DATA TILTING FOR TIME SERIES

Peter Hall[1] and Qiwei Yao[1,2]

Abstract. We develop a general methodology for tilting time series data. Attention is focused on a large class of regression problems, where errors are expressed through autoregressive processes. The class has a range of important applications, and in the context of our work may be used to illustrate the application of tilting methods to interval estimation in regression, robust statistical inference, and estimation subject to constraints. The interval estimation example includes empirical likelihood, where earlier applications to time series have involved either the multivariate "dual likelihood" approach introduced by Per Mykland, or a "Whittle likelihood" method suggested by Anna Monti. One advantage of our form of empirical likelihood is that it admits a wide range of distance, or more correctly divergence, functions. (We favour a non-traditional form of Kullback-Leibler divergence, because of its robustness properties.) Another is its simplicity; it is based directly on computed residuals, which are of course very familiar to time series analysts, and it does not involve the complexities of dual likelihood. A third advantage is its flexibility; it is readily applied to constructing confidence intervals or confidence bands in general regression problems. And a fourth is its context as a particular example of a very general methodology; our empirical likelihood approach is no more than a special case of a very broad class of tilting-based techniques for inference in time series problems.

Keywords. Autoregression, bootstrap, confidence interval, constrained inference, empirical likelihood, linear time series, power divergence, robust inference.

Short title. Data tilting for time series.

[1] Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia
[2] Department of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, UK

1 Introduction

Tilting methods in statistics are techniques for adjusting the empirical distribution by altering the data weights from their usual uniform values (identical to n^{-1} for each datum, where n denotes sample size) to a general multinomial (where weight p_i is given to the ith datum). Historically, this procedure goes back at least to methods of constrained nonparametric density estimation, introduced by Grenander (1956); and to related work on "nonparametric maximum likelihood estimation" by Kiefer and Wolfowitz (1956) and on survival analysis by Kaplan and Meier (1958). While the technique has a very wide range of applications (see e.g. Hall and Presnell, 1998, 1999ab), at its core lies the assumption that the data are independent and identically distributed. In this paper we discuss data tilting for time series, and describe their properties.

Our technique has connections to Mykland's (1995) "dual likelihood" approach, through its focus on the residuals in a time series. However, our use of general measures of discrepancy enables us to develop a wider range of applications, not least because we may take the weights p_i equal to 0. Another approach to generalising tilting methods to dependent data is that proposed by Kitamura (1997), based on blocking techniques. This method possesses the advantage that it applies directly to very general time series, but has the disadvantage that, like better known block bootstrap methods (e.g. Hall, 1985; Carlstein, 1986; Künsch, 1989) it requires very long time series and is particularly sensitive to choice of block size. By way of contrast, techniques based on linear processes require only relatively short time series and are relatively insensitive to choice of time series order. They also have effective application to nonlinear time series; see section 2.4 below.

Another attractive feature of tilting methods applied directly to time series, rather than via blocks, is that they enjoy several of the features of conventional tilting applied to independent and identically distributed data. Thus, for example, the versions of our technique that are appropriate for constructing confidence intervals have important connections to Owen's method of empirical likelihood (e.g. Owen, 2001, §8.2). In particular, one can view the method as "empirical likelihood with nuisance parameters," the latter being the autoregressive coefficients. If these coefficients are expressed through the equations that define them, they become a constraint under which an extremum of the empirical likelihood equations must be solved.

A tilting method alternative to ours would be to extend, from the special case of empirical likelihood, the spectral approach suggested by Monti (1997). This technique is in the spirit of "Whittle's method" (Whittle, 1953), in that it relies on asymptotic independence of periodogram ordinates. However, since it does not take into account the explicit nature of a linear model such as an autoregression, one would expect it to suffer from the inaccuracies of Whittle's method, relative to more conventional approaches, in time series problems where an explicit, finite-parameter linear model was valid. (Of course, the "parameters" here describe only the dependence structure; the independent disturbances, or errors, are not defined parametrically.) Moreover, the Whittle method is awkward to extend to the full range of applications of tilting, treated in the present paper.

The concept of empirical likelihood (Owen 1988, 1990) is noted, in the context of tilting and the bootstrap, by Efron (1981) and Davison, Hinkley and Worton (1992). Confidence bands, which will be discussed in §2, have featured before in the context of nonparametric or empirical likelihood primarily through the use of smoothing (e.g. Hall and Owen, 1993; Chen, 1996; Zhang, 1998), or applications to distribution functions or survival functions (e.g. Thomas and Grunkemeier, 1975; Owen, 1995; Zhang 1996; Hollander, McKeague and Yang, 1997).

2 Methods and Calculation

2.1 General Methodology

Assume data $\mathcal{D} = \{(X_i, Y_i), 1 \le i \le n\}$ are generated by the model

$$Y_i = g(X_i|\theta) + e_i\,, \tag{2.1}$$

where $g(\cdot|\theta)$ is a smooth function determined by the r-vector θ of parameters, and the errors e_i form a stationary autoregression of order s, with

$$\epsilon_i = e_i + \sum_{j=1}^{s} \alpha_j\, e_{i-j}\,,$$

$\alpha_1 > 0$ and $\alpha_s \ne 0$ and ϵ_j, $-\infty < j < \infty$, being independent and identically distributed random variables with finite variance σ^2 and zero mean. One approach to estimating $\theta = (\theta_1, \ldots, \theta_r)^T$ and $\alpha = (\alpha_1, \ldots, \alpha_s)^T$ is to use weighted least-squares, choosing $\big(\widehat{\theta}(p), \widehat{\alpha}(p)\big)$ to minimise

$$S(\theta, \alpha|p) = \sum_{j=s+1}^{n} p_j \left[Y_j - g(X_j|\theta) + \sum_{i=1}^{s} \alpha_i \left\{ Y_{j-i} - g(X_{j-i}|\theta) \right\} \right]^2, \tag{2.2}$$

where $p = (p_{s+1}, \ldots, p_n)$ denotes a multinomial distribution on the indices $s+1, \ldots, n$, with each $p_i \ge 0$ and $\sum_{s+1 \le i \le n} p_i = 1$. Based on the estimated (θ, α), we define

$$\widehat{\epsilon}_j(p) = Y_j - g\{X_j|\widehat{\theta}(p)\} + \sum_{i=1}^{s} \widehat{\alpha}_i(p) \left[Y_{j-i} - g\{X_{j-i}|\widehat{\theta}(p)\} \right], \tag{2.3}$$

being an empirical approximation to ϵ_j; and we put

$$\widehat{\sigma}(p)^2 = \sum_{j=s+1}^{n} p_j\, \widehat{\epsilon}_j(p)^2 - \left\{ \sum_{j=s+1}^{n} p_j\, \widehat{\epsilon}_j(p) \right\}^2 , \tag{2.4}$$

being an estimator of $\sigma^2 = \mathrm{var}(\epsilon_j)$.

The standard least-squares estimator $(\widetilde{\theta}, \widetilde{\alpha})$ of (θ^0, α^0), denoting the true value of (θ, α), would be obtained with $p = p_{\text{unif}} = (1/(n-s), \ldots, 1/(n-s))$. Biased-bootstrap or tilting methods would involve choosing $p = \widehat{p}$ to minimise a measure of distance or divergence, $D(p)$, between p and p_{unif}, subject to a constraint on a subvector of $\big(\widehat{\theta}(p), \widehat{\alpha}(p), \widehat{\sigma}(p)^2\big)$. Typically the subvector would be $\widehat{\theta}(p)$ or $\widehat{\sigma}(p)^2$, α being a vector of nuisance parameters. We shall consider only power divergence measures of divergence,

$$D_\rho(p) = \begin{cases} \{\rho(1-\rho)\}^{-1} \left\{1 - m^{-1} \sum_{i=s+1}^{n} (mp_i)^\rho\right\} & \text{if } \rho \neq 0,\, 1 \\ -m^{-1} \sum_{i=s+1}^{n} \log(mp_i) & \text{if } \rho = 0 \\ \sum_{i=s+1}^{n} p_i \log(mp_i) & \text{if } \rho = 1, \end{cases} \tag{2.5}$$

where $m = n - s$. See Rényi (1961), Cressie and Read (1984) and Read and Cressie (1988). We treat a class of constrained inferential problems that may be expressed in the form:

Problem: find $\bar{\theta}(\psi_\star) \equiv \widehat{\theta}(\widehat{p})$ and $\bar{\alpha}(\psi_\star) \equiv \widehat{\alpha}(\widehat{p})$, where $\widehat{p} \equiv \widehat{p}(\psi_\star)$ is chosen to minimise $D_\rho(p)$ subject to $\widehat{\psi}(p) \equiv \psi\{\widehat{\theta}(p), \widehat{\alpha}(p), \widehat{\sigma}(p)^2\} = \psi_\star$, ψ is a known, smooth, t-variate function of $r + s + 1$ variables, and $\psi_\star$ is a given t-vector. (2.6)

The method can be expected to be relatively insensitive to choice of the autoregressive order, s, particularly in comparison to the high degree of sensitivity shown by block-based methods to selection of block length. In particular, overestimation of s will usually cause only minor problems. Moreover, since the most ambitious aim of our methodology is to approximate the distribution of a statistic computed from a time series, rather than to approximate the joint distributions of the time series itself, then we need only capture main features of the way in which the dependence structure affects the statistic. This means that the AR order used in the approximation can often be relatively low.

