SJS 07-176 Revised

Kernel Likelihood Inference for Time Series

Carlo Grillenzoni
University IUAV of Venice
30135 Venezia, ITALY
(carlog@iuav.it)

Abstract. This paper develops nonparametric techniques for dynamic models whose data have unknown probability distributions. Point estimators are obtained from the maximization of a semiparametric likelihood function built on the kernel density of the disturbances. This approach can also provide Kullback-Leibler cross-validation estimates of the bandwidth of the kernel densities. Confidence regions are derived from the dual-empirical likelihood method based on nonparametric estimates of the scores. Limit theorems for martingale difference sequences support the statistical theory; moreover, simulation experiments and a real case study show the validity of the methods.

Key Words. Adaptive estimation, ARMAX models, cross validation, dual likelihood, empirical likelihood, kernel density, martingale difference.

Acknowledgments: I am grateful to the Editors and the Referees for their useful suggestions.

1. Introduction

This paper develops inferential techniques for dynamic regression models whose disturbances (noise) have a probability distribution which is totally unknown. In this case, it is well known that the maximum likelihood (ML) method cannot be applied, and least squares (LS) can produce inefficient estimates. A possible solution to this drawback is provided by *adaptive* estimation (e.g. Bickel, 1982; Kreiss, 1987), which implements a one-step ML algorithm on the basis of nonparametric estimates of the scores, computed on LS residuals. The adaptive approach has been extended to complex dynamic models (e.g. Koul and Shick, 1997 and Ling, 2003); however, it still suffers by two fundamental problems. The first is that it needs consistent initial estimates for all parameters; the second one is that it conducts inference only on the basis of the asymptotic distribution of estimates.

By embedding a kernel estimator of the noise density into the parametric likelihood function, one can obtain a semiparametric functional which can be optimized either with respect to the regression coefficients of the dynamic model, and the bandwidth of the kernel density. This approach has a close relationship with the cross-validation method discussed in nonparametric literature (e.g. Wand and Jones, 1995), and shares its optimality properties. The final result is a semiparametric ML estimator which allows for significant gain of efficiency and unbiasedness over the LS one. Simulations quantify this gain about 25% on average.

Also the finite sample inference can be treated with nonparametric techniques named empirical likelihood (EL, e.g. Owen, 1990, 2001). This method can build tests and confidence regions without knowing the sampling distribution and is an efficient alternative to bootstrap. In our context, the EL approach can be directly developed on gradient and score quantities of the semiparametric estimator, because they constitute martingale difference series. Indeed, Mykland (1995) has extended classical results of Owen, which were developed for independent sequences, to martingale series. This solution has computational and analytical advantages over existing EL approaches for time series, such as the "blockwise" one of Kitamura (1997) or the spectral one of Monti (1997).

2. Point Estimation based on Kernel Likelihood

Let $\{x_t, y_t\}$ be the input and the output of a stochastic dynamical system which is representable through an ARMAX model with orders $(p, q; d, b) \geq 0$

$$y_t = \sum_{i=1}^{p} \phi_i \, y_{t-i} + \sum_{i=0}^{d} \delta_i \, x_{t-b-i} + \sum_{i=1}^{q} \theta_i \, e_{t-i} + e_t \,, \qquad e_t \sim \text{IID}(f(e)) \tag{1}$$

where b is the delay factor and $\{e_t\}$ is a noise process. This linear model is widely used in economics, environmetrics and engineering for forecasting and control purposes (see Box, Jenkins and Reinsel, 1994). This paper, by assuming the distribution of $\{x_t, y_t\}$ unknown, investigates (1) in terms of nonparametric inference.

Using polynomials in the lag operator L, where $L^b x_t = x_{t-b}$, the model can also be written in compact form as: $\phi_p(L)\, y_t = \delta_d(L)\, x_{t-b} + \theta_q(L)\, e_t$, where $\phi_p(L) = 1 - \sum_{i=1}^{p} \phi_i L^i$, etc.. On this scheme, two assumptions are made:

Assumption A1. The sequence $\{e_t\}$ is independent and identically distributed (IID), and is independent of $\{x_t\}$. It has a stationary density $f(e_t) = f(e)$ which has zero mean, finite variance σ_e^2 and finite Fisher information τ_f. The shape of $f(\cdot)$ is differentiable but unknown and may be asymmetric and multi-modal.

