References

Ahmad, I.A. and P.-E. Lin (1976a), Nonparametric Estimation of a Density Functional, Technical Report, Florida State University.

Ahmad, I.A. and P.-E. Lin (1976b), A Nonparametric Estimation of the Entropy for Absolutely Continuous Distributions, IEEE Transformations of Information Theory 22, 372–350.

Aitchinson, J. and S.D. Silvey (1958), Maximum Likelihood Estimation of Parameters Subject to Restraints, Annals of Mathematical Statistics 29, 813–828.

Aitchinson, J. and S.D. Silvey (1960), Maximum Likelihood Estimation Procedures and Associated Tests of Significance, Journal of the Royal Statistical Society B 22, 154–171.

Amemiya, T. (1985), Advanced Econometrics, Basil Blackwell.

Bank of England, Quarterly Bulletin, Selected Issues (Bank of England, London).

Andrews, D.W.K. (1984), Non-Strong Mixing Autoregressive Processes, Journal of Applied Probability 21, 930–934.

Blum, J.R., J. Kiefer, and M. Rosenblatt (1961), Distribution Free Tests of Independence Based on the Sample Distribution Function, Annals of Mathematical Statistics 32,

485–498.

Barbe, P. (1990), Estimation Non Paremètrique de L'Entropie et de l'Information de Kullback, preprint, Institut National de la Statistique et des Etudes Economiques.

Bartels, R. (1982), The Rank Version of von Neumann's Ratio Test for Randomness", Journal of the American Statistical Association 77, 40–46.

Borges da Silveira Filho, G. (1991), Contributions to Strong Approximations in Time Series with Applications in Nonparametric Statistics and Functional Limit Theorems, Ph.D. thesis (University of London).

Box, G.E.P. and D.A. Pierce (1970), Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models, Journal of the American Statistical Association 65, 1509–1526.

Brock, W.A., Dechert, W.D. and J.A. Scheinkman (1987), A Test for Independence Based on the Correlation Dimension, preprint (University of Wisconsin, Madison WI).

Brock, W.A., Dechert, W.D., Scheinkman, J.A. and B.D. LeBaron (1987), A Test for Independence Based upon the Correlation Dimension, preprint (University of Wisconsin, Madison WI).

Buja, A., Hastie, T. and R. Tibshirani (1989), Linear Smoothers and Additive Models, Annals of Statistics 17, 453–555.

Chan, N.H. and L.T. Tran (1992), Nonparametric Tests for Serial Dependence, Journal of Time Series Analysis 13, 19–28.

Chow, Y.S. (1960), A Martingale Inequality and a Law of Large Numbers, Proceedings of the Mathematical Society 11, 107–111.

Chow, Y.S. (1967), On a Strong Law for Martingales, Mathematical Statistics 38, 610–611.

Cochrane, D. and G.H. Orcutt (1949), Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms, Journal of the American Statistical Association 33, 32–61.

Cooley, J.W. and J.W. Tukey (1965), An Algorithm for Machine Calculation of Complex Fourier Series, Mathematics of Computation 19, 297–301.

Csörgö, S. (1985), Testing for Independence by the Empirical Characteristic Function, Journal of Multivariate Analysis 16, 290–299.

Davidson, J. (1992), Conditions for Strong and Uniform Mixing in Linear Processes, preprint (London School of Economics).

Davidson, R. and J.G. MacKinnon (1993), Estimation and Inference in Econometrics (Oxford University Press, New York).

de Wet, T. (1980), Cramér-von Mises Tests for Independence, Journal of Multivariate Analysis 10, 38–50.

Dechert, W.D. (1992), An Application of Chaos Theory to Stochastic and Deterministic Observations, preprint (University of Houston, Houston).

Delgado, M. (1993) Testing Serial Independence Using the Sample Distribution Function, preprint.

Denker, M. and G. Keller (1983), On U-Statistics and Von Mises Statistics for Weakly Dependent Processes, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 64, 502–522.

Diewert, W.E. and T.J. Wales (1993), A "New" Approach to the Smoothing Problem, Discussion Paper 93–44 (University of British Columbia, Vancouver, Canada).

Drost, F.C. and B.J.M. Werker (1993), A Note on Robinson's Test of Independence, preprint (Tilburg University, Tilburg).