Example 1: *Generalised empirical likelihood.* One instance of this example is that where $\psi(\theta, \alpha, \sigma^2) \equiv \psi(\theta)$ denotes a subvector, of length $t \leq r$, of θ. More generally, an α-level generalised empirical likelihood confidence region for the true value $\psi(\theta^0)$ of $\psi(\theta)$ is the set $\{\psi_\star : L(\psi_\star) \leq u_\alpha\}$, where

$$(2m)^{-1} L(\psi_\star) = \min_{p\,:\,\widehat{\psi}(p) = \psi_\star,\ \sum_{i=s+1}^{n} p_i = 1} D_\rho(p)$$

and u_α is either the α-level critical point of the $\mathcal{X}^2$-distribution with t degrees of freedom, or is obtained by bootstrap calibration. See Section 2.3 for a discussion of calibration, and Section 4.1 for an account of asymptotic properties of the generalised likelihood, L. To obtain a simultaneous empirical likelihood confidence band for the function $g(\cdot|\theta)$, first define $\psi(\theta) \equiv \theta$. Then the region is the envelope of the class of functions $\{g(\cdot|\theta_\star) : L(\theta_\star) \le u_\alpha\}$, where

$$(2m)^{-1} L(\theta_\star) = \min_{p\,:\,\widehat{\theta}(p)=\theta_\star,\, \sum_{i=s+1}^{n} p_i = 1} D_\rho(p)\,.$$

A confidence interval for the value of $g(x|\theta)$, for a particular value of x, is obtained by taking $\psi(\theta) \equiv g(x|\theta)$, a scalar, in Problem (2.6) above. There is no difficulty in including the case that $\psi = \psi(\theta, \alpha)$, and in fact our theoretical development in section 4 will accommodate this possibility.

Example 2: *Robust inference.* To make a general statistical procedure for the data $\mathcal{D}$ more robust against unduly large values of the errors e_j or the disturbances ϵ_j, arising for example through contamination, we suggest choosing p so as to reduce the size of $\widehat{\sigma}(p)^2$ given in (2.4). We work with $\widehat{\sigma}(p)^2$, rather than with the analogous estimator of the variance of e_j, since the e_j's are not independent.

Thus, we take $\psi(\theta, \alpha, \sigma^2) \equiv \sigma^2$, a scalar, in Problem (2.6). Let $\widetilde{\sigma}^2 = \widehat{\sigma}(p_{\rm unif})^2$ denote the standard least-squares estimate of σ^2. Then, defining $\widehat{p}_\star = \widehat{p}(\psi_\star)$ to be the value of p that minimises $D_\rho(p)$ subject to $\widehat{\sigma}(p)^2 = \psi_\star$, and considering successively smaller values of $\psi_\star < \widetilde{\sigma}^2$, we are in effect "censoring" or "Winsorising" to a successively greater extent the residuals $\widehat{\epsilon}_j(p) \equiv \widehat{\epsilon}_j\{\widehat{\theta}(p), \widehat{\alpha}(p)\}$ that have most leverage on the value of $\widehat{\sigma}(p)^2$. The corresponding robust estimators of θ and α are $\widehat{\theta}_\star = \widehat{\theta}(\widehat{p}_\star)$ and $\widehat{\alpha}_\star = \widehat{\alpha}(\widehat{p}_\star)$.

The least-squares procedure defined by minimising $S(p)$, at (2.2), is equivalent to (conditional) maximum likelihood when the disturbances ϵ_j are Normally distributed. We might choose $\psi_\star$ so that a Q-Q plot for the residuals $\widehat{\epsilon}_k(\widehat{p}_\star)$, $s + 1 \le k \le n$, is approximately linear under the assumption of Normality. In addition to producing robust estimators $\widehat{\theta}(\widehat{p}_\star)$ and $\widehat{\alpha}(\widehat{p}_\star)$, this technique enables development of robust bootstrap methods for constructing confidence regions for θ^0 and α^0. We resample using the model

$$Y_j^\dagger = g\{X_j | \widehat{\theta}(\widehat{p}_\star)\} + e_j^\dagger\,,$$

where the $e_j^\dagger$'s are defined by the autoregression

$$\widehat{\epsilon}_j^\dagger = e_j^\dagger + \sum_{i=1}^{s} \widehat{\alpha}_i(\widehat{p}_\star) e_{j-i}^\dagger\,,$$

and conditional on $\mathcal{D}$ the variables $\widehat{\epsilon}_j^\dagger$ are independent and identically distributed with

$$P\{\widehat{\epsilon}_j^\dagger = \widehat{\epsilon}_k(\widehat{p}_\star) - \widehat{\mu}(\widehat{p}_\star) \big| \mathcal{D}\} = \widehat{p}_k^\star \,, \qquad s+1 \le k \le n \,,$$

for $\widehat{\mu}(p) = \sum_{k=s+1}^n p_k \widehat{\epsilon}_k(p)$. Both percentile and percentile-t confidence regions are possible in this setting.

Example 3: *Estimation subject to constraints.* Physical considerations may dictate that the true mean function, $g^0 = g^0(\cdot|\theta^0)$, satisfy a constraint such as monotonicity, in a specified direction, or convexity. In linear regression problems the constraint might be that slope be greater than or equal to a given value. If the least-squares estimator $g(\cdot|\widetilde{\theta})$ violates the constraint then we might replace $\widetilde{\theta}$ by that value of θ that minimises a least-squares criterion, say, subject to the constraint. However, the reason for $g(\cdot|\widetilde{\theta})$ failing to satisfy the constraint may be that $\widetilde{\theta}$ is substantially in error, for example due to experimental error, and then it makes little sense to choose θ close to $\widetilde{\theta}$. An alternative is to choose $p = \widehat{p}_c$, c standing for "constraint", such that $D_\rho(p)$ is minimised subject to $g\{\cdot|\widehat{\theta}(p)\}$ satisfying the constraint; and take $g\{\cdot|\widehat{\theta}(\widehat{p}_c)\}$ to be our estimator of g^0.

If we suppose that contamination caused the constraint to fail for $g(\cdot|\widetilde{\theta})$ then the approach discussed above is related to that suggested in Example 2, and $\widehat{\theta}(\widehat{p}_c)$ is in a sense a "robustified" estimator of θ. However, robustness is now enforced through the constraint, not through empirical variance of the estimated disturbances.

A qualitative constraint on $g(\cdot|\theta)$ is in general different from the "equality" constraint imposed for Problem (2.6). Instead of the identity $\psi\{\widehat{\theta}(p), \widehat{\alpha}(p), \widehat{\sigma}(p)^2\} = \psi_\star$ in (2.6) we ask that $\widehat{\theta}(p) \in \Theta_\star$, where $\Theta_\star$ denotes the set of θ such that $g(\cdot|\theta)$ satisfies the constraint. Often, "$\widehat{\theta}(p) \in \Theta_\star$" reduces to "$\widehat{\theta}_j(p) \in \mathcal{I}_\star$", where $\widehat{\theta}_j$ denotes a specific component of $\widehat{\theta}$ and $\mathcal{I}_\star$ is a semi-infinite interval. For example, when $g(\cdot|\theta)$ is linear and the constraint is monotonicity, $\widehat{\theta}_j$ is the estimate of slope and $\mathcal{I}_\star$ equals either $[0, \infty)$ or $(-\infty, 0]$, depending on the direction of the constraint.

A fourth example, related to the third and to the theory with which we shall deal briefly at the end of §4, is that of tilting the empirical distribution as a prelude to conducting a hypothesis test. It may be applied to a particularly wide range of problems; see Hall and Presnell (1999a) for the case of independent data. Taking a particular example for the purpose of illustration in a time series setting, let us suppose we wish to test the hypothesis that the error distribution has zero skewness, against the complementary alternative. Redefine ψ to be the skewness of the distribution of ϵ_j and define $\widehat{\psi}(p)$ to equal the empirical skewness of the tilted residuals:

$$\widehat{\psi}(p) = \sum_{j=s+1}^n p_j \, \widehat{\epsilon}_j(p)^3 - 3\left\{ \sum_{j=s+1}^n p_j \, \widehat{\epsilon}_j(p) \right\} \sum_{j=s+1}^n p_j \, \widehat{\epsilon}_j(p)^2 + 2\left\{ \sum_{j=s+1}^n p_j \, \widehat{\epsilon}_j(p) \right\}^3 .$$

In this instance we compute $p = \widehat{p}_\star$ to minimise $D_\rho(p)$ subject to $\widehat{\psi}(p) = 0$. The bootstrap test is then calibrated by sampling nonuniformly, with respective weights $(\widehat{p}_\star)_j$, from $\widehat{\epsilon}_j(\widehat{p}_\star)$ for $s+1 \le j \le n$, and determining the critical point which the absolute value of the bootstrap form of empirical skewness exceeds with probability π, conditional on the data (to obtain a test with nominal level π). The null hypothesis is rejected if the actual value of empirical skewness, computed from the conventional residuals $\widehat{\epsilon}_j(\widehat{p}_{\text{unif}})$, exceeds the critical point. Level accuracy of the test can be enhanced using the double bootstrap; see for example Hall (1992, Chapter 3).