Assumption A2. The polynomials $\phi_p(L), \theta_q(L)$ have no common factor and are stable (i.e. their roots lie outside the unit circle in the complex plane), and $\delta_d(L)$ has bounded coefficients. The input $\{x_t\}$ and the initial values $\boldsymbol{z}_0' = [\, y_0 \ldots y_{p-1}, x_{-b} \ldots x_{-b-d}, e_0 \ldots e_{q-1} \,]$ are fixed or come from stationary distributions.

Under these assumptions, the dynamical system (1) is stationary and invertible (that is, e_t can be uniquely derived from y_t) and is parametrically identified.

2.1. Quasi ML

Using a vector notation, the model (1) can be written in "regression" form as:

$$\begin{aligned} y_t &= \boldsymbol{\beta}' \boldsymbol{z}_t + e_t \,, \qquad t = 1, 2 \ldots n \,, \\ \boldsymbol{\beta}' &= [\, \phi_1 \ldots \phi_p, \delta_0 \ldots \delta_d, \theta_1 \ldots \theta_q \,] \,, \\ \boldsymbol{z}_t' &= [\, y_{t-1} \ldots y_{t-p}, x_{t-b} \ldots x_{t-b-d}, e_{t-1} \ldots e_{t-q} \,] \,, \end{aligned} \tag{2}$$

where $\boldsymbol{z}_t$ is the vector of pseudo-regressors and depends on $\boldsymbol{\beta}$. In this framework, it

is natural to apply nonlinear least squares (NLS) to estimate the parameters

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_n &= \arg\min \left[\frac{1}{n} \sum_{t=1}^{n} e_t^2(\boldsymbol{\beta}) \right] \\ e_t(\boldsymbol{\beta}) &= y_t - \boldsymbol{\beta}' \boldsymbol{z}_t(\boldsymbol{\beta}) \end{aligned} \tag{3}$$

where n is the sample size. Under the condition $f(e)$ Gaussian, it is well known that NLS estimator is equivalent to the ML method. For this reason, (3) is also known as *quasi* or *pseudo* ML estimator (see White, 1996).

Assuming differentiability of the residual function $e_t(\boldsymbol{\beta})$, it is easy to show (e.g. Grillenzoni, 1991) that first derivatives satisfy a the dynamic relationship

$$\begin{aligned} \boldsymbol{\zeta}_t(\boldsymbol{\beta}) &= -\frac{\partial\, e_t(\boldsymbol{\beta})}{\partial\, \boldsymbol{\beta}} = \frac{1}{\theta_q(L)} \boldsymbol{z}_t(\boldsymbol{\beta}) \\ &= \boldsymbol{z}_t(\boldsymbol{\beta}) + \sum_{i=1}^{q} \theta_i\, \boldsymbol{\zeta}_{t-i}(\boldsymbol{\beta}) \end{aligned} \tag{4}$$

Hence, the gradient $\{\boldsymbol{\zeta}_t\}$ is a stationary process under the condition of invertibility of the model (1); moreover, its past and present values $\boldsymbol{\zeta}_{t-k}$, $k \geq 0$ are independent of the innovations e_{t+k}. On the basis of the expression (4), one can easily compute the explicit iterative version of the estimator (3)

$$\tilde{\boldsymbol{\beta}}_n^{(i+1)} = \tilde{\boldsymbol{\beta}}_n^{(i)} + \left(\sum_{t=1}^{n} \tilde{\boldsymbol{\zeta}}_t^{(i)}\, \tilde{\boldsymbol{\zeta}}'^{(i)}_t \right)^{-1} \sum_{t=1}^{n} \tilde{\boldsymbol{\zeta}}_t^{(i)}\, \tilde{e}_t^{(i)} \tag{5}$$

where $\tilde{\boldsymbol{\zeta}}_t, \tilde{\boldsymbol{z}}_t, \tilde{e}_t$ are evaluated at the point $\tilde{\boldsymbol{\beta}}_n^{(i)}$. This expression is useful for obtaining the dispersion matrix of the NLS estimator.