Dudley, R.M. (1989), Real Analysis and Probability (Wadsworth & Brooks/Cole, Pacific Grove CA).

Dufour, J.-M. (1981), Rank Tests for Serial Dependence, Journal of Time Series Analysis 2, 117–127.

Durbin, J. (1970), Testing for Serial Correlation in Least Squares Regression when some of the Regressors are Lagged Dependent Variables, Econometrica 38, 410–421.

Durbin, J. and G.S. Watson (1950), Testing for Serial Correlation in Least Squares Regression I, Biometrika 37, 409–428.

Durbin, J. and G.S. Watson (1951), Testing for Serial Correlation in Least Squares Regression II, Biometrika 38, 159–178.

Durbin, J. and G.S. Watson (1971), Testing for Serial Correlation in Least Squares Regression III, Biometrika 58, 1–19.

Engle, R. (1982), Autoregressive Conditional Heteroskedasticity with estimates of the variance of United Kingdom inflation, Econometrica 50, 987–1007.

Epanechnikov, V. (1969), Nonparametric Estimates of a Multivariate Probability Density, Theory of Probability and its Applications 14, 153–158.

Gasser, T., Müller, H.-G., Köhler, W., Molinari, L., and A. Prader (1984), Annals of Statistics 12, 210–229.

Gasser, T., Müller, H.G. and V. Mammitzsch (1985),

Kernels for Nonparametric Curve Estimation, Journal of the Royal Statistical Society, Series B, 47, 238–252.

Godfrey, L.G. (1978a), Testing against Autoregressive and Moving Average Error Models When the Regressors Include Lagged Dependent Variables, Econometrica 46, 1293–1302.

Godfrey, L.G. (1978b), Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables, Econometrica 46, 1303–1310.

Godfrey, L.G. (1988), Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches (Cambridge University Press, Cambridge MA).

Gorodetskii, V.V. (1977), On the Strong Mixing Property for Linear Sequences, Theory of Probability and its Applications 22, 411–413.

Guerre, E. (1991), Estimating Kullback Contrast by the Kernel Method: Asymptotic Distribution, preprint (ENSAE, Paris).

Hall, P. and M. Titterington (1988), On Confidence Bands in Nonparametric Density Estimation and Regression, Journal of Multivariate Analysis, 27, 228–254.

Hallin, M., J.-F. Ingenbleek, and M.L. Puri (1985), Linear Serial Rank Test for Randomness against ARMA Alternatives, Annals of Statistics 12, 1156–1181.

Hallin, M. and G. Mélard (1988), Optimal Rank Based Tests for Randomness against First-Order Serial Dependence, Journal of the American Statistical Association 83, 1117–1122.

Hallin, M. and M.L. Puri (1988), Optimal Rank-Based Procedures for Time Series Analysis: Testing an ARMA model against other ARMA models, Annals of Statistics 16,

403–432.

Hannan, E.J. (1970), Multiple Time Series (Wiley, New York).

Härdle, W. (1987), Resistant Smoothing Using the Fast Fourier Transform, AS 222, Applied Statistics 36, 104–111.

Härdle, W. (1990), Applied Nonparametric Regression (Cambridge University Press, Cambridge MA).

Härdle, W., Hall, P. and H. Ichimura (1993), Optimal Smoothing in Single Index Models, Annals of Statistics 21, 157–178.

Härdle, W. and J.S. Marron (1990), Semiparametric Comparison of Regression Curves, Annals of Statistics, 18, 63–89.

Hidalgo, F.J. (1992), Adaptive Semiparametric Estimation in the Presence of Autocorrelation of Unknown Form, Journal of Time Series Analysis 13, 47–78.

Hidalgo, F.J. (1993), A Nonparametric Test for Structural Stability, preprint.

Hinich, M.J. (1982), Testing for Gaussianity and Linearity of a Stationary Time Series, Journal of Time Series Analysis 3, 169–176.

Hoeffding (1958), A Nonparametric Test of Independence, Annals of Mathematical Statistics 30, 420–447.

Ibragimov, I.A. and Y. Linnik (1971), Independent and Stationary Sequences of Random Variables (Wolters-Noordhoff, Groningen).