2.2 Solving Problem (2.6)

Let $f_1(p_k)$ denote any function proportional to $\partial D_\rho(p)/\partial p_k$, and put $f \equiv f_1^{-1}$. Then, a Lagrange multiplier argument shows that

$$p_k = m^{-1} f\left(\lambda_1 + \lambda_2^T \frac{\partial \psi}{\partial \theta}\frac{\partial \theta}{\partial p_k} + \lambda_2^T \frac{\partial \psi}{\partial \alpha}\frac{\partial \alpha}{\partial p_k} + \lambda_2^T \frac{\partial \psi}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial p_k}\right), \qquad s+1 \le k \le n, \tag{2.7}$$

where λ_1 and $\partial\sigma^2/\partial p_k$ are scalars, λ_2 and $\partial\psi/\partial\sigma^2$ are t-vectors, $\partial\psi/\partial\theta$ and $\partial\psi/\partial\alpha$ are t-by-r and t-by-s matrices, respectively, and $\partial\theta/\partial p_k$ and $\partial\alpha/\partial p_k$ are r- and s-vectors, respectively. For given k, the $(r+s)$-vector $w \equiv ((\partial\theta/\partial p_k)^T, (\partial\alpha/\partial p_k)^T)^T$ may be represented as a function of θ, α and p, through the $r+s$ equations

$$A_k(\theta,\alpha,p)\,\frac{\partial \theta}{\partial p_k} + B_k(\theta,\alpha,p)\,\frac{\partial \alpha}{\partial p_k} = V_k\,, \tag{2.8}$$

where A_k and B_k are $(r+s)$-by-r and $(r+s)$-by-s matrices and V_k is an $(r+s)$-vector. The components of A_k, B_k and V_k are given in Appendix 1. By (2.4),

$$\frac{\partial \sigma^2}{\partial p_k} = \widehat{\epsilon}_k^2 - 2\widehat{\epsilon}_k \sum_{j=s+1}^{n} p_j\,\widehat{\epsilon}_j + 2\sum_{j=s+1}^{n} p_j\,\widehat{\epsilon}_j\,\frac{\partial \widehat{\epsilon}_j}{\partial p_k} - 2\sum_{j=s+1}^{n} p_j\,\widehat{\epsilon}_j \sum_{j=s+1}^{n} p_j\,\frac{\partial \widehat{\epsilon}_j}{\partial p_k}\,, \tag{2.9}$$

where $\partial\widehat{\epsilon}_j/\partial p_k$ can be calculated in terms of $\partial\theta/\partial p_k$ and $\partial\alpha/\partial p_k$ via (2.3). Along with the components of θ, the quantities α, p, λ_1 and λ_2 are given by the $n-s$ equations at (2.7) and the $r+s+t+1$ equations

$$\sum_{j=s+1}^{n} Z_j(\theta,\alpha)\,p_j\left\{\dot{g}_u(X_j|\theta) + \sum_{i=1}^{s}\alpha_i\,\dot{g}_u(X_{j-i}|\theta)\right\} = 0\,, \qquad 1 \le u \le r\,, \tag{2.10}$$

$$\sum_{j=s+1}^{n} Z_j(\theta,\alpha)\,p_j\,\{Y_{j-v} - g(X_{j-v}|\theta)\} = 0\,, \qquad 1 \le v \le s\,, \tag{2.11}$$

$$\psi(\theta,\alpha) = \psi_\star\,, \tag{2.12}$$

and

$$\sum_{j=s+1}^{n} p_j = 1\,, \tag{2.13}$$

where

$$Z_j(\theta,\alpha) = Y_j - g(X_j|\theta) + \sum_{i=1}^{s} \alpha_i \,\{Y_{j-i} - g(X_{j-i}|\theta)\} \tag{2.14}$$

and $\dot{g}_u(x|\theta)$ denotes the u-th element of the r-vector $\partial g(x|\theta)/\partial\theta$.

The value of $f(u)$ in (2.7) may be taken equal to $u^{-1/(1-\rho)}$ if $\rho \neq 0$ or 1, u^{-1} if $\rho = 0$ and e^u if $\rho = 1$. In this notation, we have $\lambda_1 = 1$ at (2.7) when $\rho = 0$, and λ_1 reduces to a scaling constant when $\rho = 1$.

2.3 Calibrating empirical likelihood

Bootstrap methods, instead of the χ^2 approximation, may be used to calibrate empirical likelihood confidence regions, as follows. Let $(\widetilde{\theta}, \widetilde{\alpha}, \widetilde{\sigma}^2)$ be the standard least-squares estimators, put $\widetilde{\psi} = \psi(\widetilde{\theta}, \widetilde{\alpha}, \widetilde{\sigma}^2)$, and let

$$\widetilde{\epsilon}'_j = Y_j - g(X_j|\widetilde{\theta}) + \sum_{i=1}^{s} \widetilde{\alpha}_i \,\{Y_{j-i} - g(X_{j-i}|\widetilde{\theta})\}\,.$$

For $s+1 \leq j \leq n$, let $\widetilde{\epsilon}_1, \cdots, \widetilde{\epsilon}_n$ be the centred values of $\widetilde{\epsilon}'_j$. We generate data (X_i, Y_i^*), $1 \leq i \leq n$, from the model $Y_i^* = g(X_i|\widetilde{\theta}) + e_i^*$, where the e_i^*'s form a bootstrap autoregression and are resampled from $\widetilde{\epsilon}_{s+1}, \cdots, \widetilde{\epsilon}_n$. (The design variables X_i are not resampled.) From those data, compute the bootstrap version $L^*(\psi_\star)$ of $L(\psi_\star)$. Given a nominal coverage probability γ for an empirical likelihood confidence region, define $u = \widehat{u}_\gamma$ to be the nearest solution of $P\{L^*(\widetilde{\psi}) \leq u|\mathcal{D}\} = \gamma$. Then the bootstrap-calibrated empirical likelihood confidence region for θ is $\{\psi_\star :\ L(\psi_\star) \leq \widehat{u}_\gamma\}$.

The bootstrap method described above involves drawing samples from the estimated autoregression model, which should be constrained to be causal in the sense that the equation

$$1 + \sum_{j=1}^{s} \widetilde{\alpha}_j \, z^j = 0 \tag{2.15}$$

has its all roots outside of the unit circle $|z| = 1$. In practice this can be achieved as follows. Suppose $z_1, \cdots, z_s$ are the roots of (2.15) with $|z_j| < 1$ for $1 \leq j \leq k$, and $|z_j| > 1$ for $k < j \leq s$. We generate e_j^* from model

$$e_i^* + \sum_{j=1}^{s} \breve{\alpha}_j \, e_{i-j}^* = \epsilon_i^* \prod_{l=1}^{k} |z_l|\,, \tag{2.16}$$

where the $\breve{\alpha}_j$'s are determined by

$$1+\sum_{j=1}^{s} \breve{\alpha}_j\, z^j = \prod_{j=1}^{k} (1 - z z_j) \prod_{i=k+1}^{s} (1 - z/z_i)\,.$$

Note that model (2.16) admits the same autocovariance function as the estimated autoregression model; see Proposition 4.4.2 of Brockwell and Davis (1991).

2.4 An extension

The general methodology outlined in section 2.1 may be readily extended to deal with model (2.1) in which e_i follows a parametric nonlinear autoregressive model. We illustrate this idea when e_i is an autoregressive conditional heteroscedastic process, i.e.

$$e_i = \epsilon_i \left\{ \alpha_0 + \sum_{j=1}^{s} \alpha_j e_{i-j}^2 \right\}^{1/2},$$

where ϵ_i are independent and identically distributed random variables with mean 0 and variance 1, all α_j are non-negative, and $\alpha_0 > 0$ and $\sum_{1\le j\le s} \alpha_j < 1$. Under those conditions, the above equation defines a unique strictly stationary process $\{e_i\}$ with $Ee_i^2 < \infty$.

The statements in section 2.1 are still valid if we replace the weighted least-squares (2.2) by

$$S(\theta, \alpha|p) = \sum_{j=s+1}^{n} p_i \, \frac{\{Y_j - g(X_j|\theta)\}^2}{\alpha_0 + \sum_{i=1}^{s} \alpha_i \{Y_{j-i} - g(X_{j-i}|\theta)\}^2}\,,$$

and the residual (2.3) by

$$\widehat{\epsilon}_j(p) = \frac{Y_j - g\{X_j|\widehat{\theta}(p)\}}{[\widehat{\alpha}_0(p) + \sum_{i=1}^{s} \widehat{\alpha}_i(p)\{Y_{j-i} - g(X_{j-i}|\widehat{\theta}(p))\}^2]^{1/2}}.$$

3 Numerical Study

In this section we use simulation studies to illustrate each of the three problems treated in §2.1. Throughout we employ the power divergence D_ρ with $\rho = 1$. This choice was made because of the good robustness properties of D_1. In particular, $D_1(p)$ remains finite even if some values of p_i are zero. Otherwise, the algorithm becomes (in effect) obsessed with reweighting a small number of data that have considerable leverage on the value of D_ρ, due to little more than artifacts in the definition of distance.