Under the assumptions A1, A2, the estimators (3) and (5) are consistent for the population value $\boldsymbol{\beta}$ which minimizes the expectation $\mathrm{E}[\, e_t^2(\boldsymbol{\beta})\,]$; moreover, it converges in law as $n \to \infty$ (see Box *et al.*, 1994 Chap. 7)

$$\sqrt{n}\left(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta} \right) \quad \xrightarrow{\mathrm{L}} \quad \mathrm{N}\Big[\mathbf{0}\,,\, \mathrm{E}\big(\boldsymbol{\zeta}_t\, \boldsymbol{\zeta}_t' \big)^{-1} \sigma_e^2 \Big] \tag{6}$$

The asymptotic dispersion in (6) is equivalent to that of the ML estimator in the case of $f(e)$ Gaussian; it does not provide, however, the lower bound in the general case. One may wonder if the efficiency of the NLS method can be improved by means of nonparametric techniques.

2.2. Kernel ML

Given the NLS residuals $\tilde{e}_t = (y_t - \tilde{\boldsymbol{\beta}}_n{}' \tilde{\boldsymbol{z}}_t)$, smoothing techniques can be used for testing, see Azzalini, Bowman and Härdle (1989). For example, the kernel density

$$\tilde{f}_\kappa(e) = \frac{1}{n\kappa} \sum_{t=1}^{n} K\Big(\frac{e - \tilde{e}_t}{\kappa}\Big), \quad e \in \Re$$

can tentatively is useful to identify the functional form of $f(e)$. In the above, we just recall that $K(\cdot)$ is the kernel function (a symmetric density with mean zero and unit variance) and $\kappa \geq 0$ is the bandwidth coefficient.

Kernel techniques have also been used in *Adaptive* ML estimation to improve the efficiency of the initial NLS method (e.g. Kreiss, 1987 and Ling, 2003). In this context, the density $\tilde{f}_\kappa(\tilde{e}_t)$ is used for computing the scores of the likelihood function, so that a *one-step* Newton-Raphson estimator could be implemented. However, the entire procedure requires a number of adjustments (such as discretization of the initial $\tilde{\boldsymbol{\beta}}_n$, sample splitting of the residuals, a-priori selection of the bandwidth κ, trimming of the kernel scores ψ_t, outer-product of gradient computation of the Hessian, etc.), which actually hinder the efficacy of the method.

To avoid these drawbacks altogether, one can imbed the kernel density in the likelihood function, and directly optimize the resulting functional. The implied solution can be termed kernel maximum likelihood (KML) estimator

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{n\kappa} &= \arg\max \left[\frac{1}{n} \sum_{t=1}^{n} \log\, f_\kappa(e_t(\boldsymbol{\beta})) \right] \\ f_\kappa(e_t(\boldsymbol{\beta})) &= \frac{1}{n\kappa} \sum_{j=1}^{n} K\left(\frac{e_t(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})}{\kappa} \right) \end{aligned} \tag{7}$$

In this context, the likelihood function can be viewed as a mixture of densities defined by the kernel functions. Because the nuisance parameter κ is invariant with t, it is parametrically identified and can be included in the estimation framework (7). However, to avoid the trivial solution $\kappa \to 0$, the term $K(0/\kappa)$ must be excluded from the loss function, so that the estimator becomes

$$\Big[\hat{\boldsymbol{\beta}}_{n\kappa}\,, \hat{\kappa}_n \Big] = \arg\max \left\{ \frac{1}{n} \sum_{t=1}^{n} \log \left[\sum_{j \neq t}^{n} K\Big(\frac{e_t(\boldsymbol{\beta}) - e_j(\boldsymbol{\beta})}{\kappa} \Big) \right] - \log((n-1)\kappa) \right\} \tag{8}$$

In the absence of $\boldsymbol{\beta}$, this approach coincides with the maximum likelihood cross-validation (MLCV) selection of the bandwidth (e.g. Wand and Jones, 1995, Chap.3), which is closely related to the minimum Kullback-Leibler distance method discussed in Bowman, Hall and Titterington (1984). Given possible non-smoothness of the objective function (8), optimization can be carried out with stochastic search methods, such as simulated annealing or genetic algorithms.