Ibragimov, I.A. and I.A. Rozanov (1978), Gaussian Random Processes (Springer Verlag, New York).

Joe, H. (1989), Relative Entropy Measures of Multivariate Dependence, Journal of the American Statistical Association 84, 157–163.

Kendall, M.G. and A. Stuart (1967), The Advanced Theory of Statistics, volume 2, 2nd edition.

Kneip, A. and T. Gasser (1988), Convergence and Consistency Results for Self-Modeling Nonlinear Regression, Annals of Statistics 16, 82–112.

Knoke, J.D. (1977), Testing for Randomness Against Autocorrelation: Alternative Tests, Biometrika 64, 523–329.

Kullback, S. (1959), Information Theory and Statistics (Wiley, New York).

Kullback, S. and R.A. Leibler (1951), On Information and Sufficiency, Annals of Mathematical Statistics 22, 79–86.

Lafontaine, F. and K.J. White (1986), Obtaining any Wald Statistic You Want, Economics Letters 21, 35–40.

Lawton, W.H., E.A. Sylvestre and M.S. Maggio (1972), Self-Modelling Nonlinear Regression, Technometrics 14, 513–532.

Ljung, G.M. and G.E. Box (1978), On a Measure of Lack of Fit in Time Series Models, Biometrika 65, 297–303.

Lukacs, E. (1970), Characteristic Functions, second edition (Griffin, London).

Mack, Y.P. (1981), Local Properties of k-NN Regression Estimates, SIAM Journal of Algebraic and Discrete Methods, 2, 311–323.

McLeish, D.L. (1974), Dependent Central Limit Theorems and Invariance Principles, Annals of Probability 2, 620–628.

Nadaraya, E.A. (1964), On Estimating Regression, Theory of Probability and its Applications 10, 186–190.

Neyman, J. (1959), Optimal Asymptotic Tests of Composite Statistical Hypotheses, Probability and Statistics, the Harald Cramer Volume, U. Grenander ed. (Wiley, New York).

Noether, G.E. (1955), On a Theorem of Pitman, Annals of Mathematical Statistics 26, 64–68.

Pemberton, J. and H. Tong (1981), A Note on the Distributions pf Non-Linear Autoregressive Stochastic Models, Journal of Time Series Analysis 2, 49–52.

Pham, T.D. and L.T. Tran (1985), Some Mixing Properties of Time Series Models, Stochastic Processes and their Applications 19, 297–303.

Pinkse, C.A.P. (1993a), On the Computation of Semiparametric Estimates in Limited Dependent Variable Models, Journal of Econometrics 58, 185–205.

Pinkse, C.A.P. (1993b), Entropy Based Testing Revisited, preprint.

Pinkse, C.A.P. (1993c), A General Characteristic Function Based Measure Applied to Serial Independence Testing, preprint.

Pinkse, C.A.P. and P.M. Robinson (1993), Pooling Nonparametric Estimates of Regression Functions with a Similar Shape, Statistical Methods of Econometrics and Quantitative Econometrics: a Volume in Honour of C.R. Rao, G.S. Maddala, P.C.B. Phillips and T.N. Srinivisan, eds., forthcoming.

Pitman, E.J.G. (1948), Nonparametric Statistical Inference, Lecture Notes (University of North Carolina Institute of Statistics).

Prakasa Rao, B.L.S. (1983), Nonparametric Functional Estimation, Academic Press.

Rao, C.R. (1948), Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation, Proceedings of the Cambridge Philosophical Society 44, 50–57.

Rao, C.R. (1977), Prediction of Future Observations with Special Reference to Linear Models, Multivariate Analysis IV, 193–208.

Robinson, P.M. (1984), Kernel Estimation and Interpolation for Time Series Containing Missing Observations, Annals of the Institute of Mathematical Statistics 36, 401–412.

Robinson, P.M. (1985), Testing for Serial Correlation in Regression with Missing Observations, Journal of the Royal Statistical Society B 47-3, 429–437.

Robinson, P.M. (1987), Time Series Residuals with Application to Probability Density Estimation, Journal of Time Series Analysis 8-3, 329–344.

Robinson, P.M. (1988), Root-N-Consistent Semiparametric Regression, Econometrica 56, 931–954.