Example 1: *Empirical likelihood.* For the model

$$Y_j = \theta\, X_j\, (2 - X_j) + e_j\,, \quad \text{and} \quad e_j = \epsilon_j - \alpha_1\, e_{j-1} - \alpha_2\, e_{j-2}\,,$$

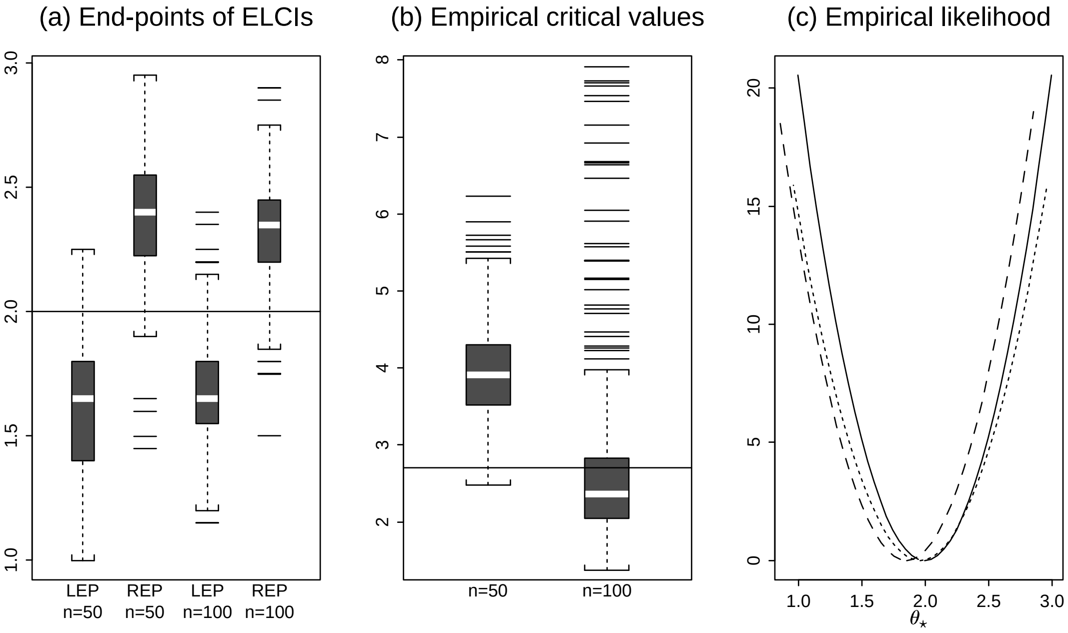


Figure 1: *(a) Boxplots of left end points (LEP) and right end points (REP) of the simulated generalised empirical likelihood confidence interval for θ at level $\gamma = 0.9$. The horizontal line indicates the true value $\theta = 2$. (b) Boxplots of bootstrap critical values $\widehat{u}_\gamma$. The horizontal line indicates the critical value derived from a χ^2 approximation. (c) Three examples of generalised empirical likelihood $D_1\{\widehat{p}(\theta_\star)\}$, plotted against $\theta_\star$, with the critical value equal to, respectively, 25% (dotted curve), 50% (solid curve) and 75% (dashed curve) quantiles in the 400 replications.*

we constructed the generalised empirical likelihood confidence interval for θ. We took the X_j's to be independent $U(0,2)$ random variables, and the ϵ_j's to be independent and standard normal. Setting $\theta = 2$, $\alpha_1 = 1$ and $\alpha_2 = 0.25$, we drew 400 samples for $n = 50$ and $n = 100$. We set the nominal confidence level at $\gamma = 0.9$. The bootstrap method, instead of a χ^2 approximation, was used to determine the critical value $\widehat{u}_\gamma$. For each sample, we repeated bootstrap sampling 400 times. In this example, the chi-squared approximation — noted in Theorem 4.1 below — gave inferior results. There are alternative methods of calibration, but the main ones are based on asymptotic approximations that capture high-order terms, and are quite tedious to apply in the time-series setting.

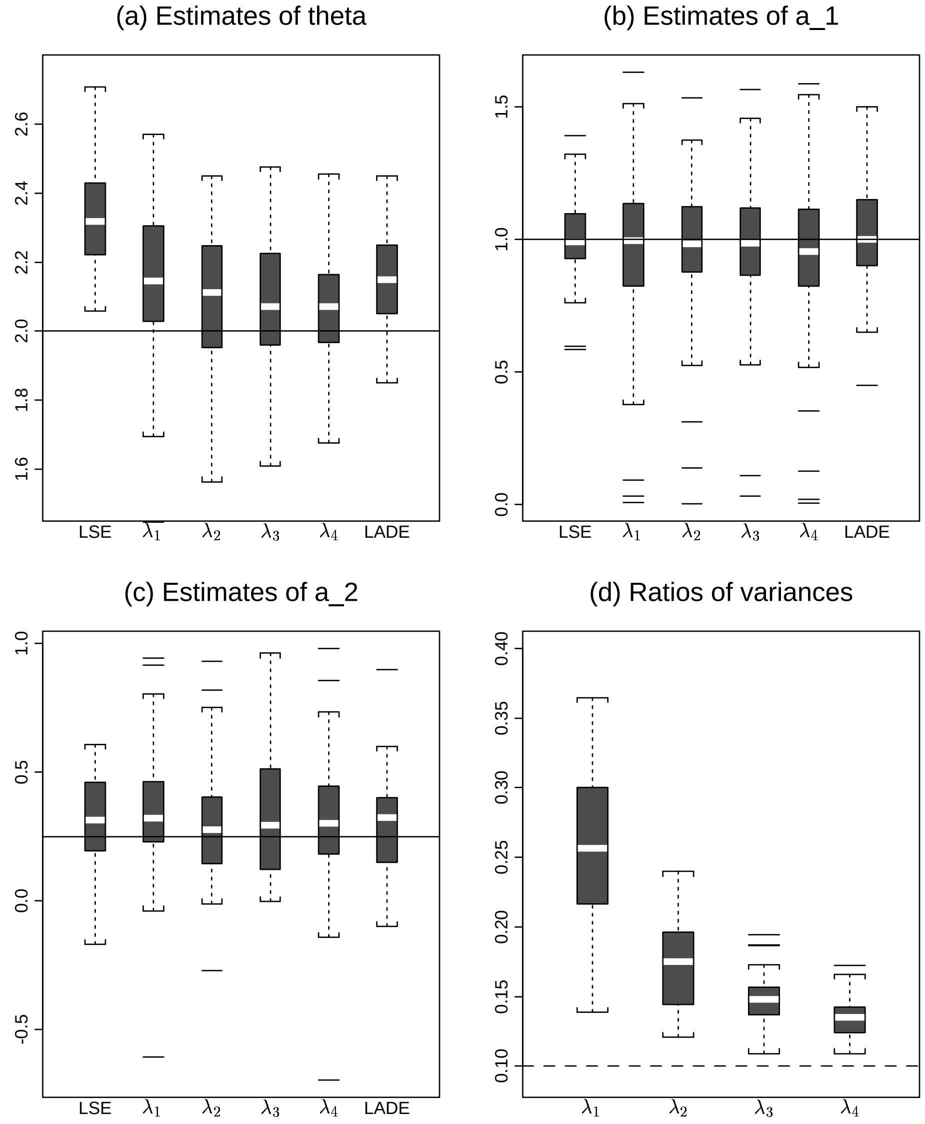
The generalised empirical likelihood function was calculated in terms of an iterative algorithm, as follows. With given initial values of θ, α_i's and p_i's, find λ_2 based on (2.12) and (2.7)–(2.9). (Note that λ_1 can be obtained through a simple scaling.) Then update θ, the α_i's in terms of (2.10) and (2.11), and the p_i's in terms of (2.7). We repeated the above process until the estimated value of θ differed from $\theta_\star$ by only 10^{-4}. We started $\theta_\star$ at the value of the standard lease squares estimate. Using the estimates in the previous setting as initial values, we gradually increased (or decreased) the value of $\theta_\star$. The two end-points of the confidence intervals are displayed in Fig. 1(a). The relative frequency of the interval covering the true value $\theta = 2$ (over 400 replications) was 0.920 for $n = 50$, and 0.905 for $n = 100$. The average length of the interval was 0.789 for $n = 50$, and 0.645 for $n = 100$. Fig. 1(b) presents boxplots of the bootstrap-determined critical values $\widehat{u}_\gamma$, which are closer to 2.706 (the 90% quantile of the χ^2-distribution with 1 degree of freedom) for $n = 100$ than for $n = 50$. Fig. 1(c) plots three typical examples of the generalised empirical likelihood function $D_1\{\widehat{p}(\theta_\star)\}$ against $\theta_\star$, with sample size $n = 100$. Those examples were selected such that the corresponding bootstrap critical values were at the 25%, 50% and 75% quantiles respectively during simulations with 400 replications.

In conclusion it can be seen that coverage accuracy is very good, variability of interval endpoints is low, and the generalised empirical likelihood function is convex.