Having assumed differentiability of $f(e_t)$, the analytical expression of the scores of the kernel likelihood function (7) is given by

$$\begin{aligned} \boldsymbol{\xi}_t(\boldsymbol{\beta}) &= \frac{\partial \log f(e_t(\boldsymbol{\beta}))}{\partial \boldsymbol{\beta}} = \frac{\partial f(e_t)/\partial e_t}{f(e_t)} \frac{\partial e_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\ &= \psi(e_t)\, \boldsymbol{\zeta}_t(\boldsymbol{\beta}) \end{aligned} \tag{9}$$

where $\psi(e_t) = \dot{f}(e_t)/f(e_t)$ is the score of the noise density and $\boldsymbol{\zeta}_t$ is the gradient of the NLS estimator. Since $\psi(\cdot)$ only depends on e_t, it is independent of the values $\boldsymbol{\zeta}_{t-k}$, $k \geq 0$ and has zero mean. It follows that the score $\{\boldsymbol{\xi}_t\}$ forms a martingale difference sequence, that is $\mathrm{E}(\boldsymbol{\xi}_t \,|\, y_{t-k}, x_{t-k}; k \geq 1) = \mathbf{0}$.

As regards the second derivatives $\boldsymbol{\Xi}_t(\boldsymbol{\beta}) = \partial \boldsymbol{\xi}_t(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$, it is well know that the Fisher information matrix $-\mathrm{E}\big(\boldsymbol{\Xi}_t\big) = \mathrm{E}\big(\boldsymbol{\xi}_t \boldsymbol{\xi}_t'\big)$. Therefore, we have that $n^{-1} \sum_t \boldsymbol{\Xi}_t = -\, n^{-1} \sum_t \boldsymbol{\xi}_t \boldsymbol{\xi}_t' + o_p(1)$, and the iterative expression of the estimator of (7) can be based on the outer-product of gradient computation of the Hessian matrix:

$$\hat{\boldsymbol{\beta}}_{n\kappa}^{(i+1)} = \hat{\boldsymbol{\beta}}_{n\kappa}^{(i)} + \left(\sum_{t=1}^{n} \hat{\boldsymbol{\xi}}_t^{(i)}\, \hat{\boldsymbol{\xi}}'^{(i)}_t \right)^{-1} \sum_{t=1}^{n} \hat{\boldsymbol{\xi}}_t^{(i)} \tag{10}$$

where $\hat{\boldsymbol{\xi}}_t = \psi(\hat{e}_t)\, \boldsymbol{\zeta}_t(\hat{\boldsymbol{\beta}}_{n\kappa})$. This algorithm requires the equations (3), (4) and a nonparametric estimator for $\psi(\cdot)$; in the case of Gaussian kernels, we have

$$\hat{\psi}_\kappa(e) = \frac{\hat{\dot{f}}_\kappa(e)}{\hat{f}_\kappa(e)} = \frac{\sum_{j=1}^{n} K\Big(\frac{e-\hat{e}_j}{\kappa}\Big) \Big(\frac{\hat{e}_j - e}{\kappa^2}\Big)}{\sum_{j=1}^{n} K\Big(\frac{e-\hat{e}_j}{\kappa}\Big)} \tag{11}$$

which looks like a kernel regression between independent processes.

In practice, however, the computation of $\hat{\boldsymbol{\beta}}_{n\kappa}$ is usually performed by direct minimization of the functionals (7)-(8) with search methods; the algorithm (10)-(11) is useful for estimating the dispersion matrix and for obtaining the *recursive* (on-line) version of KML : $\hat{\boldsymbol{\beta}}_\kappa(t)$. As in Grillenzoni (1991), this can be obtained by

equating number of iterations and number of processed data, namely $(i = n) = t$; by recursively computing the gradient $\hat{\boldsymbol{\zeta}}(t) = \hat{\boldsymbol{z}}(t) + \sum_{j=1}^{q} \hat{\theta}_j(t)\, \hat{\boldsymbol{\zeta}}(t-j)$, and estimating the score $\hat{\psi}(t)$ with a sequential kernel density $\hat{f}_\kappa(t)$ (see Grillenzoni, 2000). In this context, the bandwidth can heuristically be selected on the basis of a error variance, as $\hat{\kappa}(t) = \hat{\sigma}_e(t)/t^{1/5}$ (see Wand and Jones, 1995).