Robinson, P.M. (1989), Hypothesis Testing in Semiparametric and Nonparametric Models for Econometric Time Series. Review of Economic Studies 56, 511–534.

Robinson, P.M. (1991a), Consistent Nonparametric Entropy Based Testing, Review of Economic Studies 58, 437–453.

Robinson, P.M. (1991b), Nonparametric Function Estimation for Long Memory Time Series, Proceedings of Fifth International Symposium in Economic Theory and Econometrics, W. Barnett, J. Powell, and G.E. Tauchen, eds., 437–457.

Robinson, P.M. (1991c), Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models, Econometrica 59, 1329–1364.

Robinson, P.M. (1992), The Approximate Distribution of Nonparametric Regression Estimates, preprint.

Rosenblatt, M. (1956), A Central Limit Theorem and a Strong Mixing Condition, Proceedings of the National Academy of Science U.S.A., 42, 43–47.

Rosenblatt, M. (1961), Independence and Dependence, Proceedings of the 4th Berkeley Symposium, 431–443.

Rosenblatt, M. (1975), A Quadratic Measure of Deviation of Two-Dimensional Density Estimates and a Test of Independence, Annals of Statistics 3, 1–14.

Rosenblatt, M. and B.E. Wahlen (1992), A Nonparametric Measure of Independence Under a Hypothesis of Independent Components, to appear in Statistics and Probability Letters 15.

Samuel-Cahn, E. (1994), Combining Unbiased Estimators, The American Statistician 48, 34–36.

Serfling, R. (1980), Approximation Theorems of Mathematical Statistics (Wiley, New York).

Silverman, B.W. (1982), Kernel Density Estimation Using the Fast Fourier Transform, Statistical Algorithm AS170, Applied Statistics 31, 93–97.

Silvey, S.D. (1959), The Lagrange Multiplier Test, Annals of Mathematical Statistics 30, 389–407.

Skaug, H.J. and D. Tjøstheim (1992a), Nonparametric Tests for Serial Independence, The M.B. Priestley Birthday Volume, forthcoming.

Skaug, H.J. and D. Tjøstheim (1992b), A Nonparametric Test for Serial Independence Based on the Empirical Distribution Function, preprint (University of Bergen, Bergen).

Stuetzle, W., T. Gasser, L. Molinari, R.H. Largo, A. Prader and P.J. Huber (1980), Self-Invariant Modeling of Human Growth, Annals of Human Biology 7, 507–528.

van der Genugten, B.B. (1988), Inleiding tot de Waarschijnlijkheidsrekening en Mathematische Statistiek, Deel 2 (Stenfert Kroese, Leiden, The Netherlands).

van Es, B. (1988), Aspects of Nonparametric Density Estimation, Ph.D. thesis (Faculteit der Wiskunde en Informatica, University of Amsterdam, Amsterdam).

Vasicek, O. (1976), A Test for Normality Based on Sample Entropy, Journal of the Royal Statistical Society Series B 31, 632–636.

Von Neuman, J. (1941), Distribution of the Ratio of the Mean Square Successive Difference to the Variance, Annals of Mathematical Statistics 12, 367–395.

Von Neuman, J., Kent, R.H., Belbinson, H.R. and B.I. Hart (1941), The Mean Square Successive Difference, Annals of Mathematical Statistics 12, 153–162.

Wahlen, B.E. (1991), A Nonparametric Measure of Independence, Ph.D. thesis (University of California at San Diego, La Jolla).

Wald, A. (1943), Tests of Statistical Hypotheses Concerning Several Parameters when the Number of Observations is Large, Transaction of the American Mathematical Society 54, 426–482.

Watson, G.S. (1964), Smooth Regression Analysis, Sankhyā Series A 26, 359–372.

Whistler, D.E.N. (1990), Semi-parametric Models of Daily and Intra-Daily Exchange Rate Volatility, Ph.D. thesis (University of London, London).

White, K. (1992), The Durbin-Watson Test for Autocorrelation in Nonlinear Models, Review of Economics and Statistics May 1992, 370–373.

Yoshihara, K. (1976), Limiting Behaviour of U-statistics for Stationary Absolutely Regular Processes, Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 35, 237–252.