Example 2: *Robust estimation of location.* Consider the model

$$Y_j = \theta + e_j \quad \text{where} \quad e_j = \epsilon_j - \alpha_1\, e_{j-1} - \alpha_2\, e_{j-2}\,,$$

where the ϵ_j's, θ and α_i's are as in Example 1. With contaminated observations we seek estimators $\widehat{\theta}(\widehat{p})$, $\widehat{\alpha}_1(\widehat{p})$ and $\widehat{\alpha}_2(\widehat{p})$, where $\widehat{p}$ minimises $D_1(p)$ subject to $\widehat{\sigma}(p)^2$, defined as in (2.4), equal to a fixed value $\sigma_\star^2$. Having encountered numerical problems with an iterative algorithm based on the equations in §2.2, caused by the complex function form of ψ, we adopted a simulated annealing



(a) Estimates of theta

(b) Estimates of a_1

(c) Estimates of a_2

(d) Ratios of variances

Figure 2: *Contamination of Y_j's. Boxplots show variability of the standard least squares estimates (LSE), the estimates derived from minimising (3.1) with $\lambda = \lambda_i$ $(i = 1, \cdots, 4)$, and the least absolute deviations estimates (LADE) of (a) θ, (b) α_1 and (c) α_2. Horizontal lines indicate true values of parameters. (d) Boxplots of ratios $\widehat{\sigma}(\widehat{p})^2/\widehat{\sigma}(p_{\text{unif}})^2$. The dashed line indicates the target value 0.1 for the ratios.*

approach. To this end we searched for $\widehat{p}$ that minimised

$$L(p) \equiv D_1(p) + \lambda \, |\widehat{\sigma}(p)^2 - \sigma_\star^2| \, , \tag{3.1}$$

where $\lambda > 0$ controlled the penalty for the discrepancy between $\widehat{\sigma}(p)^2$ and its target value $\sigma_\star^2$. Letting $n = 30$, we considered two types of contamination: either each Y_j or each ϵ_j received an added value 3 with probability 0.25.

Setting $\sigma_\star^2 = 0.1 \, \widehat{\sigma}(p_{\rm unif})^2$, we started the search with the initial p being the uniform distribution over those j's such that $|\widetilde{\epsilon}_j| \le \widehat{\sigma}(p_{\rm unif})$ (therefore $p_i = 0$ if $|\widetilde{\epsilon}_i| > \widehat{\sigma}(p_{\rm unif})$), where the $\widetilde{\epsilon}_j$'s were the residuals based on the standard least squares estimates. We added an independent $N(0, \delta^2)$ perturbation to each p_j, setting $p_j = 0$ whenever it took a negative value. The new p was obtained by the standard normalisation. We kept the new p if $L(p)$ was not greater than the minimum value of L (plus a small allowance τ) at that stage. For each $\lambda = \lambda_i \equiv 40Ni$ and $i = 1, \cdots, 4$, we repeated the above procedure 50 million times with δ fixed at $m^{-1}/2^{j-1}$ for $j = 1, \cdots, 4$ in turn. We proceeded to the next stage if the minimum value of L failed to reduce in 50,000 successive searches. We took τ to equal 30% of the minimum value in the previous stage, starting at $0.3 \, \lambda \, |\widehat{\sigma}(p_{\rm unif})^2 - \sigma_\star^2|$. The overall minimiser of L was taken as the value of $\widehat{p}$. Each replication took about 5.4 hours using a Linux PC equipped with a Pentium III 1GHz processor.

Fig. 2 gives boxplots of estimates computed from 50 samples with contamination of the Y_j's. For the sake of comparison we also calculated the standard least squares estimates (LSE) and least absolute deviations estimates (LADE). The former method is of course known to be sensitive to outliers, while the latter is robust against outliers. Fig. 2(a) shows that our data tilting approach substantially reduces bias of estimators of θ, relative to LADE, and reduces variability too for larger λ. Fig. 2(d) indicates that for larger λ the ratio $\widehat{\sigma}(\widehat{p})^2/\widehat{\sigma}(p_{\rm unif})^2$ was closer to its target value, 10%. Estimates of the autoregressive parameters α_1 and α_2 show less of a pattern, but of course those quantities are not the target of our method.

Similar results are also evident from Fig. 3, where the estimates were obtained from 50 samples with contamination on the ϵ_j's.

The value of $\sigma_\star^2$ is often determined by the nature of problem and prior experience. Typically we may let $\sigma_\star^2 = a \, \widehat{\sigma}(\widehat{p}_{\rm unif})^2$ for $a \in [0.1, 0.25]$. In practical implementation we may take, in (3.1), $\sigma_\star^2$ to be smaller than a target value, and then choose the value λ such that the ratio $\widehat{\sigma}(\widehat{p})^2/\widehat{\sigma}(\widehat{p}_{\rm unif})^2$ equals a; see Figs.2(d) and 3(d).

Example 3: *Estimation in regression subject to sufficiently steep slope.* Consider the linear

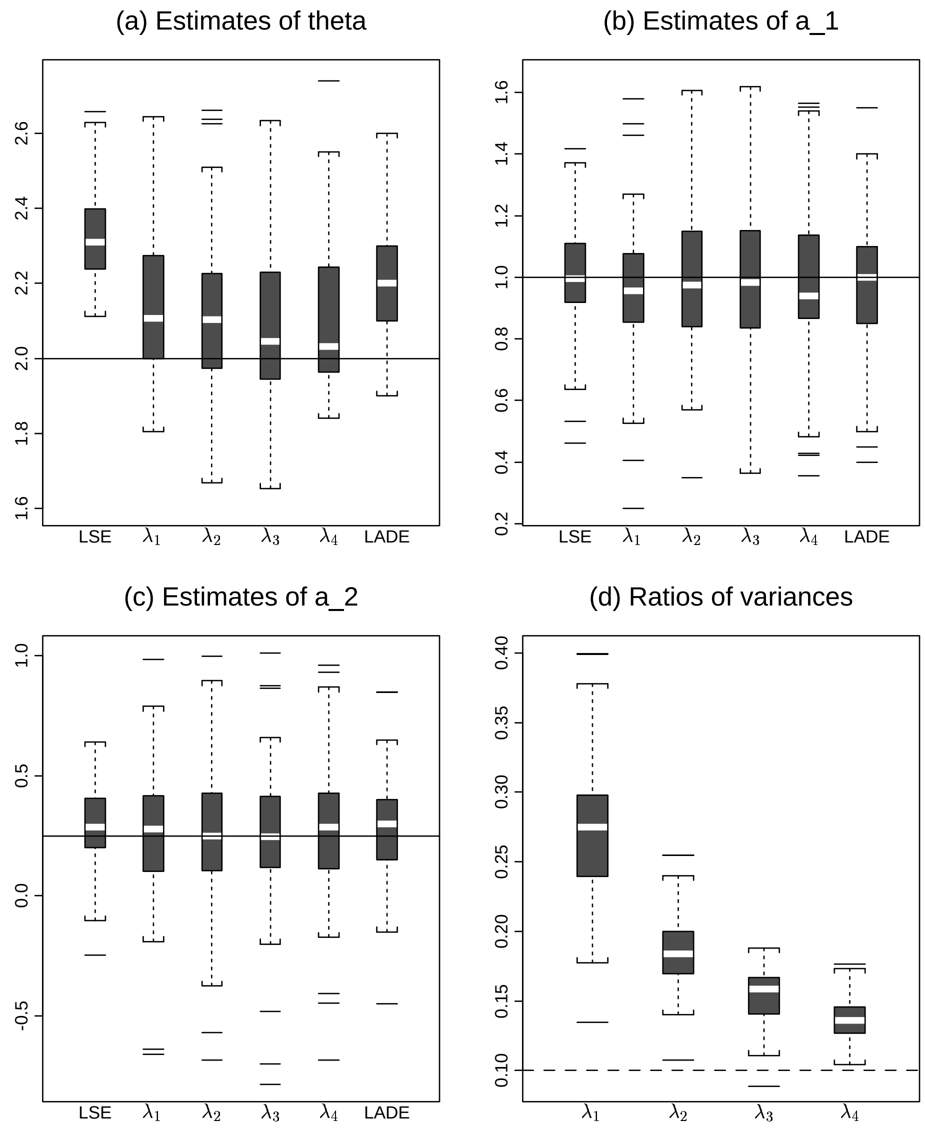


Figure 3: *Contamination on ϵ_j's: Panel information is as for Fig. 2.*

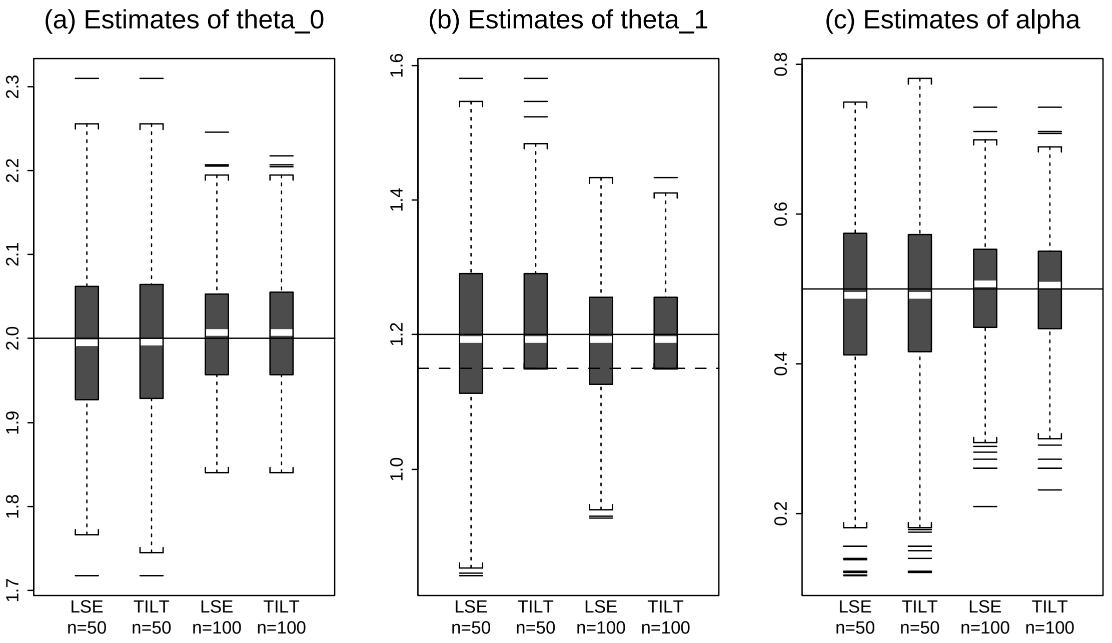


Figure 4: *Boxplots show variability of the standard least squares estimates (LSE) and estimates derived by constrained tilting (TILT) when the target was (a) θ_0, (b) θ_1 and (c) α. Unbroken horizontal lines indicate true parameter values, and the dashed line in panel (b) indicates the constraint imposed on the tilted estimates of slope.*

regression model

$$Y_j = \theta_0 + \theta_1 X_j + e_j\,, \quad \text{and} \quad e_j = \epsilon_j - \alpha\, e_{j-1}\,,$$

where both the X_j's and the ϵ_j's are independent and standard normal. Setting $\theta_0 = 2$, $\theta_1 = 1.2$ and $\alpha = 0.5$, we compared the standard least squares estimators of θ_0, θ_1 and α with the estimators obtained by minimising generalised empirical likelihood subject to the constraint that the regression line is sufficiently strongly increasing — specifically, $\theta_1 \geq 1.15$. In practice a constraint such as this would often be determined through prior experience with similar problems in the past. In a sense, the methodology provides something like a Bayesian option to nonparametric inference; it allows the experimenter to impose his or her prior belief on the value of estimator.

As in Example 1 we calculated the estimates in terms of an iterative algorithm based on the equations in §2.2. The boxplots of those estimates are depicted in Fig. 4. Note that the estimates derived from the generalised empirical likelihood are only different from the least squares estimates when $\widetilde{\theta}_1 < 1.15$.

Of course, imposing the constraint removes all instances where $\widehat{\theta}_1$ takes values less than 1.15. However, it has little impact on distribution of the estimator on the upper side of 1.15, and neither does it have a noticeable effect on either bias or variability of $\widehat{\theta}_2$ or $\widehat{\theta}_3$.

4 Theoretical Properties

In the context of Example 1 in §2, we show that if $\psi_\star$ equals the true value of $\psi(\theta, \alpha)$ then, under regularity conditions, the generalised empirical likelihood ratio statistic $(2m)^{-1}L(\psi_\star)$ is asymptotically distributed as χ^2 with t degrees of freedom. At the end of this section we discuss theory for the other two examples from §2.

Let θ^0, α^0 denote the true values of θ, α respectively, and put $\psi^0 = \psi(\theta^0, \alpha^0)$. Suppose that for each n,

> (C1) either $X_1 \le \cdots \le X_n$ are regularly spaced on a fixed interval $\mathcal{I}$, or they represent values of a sample of size n from a fixed, continuous distribution F; and in the latter case, assume that for each n, $X_1, \cdots, X_n$ are stochastically independent of the perturbations ϵ_i's.

If the X_i's are regularly spaced on $\mathcal{I}$ then, for future reference, we take F to be the uniform distribution on $\mathcal{I}$. In this notation, assume that

> (C2) $g(x|\theta)$ has three derivatives with respect to θ, bounded uniformly in x and in θ in a neighbourhood of θ^0; the $r \times r$ matrix $A_1(\theta)$, of which the (k, l)-th element is
>
> $$\int \frac{\partial g(x|\theta)}{\partial \theta_k} \, \frac{\partial g(x|\theta)}{\partial \theta_l} \, dF(x) \, ,$$
>
> is nonsingular when $\theta = \theta^0$; the time series $\{e_i\}$ is causal in the sense that $1 + \sum_{1 \le j \le s} \alpha_j^0 z^j \ne 0$ for all $|z| \le 1$; $\psi(\theta, \alpha)$ has two bounded derivatives in a neighbourhood of (θ^0, α^0), and the $t \times t$ matrix
>
> $$\left(1 + \sum_{i=1}^{s} \alpha_i\right)^{-2} \frac{\partial \psi}{\partial \theta} A_1(\theta)^{-1} \left(\frac{\partial \psi}{\partial \theta}\right)^T + \frac{\partial \psi}{\partial \alpha} A_2(\alpha)^{-1} \left(\frac{\partial \psi}{\partial \alpha}\right)^T \qquad (4.1)$$
>
> evaluated at (θ^0, α^0), is nonsingular.

In our technical arguments we do not address the possibility that in the definition of $L(\psi_\star)$ at (2.6), local minima occur at more than one value of p. To circumvent this issue we assume, in the theorem below, that $\widehat{p} = \widehat{p}(\psi_\star)$ is taken to be the multinomial distribution which, of all those that produce a local minimum of $D_\rho(p)$ subject to $\widehat{\psi}(p) = \psi_\star$, minimises $||\widehat{\theta}(p) - \widehat{\theta}(p_{\text{unif}})|| + ||\widehat{\alpha}(p) - \widehat{\alpha}(p_{\text{unif}})||$. A proof of Theorem 4.1 is given in Appendix 2.

Theorem 4.1. Under conditions (C1) and (C2), $(2m)^{-1}L(\psi^0)$ is asymptotically distributed as χ^2 with t degrees of freedom.

If in addition to (C1) and (C2) we assume that the distribution of ϵ_t is absolutely continuous and has sufficiently many finite moments, and that $\psi(\theta, \alpha)$ and $g(\cdot|\theta)$ have sufficiently many derivatives with respect to θ and α, then we may derive an Edgeworth expansion of the distribution of $L(\psi^0)$:

$$P\{L(\psi^0) \leq u\} = P(\chi_t^2 \leq u) + n^{-1}\, q_1(x)\, \xi(x) + \cdots + n^{-k}\, q_k(x)\, \xi(x) + O(n^{-(k+1)})\,, \qquad (4.2)$$

where ξ denotes the χ_t^2 density and $q_1, \cdots, q_k$ are polynomials. It follows that calibration using the χ^2 approximation produces a coverage error of order n^{-1}. A bootstrap version of (4.2) may likewise be developed, from which it may be shown that the bootstrap-calibrated approach suggested in §2.3 produces confidence regions with coverage errors $O(n^{-2})$.

We shall not give formal theoretical results in the context of applications to robustness, or for inference under constraints. In those cases the mathematical proofs needed are generally simpler, since they relate to convergence in probability to points or deterministic functions, rather than to convergence in distribution. An instance where the argument is relatively complex, and similar to that in the case of empirical likelihood, is that of implementing a bootstrap test using tilting methods, mentioned at the end of section 2.

The power of such a test, against local alternatives, can be described by developing an Edgeworth expansion analogous to that at (4.2). In this way it can be shown that the level of the test is in error by only $O(n^{-1})$, and the test is capable of distinguishing local alternatives that are distant $n^{-1/2}$ from the null. In particular, if the distribution $F_{n,c}$ of the errors ϵ_i can be written as $F_{n,c} = (1 - cn^{-1/2})\, G + cn^{-1/2} H$, where G and H are fixed distributions, G has zero skewness, H has nonzero skewness, and $c > 0$; if both G and H are continuous with sufficiently many finite moments; and if the probability that the bootstrap test rejects the null hypothesis of zero skewness is denoted by $\pi(n, c)$; then the methods leading to Theorem 4.1 may be employed to prove that

$$\pi(n, 0) = \pi + O(n^{-1})\,, \quad \lim_{c\to\infty} \liminf_{n\to\infty} \pi(n, c) = 1\,.$$

Appendix 1: Derivation of (2.8)

Differentiating (2.10) and (2.11) with respect to p_k we obtain,

$$A_{vk}(\theta, \alpha, p)\, \frac{\partial\theta}{\partial p_k} + B_{vk}(\theta, \alpha, p)\, \frac{\partial\alpha}{\partial p_k} = V_{vk}\,, \qquad v = 1, 2, \qquad (A.1)$$

where the cases $v = 1$, 2 correspond to (2.10), (2.11) respectively, the matrices A_{1k}, B_{1k}, A_{2k} and B_{2k} are r-by-r, r-by-s, s-by-r and s-by-s, respectively, and V_{1k} and V_{2k} are r- and s-vectors, respectively. Specifically, the u-th rows of A_{1k} and B_{1k} are

$$\sum_{j=s+1}^{n} \left[\xi_{ju}(\theta,\alpha)\, \xi_j(\theta,\alpha)^T - Z_j(\theta,\alpha) \left\{\ddot{g}_u(X_j|\theta) + \sum_{i=1}^{s} \alpha_i\, \ddot{g}_u(X_{j-i}|\theta)\right\}^T\right] p_j$$

and

$$-\sum_{j=s+1}^{n} \left\{\xi_{ju}(\theta,\alpha)\, \eta_j(\theta)^T + Z_j(\theta,\alpha)\, \zeta_{ju}(\theta)^T\right\}\, p_j$$

respectively, where $\xi_{ju}(\theta,\alpha) = \dot{g}_u(X_j|\theta) + \sum_{i=1}^{s} \alpha_i\, \dot{g}_u(X_{j-i}|\theta)$ and $\xi_j = (\xi_{j1},\ldots,\xi_{jr})^T$; $\eta_j(\theta)$ and $\zeta_{ju}(\theta)$ are the s-vectors with i-th elements $\eta_{ji}(\theta) = Y_{j-i} - g(X_{j-i}|\theta)$ and $\dot{g}_u(X_{j-i}|\theta)$, respectively; and $\ddot{g}_u(x|\theta)$ is the r-vector $\partial \dot{g}_u(x|\theta)/\partial\theta$. The u-th element of V_{1k} equals $Z_k(\theta,\alpha)\, \xi_{ku}(\theta,\alpha)$. The v-th rows of A_{2k} and B_{2k} are

$$\sum_{j=s+1}^{n} \left\{\eta_{jv}(\theta)\, \xi_j(\theta,\alpha)^T + Z_j(\theta,\alpha)\, \dot{g}(X_{j-v}|\theta)^T\right\} p_j \quad \text{and} \quad -\sum_{j=s+1}^{n} \eta_{jv}(\theta)\, \eta_j(\theta)^T\, p_j$$

respectively, and the v-th element of V_{2k} equals $Z_k(\theta,\alpha)\, \eta_{kv}(\theta)$. Recall that $Z_j(\theta,\alpha)$ was defined at (2.14). Formula (2.8) follows from (A.1).

Appendix 2: Derivation of Theorem 4.1

For the sake of brevity we treat only the case $0 \le \rho < 1$ and $\psi = \psi(\theta,\alpha)$. Then a Lagrange multiplier argument shows that

$$p_k = m^{-1}\,(1 + \lambda_1 + \lambda_2^T \delta_k)^{-1/(1-\rho)}, \tag{A.2}$$

where λ_1 is a scalar, λ_2 and δ_k are t-vectors, $\delta_k = \partial\psi\{\widehat{\theta}(p), \widehat{\alpha}(p)\}/\partial p_k$, and λ_1, λ_2 are determined by the constraints $\psi\{\widehat{\theta}(p), \widehat{\alpha}(p)\} = \psi_\star$ and $\sum_k p_k = 1$.

Let $||\cdot||$ denote the Euclidean norm applied to a vector or matrix. Put $\widehat{\mu} = m^{-1}\sum_k \delta_k$, $\widehat{M} = m^{-1}\sum_k \delta_k\delta_k^T$ and $\widetilde{\psi} = \psi(\widetilde{\theta}, \widetilde{\alpha})$. Writing ψ for $\psi(\theta,\alpha)$, let $\psi_1 = \partial\psi/\partial\theta$ and $\psi_2 = \partial\psi/\partial\alpha$. The results in (A.3) below may be proved to hold, the first for uniformly in

$$p \in \mathcal{P} = \left\{p \,:\, \left\|\widehat{\theta}(p) - \theta^0\right\| \le Cn^{-1/2} \quad \text{and} \quad \left\|\widehat{\alpha}(p) - \alpha^0\right\| \le Cn^{-1/2}\right\},$$

where $C > 0$ is arbitrary but fixed:

$$\begin{aligned} &\sup_k ||\delta_k|| = O_p(n^{-1/2}), \quad ||\widehat{\mu}|| = O_p(n^{-1/2}), \quad m^{-1}\sum_k ||\delta_k||^3 = O_p(n^{1/2}), \\ &||\widehat{M}|| = O_p(1), \qquad \lambda_1 = O_p(n^{-1}), \qquad ||\lambda_2|| = O_p(n^{-1/2}). \end{aligned} \tag{A.3}$$

Indeed, the first bound in (A.3) follows by Taylor expansion and the smoothness of ψ, and implies the next three bounds. To obtain the two remaining formulae one initially assumes their correctness and carries the proof below to its conclusion, obtaining (after only a minor elaboration of the argument) proper "in probability" limits for each of $n\lambda_1$ and $n^{1/2}\lambda_2$. This proves the existence of Lagrange multipliers (λ_1, λ_2) which satisfy the desired constraints and are such that (A.3) holds. Since the multipliers are uniquely defined then the fifth and sixth results in (A.3) must hold without precondition.

In view of (A.2) and (A.3),

$$1 = \sum_k p_k = 1 - (1-\rho)^{-1}\,(\lambda_1 + \lambda_2^T \widehat{\mu}) + \tfrac{1}{2}\,(1-\rho)^{-2}\,(2-\rho)\,\lambda_2^T\,\widehat{M}\,\lambda_2 + o_p(n^{-1})\,, \tag{A.4}$$

$$\psi_\star = \psi\{\widehat{\theta}(p), \widehat{\alpha}(p)\} = \widetilde{\psi} - (1-\rho)^{-1}\,\widehat{M}\,\lambda_2 + o_p(n^{-1/2})\,. \tag{A.5}$$

We shall prove shortly that

$$\widehat{M} = M + o_p(1)\,, \tag{A.6}$$

where M is a fixed, nonsingular matrix. Therefore, by (A.4) and (A.5),

$$\lambda_1 = (1-\rho)\,\widehat{\mu}^T\,M^{-1}\,(\psi_\star - \widetilde{\psi}) + \tfrac{1}{2}\,(1-\rho)\,(2-\rho)\,(\psi_\star - \widetilde{\psi})^T\,M^{-1}\,(\psi_\star - \widetilde{\psi}) + o_p(n^{-1})\,, \tag{A.7}$$

$$\lambda_2 = -(1-\rho)\,M^{-1}\,(\psi_\star - \widetilde{\psi}) + o_p(n^{-1/2})\,. \tag{A.8}$$

By (A.2), (A.3) and (A.6),

$$a \equiv 1 - m^{-1} \sum_k (mp_k)^\rho = \rho\,(1-\rho)^{-1}\,(\lambda_1 + \lambda_2^T\widehat{\mu}) - \tfrac{1}{2}\,\rho\,(1-\rho)^{-2}\,\lambda_2^T\,M\,\lambda_2 + o_p(n^{-1})\,.$$

Substituting for λ_1 and λ_2 from (A.7) and (A.8) we deduce that

$$D_\rho(p) = \{\rho(1-\rho)\}^{-1}\,a = \tfrac{1}{2}\,(\psi_\star - \widetilde{\psi})^T\,M^{-1}\,(\psi_\star - \widetilde{\psi}) + o_p(n^{-1})\,. \tag{A.9}$$

Next we derive (A.6) and identify M. Put

$$\Delta_j = Y_j - g(X_j|\theta) + \sum_{i=1}^{s} \alpha_i\{Y_{j-i} - g(X_{j-i}|\theta)\}.$$

Let $\dot{g}(\cdot|\theta)$ (respectively, $\ddot{g}(\cdot|\theta)$) denote the r-vector ($r \times r$ matrix) of the first (second) derivatives of $g(\cdot|\theta)$ with respect to θ, and define (a) $U_j = \dot{g}(X_j|\theta) + \sum_i \alpha_i\,\dot{g}(X_{j-i}|\theta)$, (b) V_j to be the s-vector with i-th element $Y_{j-i} - g(X_{j-i}|\theta)$, (c) $W_j = \ddot{g}(X_j|\theta) + \sum_i \alpha_i\,\ddot{g}(X_{j-i}|\theta)$, and (d) Z_j to be the $r \times s$ matrix with i-th column $\dot{g}(X_{j-i}|\theta)$. Then for S defined in (2.2),

$$\frac{\partial S}{\partial \theta} = -2 \sum_{j=s+1}^{n} p_j\,\Delta_j\,U_j\,, \quad \frac{\partial S}{\partial \alpha} = 2 \sum_{j=s+1}^{n} p_j\,\Delta_j\,V_j\,. \tag{A.10}$$

If $\theta = \widehat{\theta}(p)$ and $\alpha = \widehat{\alpha}(p)$ are chosen to minimise $S = S(\theta, \alpha | p)$, then both the derivatives at (A.10) vanish, and so

$$0 = -\frac{1}{2}\, \frac{\partial^2 S}{\partial p_k \partial \theta} = \sum_{j=s+1}^{n} p_j \left(V_j^T \frac{\partial \alpha}{\partial p_k} - U_j^T \frac{\partial \theta}{\partial p_k} \right) U_j + \sum_{j=s+1}^{n} p_j\, \Delta_j \left(W_j \frac{\partial \theta}{\partial p_k} + Z_j \frac{\partial \alpha}{\partial p_k} \right) + \Delta_k\, U_k \,,$$

$$0 = \frac{1}{2}\, \frac{\partial^2 S}{\partial p_k \partial \alpha} = \sum_{j=s+1}^{n} p_j \left(V_j^T \frac{\partial \alpha}{\partial p_k} - U_j^T \frac{\partial \theta}{\partial p_k} \right) V_j - \sum_{j=s+1}^{n} p_j\, \Delta_j\, Z_j^T \frac{\partial \theta}{\partial p_k} + \Delta_k\, V_k \,.$$

Equivalently,

$$M_{11} \frac{\partial \theta}{\partial p_k} + M_{12} \frac{\partial \alpha}{\partial p_k} = \Delta_k\, U_k \,, \quad M_{21} \frac{\partial \theta}{\partial p_k} + M_{22} \frac{\partial \alpha}{\partial p_k} = \Delta_k\, V_k \,, \tag{A.11}$$

where

$$M_{11} = \sum_{j=s+1}^{n} p_j\, (U_j\, U_j^T - \Delta_j\, W_j) \,, \quad M_{12} = - \sum_{j=s+1}^{n} p_j\, (U_j\, V_j^T + \Delta_j\, Z_j) \,,$$

$$M_{21} = \sum_{j=s+1}^{n} p_j\, (V_j\, U_j^T + \Delta_j\, Z_j^T) \,, \quad M_{22} = - \sum_{j=s+1}^{n} p_j\, V_j\, V_j^T \,.$$

Let $\Delta_j^0, U_j^0, V_j^0, W_j^0$ and Z_j^0 denote the values assumed by Δ_j, U_j, V_j, W_j and Z_j when (θ, α) is replaced by (θ^0, α^0); let Q_j be the s-vector with i-th element e_{j-i}; and observe that $\Delta_j^0 = \epsilon_j$ and $V_j^0 = Q_j$. Using Taylor expansion we may prove that, uniformly in θ and α such that $||\theta - \theta^0|| \leq C n^{-1/2}$ and $||\alpha - \alpha^0|| \leq C n^{-1/2}$, and uniformly in $s + 1 \leq j \leq n$, we have $\Delta_j = \epsilon_j + Q_j^T\, (\alpha - \alpha^0) + O_p(n^{-1/2})$, $U_j = U_j^0 + O_p(n^{-1/2})$, $V_j = Q_j + O_p(n^{-1/2})$, $W_j = W_j^0 + O_p(n^{-1/2})$ and $Z_j = Z_j^0 + O_p(n^{-1/2})$. From these formulae, (A.2), (A.3) and the fact that the design points X_j are independent of the perturbations ϵ_k, we may deduce that $M_{11} = M_1 + o_p(1)$, $M_{22} = M_2 + O_p(n^{-1/2})$, $||M_{12}|| + ||M_{21}|| = O_p(n^{-1/2})$, where $M_1 = \operatorname{plim} m^{-1} \sum_j E\{U_j^0 (U_j^0)^T\}$ and $M_2 = E(Q_0 Q_0^T)$. Both M_1 and M_2 are nonsingular. (In notation of §4, $M_1 = (1 + \sum_i \alpha_i)^2 A_1$ and $M_2 = A_2$.) Hence, by (A.11), and defining $\widehat{\psi}_j(p) = \psi_j\{\widehat{\theta}(p), \widehat{\alpha}(p)\}$, we have

$$\begin{aligned} \delta_k &= \widehat{\psi}_1(p)\, \frac{\partial \widehat{\theta}(p)}{\partial p_k} + \widehat{\psi}_2\, \frac{\partial \widehat{\alpha}(p)}{\partial p_k} \\ &= \{\widehat{\psi}_1(p)\, M_{11}^{-1}\, U_k + \widehat{\psi}_2(p)\, M_{22}^{-1}\, V_k\}\, \Delta_k + O_p \left\{ n^{-1/2} \left(\left|\left| \frac{\partial \widehat{\theta}(p)}{\partial p_k} \right|\right| + \left|\left| \frac{\partial \widehat{\alpha}(p)}{\partial p_k} \right|\right| \right) \right\} . \end{aligned}$$

Therefore, writing $\psi_j^0 = \psi_j(\theta^0, \alpha^0)$ and $\xi_k = \psi_1^0\, M_1^{-1}\, U_k^0 + \psi_2^0\, M_2^{-1}\, Q_k$, we have

$$\widehat{M} = m^{-1} \sum_{k=s+1}^{n} \xi_k\, \xi_k^T\, \epsilon_k^2 + o_p(1). \tag{A.12}$$

Here we have used the property,

$$\sum_{k=s+1}^{n} \left(\left|\left| \frac{\partial \widehat{\theta}(p)}{\partial p_k} \right|\right|^2 + \left|\left| \frac{\partial \widehat{\alpha}(b)}{\partial p_k} \right|\right|^2 \right) = o_p(n^{1/2}) \,,$$

which may be proved from (A.11). The remainders here and at (A.12) are of the stated orders uniformly in $p \in \mathcal{P}$.

The autoregression $\epsilon_i = \sum_{0 \le j \le s} \alpha_j \, e_{i-j}$, in which $\alpha_0 = 1$, may be inverted to give a moving average, $e_i = \sum_{j \ge 0} \beta_j \, \epsilon_{i-j}$, say. It follows that the variables $\{e_{j-i}, 1 \le i \le s\}$ are independent of ϵ_j. This property and the law of large numbers may be used to prove that (A.12) implies (A.6), with

$$M = \{\psi_1^0 \, M_1^{-1} \, (\psi_1^0)^T + \psi_2^0 \, M_2^{-1} \, (\psi_2^0)^T\} \, E(\epsilon_0^2) \, . \tag{A.13}$$

Note that $M/E(\epsilon_0^2)$ is the matrix defined at (4.1), evaluated at (θ^0, α^0).

Finally we derive the required asymptotic distribution. Note that

$$\widetilde{\psi} = \psi^0 + \psi_1^0 \, (\widetilde{\theta} - \theta^0) + \psi_2^0 \, (\widetilde{\alpha} - \alpha^0) + o_p(n^{-1/2}) \, . \tag{A.14}$$

Taking $p = p_{\mathrm{unif}}$ in (A.10), equating $\partial S/\partial\theta$ and $\partial S/\partial\alpha$ to 0, Taylor expanding the right hand sides (which are functions of $(\widetilde{\theta}, \widetilde{\alpha})$) about (θ^0, α^0), and solving for $(\widetilde{\theta}, \widetilde{\alpha})$, we deduce that

$$M_1 \, (\widetilde{\theta} - \theta^0) = \frac{1}{m} \sum_{k=s+1}^{n} \epsilon_k \, U_k^0 + o_p(n^{-1/2}) \, , \quad M_2 \, (\widetilde{\alpha} - \alpha^0) = \frac{1}{m} \sum_{k=s+1}^{n} \epsilon_k \, Q_k + o_p(n^{-1/2}) \, .$$

From these results and (A.14) we may deduce that $\widetilde{\psi} = \psi^0 + n^{-1/2} N_n'$, where N_n' is asymptotically Normally distributed with mean zero and covariance matrix M defined at (A.13). Combining this result with (A.9) we deduce that

$$2 \, n \, D(\widehat{p}) = (N_n - \gamma_n)^T \, (N_n - \gamma_n) \, , \tag{A.15}$$

where $\gamma_n = n^{1/2}(\psi_\star - \psi^0)$ and N_n is asymptotically Normal $N(0, I_t)$. Theorem 4.1 follows from (A.15).

References

Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods.* Springer, New York.

Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann. Statist. **14**, 1171–1179.

Chen, S.X. (1996). Empirical likelihood confidence intervals for nonparametric density estimation. *Biometrika* **83**, 329–341.

Cressie, N.A.C., and read, T.R.C. (1984). Multinomial goodness-of-fit tests. *J. Roy. Statist. Soc. Ser. B* **46**, 440–464.

Davison, A.C., Hinkley, D.V. and Worton, B.J. (1992). Bootstrap likelihoods. *Biometrika* 79, 113–130.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. (With Discussion.) *Canad. J. Statist.* **36**, 369–401.

Grenander, U. (1956). On the theory of mortality measurement, II. *Skand. Akt.* **39**, 125–153.

Hall, P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl.* **20**, 231–246.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer, New York.

Hall, P. and Owen, A.B. (1993). Empirical likelihood confidence bands in density estimation. *J. Comput. Graph. Statist.* **2**, 273–289.

Hall, P. and Presnell, B. (1998). Applications of intentionally biased bootstrap methods. *Proceedings of the International Congress of Mathematicians*, Vol. III (Berlin, 1998). *Doc. Math.* Extra Vol. III, 257–266.

Hall, P. and Presnell, B. (1999a). Intentionally biased bootstrap methods. *J. R. Stat. Soc.* Ser. B **61**, 143–158.

Hall, P. and Presnell, B. (1999b). Biased bootstrap methods for reducing the effects of contamination. *J. R. Stat. Soc. Ser.* B **61**, 661–680.

Hollander, M., McKeague, I.W. and Yang, J. (1997). Likelihood ratio-based confidence bands for survival functions. *J. Amer. Statist. Assoc.* **92**, 215–226.

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **33**, 457–481.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887–906.

Kitamura, Y. (1997). Empirical likelihood methods with weakly dependent processes. *Ann. Statist.* **25**, 2084–2102.

Künsch, H.R. (1989). The jackknife and the bootstrap for general stationary observations. Ann. Statist. **17**, 1217–1241.

Monti, A.C. (1997). Empirical likelihood confidence regions in time series models. *Biometrika* **84**, 395–405.

Mykland, P.A. (1995). Dual likelihood. *Ann. Statist.* **23**, 396–421.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90–120.

Owen, A. (1995). Nonparametric likelihood confidence bands for a distribution function. *J. Amer. Statist. Assoc.* **90**, 516–521.

Owen, A. (2001). *Empirical Likelihood.* Chapman & Hall/CRC, London.

Rényi, A. (1961). On measures of entropy and information. *Proc. 4th Berkeley Symp. Math. Statist. Prob.* **1**, 547–561.

Thomas, D.R. and Grunkemeier, G.L. (1975). Confidence interval estimation of survival probabilities for censored data. *J. Amer. Statist. Assoc.* **70**, 865–871.

Read, T. and Cressie, N. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data.* Springer, New York.

Whittle, P. (1953). Estimation and information in stationary time series. *Ark. Mat.* **2**. 423–434.

Zhang, B. (1996). Confidence intervals for a distribution function in the presence of auxiliary information. *Comput. Statist. Data Anal.* **21**, 327–342.

Zhang, B. (1998). A note on kernel density estimation with auxiliary information. *Comm. Statist. Theory Methods* **27**, 1–11.