2.3. Asymptotics

The analysis we outline in this section is different from those available in the adaptive ML literature (e.g. Bickel, 1982; Kreiss, 1987), which require the symmetry constraint on $f(e)$, a consistent initial estimator of $\boldsymbol{\beta}$ and perform various numerical adjustments, such as discretization, sample splitting and trimming (see Ling, 2003) . Our heuristic idea is that because $\hat{\boldsymbol{\beta}}_{n\kappa}$ maximizes the functional $\ell_{n\kappa}(\boldsymbol{\beta}) = n^{-1} \sum_{t=1}^{n} \log [\, f_\kappa(e_t(\boldsymbol{\beta}))\,]$, which tends to $\mathrm{E}(\ell_{n\kappa})$, then $\hat{\boldsymbol{\beta}}_{n\kappa}$ should converge to the value $\boldsymbol{\beta}_0$ which maximizes $\mathrm{E}(\ell_{n\kappa})$. As in the classical nonparametric estimation, analysis must be performed under the conditions $\kappa \to 0$, $n\kappa \to \infty$, which allow the kernel estimate $f_\kappa(\cdot)$ to converge uniformly to the unknown noise density. When this happens, the kernel likelihood function converges to its parametric version, and the KML estimator converges to the parametric ML solution.

We start by assuming *identifiability*, on the parameter space $\boldsymbol{B} \subset \Re^{p+q+d+1}$, of a stationary and invertible ARMAX model in the estimation framework (7).

Definition 1. Suppose that $n^{-1} \sum_{t=1}^{n} \mathrm{E}\big[\, \log f_\kappa(e_t(\boldsymbol{\beta}))\,\big]$ has a maximum at $\boldsymbol{\beta}_0$ for each n. Let $\boldsymbol{\Theta}_0$ be an open sphere centred at $\boldsymbol{\beta}_0$ with fixed radius $\rho > 0$, and let $\boldsymbol{\Theta}_0^c$ be its compact complement in $\boldsymbol{B}$. The maximizer $\boldsymbol{\beta}_0$ is *identifiably unique* iff

$$\inf_n \left[\min_{\boldsymbol{\beta} \in \boldsymbol{\Theta}_0^c} \left(\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\big[\log f_\kappa(e_t(\boldsymbol{\beta}_0))\big] - \frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\big[\log f_\kappa(e_t(\boldsymbol{\beta}))\big] \right) \right] > 0$$

This statement is inspired by Definition 3.3 of White (1996, p.28).

Now, let us assume that the kernel likelihood function in (7) satisfies the uniform law of large numbers (ULLN) for dependent processes.

Lemma 1. If there exists a function $D : \Re \to \Re^+$ such that $|\log f_\kappa(e(\boldsymbol{\beta}))| \leq D(e(\boldsymbol{\beta}))$ for all $\boldsymbol{\beta} \in \boldsymbol{B}$ and $e \in \Re$, and such that $\mathrm{E}(D) = \int_\Re D(e) f(e)\, \mathrm{d}e < \infty$, then

it follows that $\mathrm{E}\Big[\, \log f_\kappa(e(\boldsymbol{\beta}))\,\Big]$ is continuous on $\boldsymbol{B}$ and, uniformly on $\boldsymbol{B}$, one has

$$\sup_n \left(\frac{1}{n}\sum_{t=1}^{n} \log f_\kappa(e_t(\boldsymbol{\beta})) - \mathrm{E}\Big[\, \log f_\kappa(e(\boldsymbol{\beta}))\,\Big] \right) = o_p(1)$$

Proof. This result follows from Theorem A.2.2 in White (1996, p.351).

The boundedness condition stated in the Lemma is generally assumed in the likelihood literature to ensure consistency of ML estimates. We assume that it holds even for the model (1). The consistency property can now be established.

Theorem 1. Under the assumptions A1, A2 for model (1), the boundedness condition stated in the Lemma 1, and assuming that $\boldsymbol{\beta}_0$ is the identifiably unique maximizer of $n^{-1}\sum_{t=1}^{n} \mathrm{E}\Big[\, \log f_\kappa(e_t(\boldsymbol{\beta}))\,\Big]$ for every n, then the KML estimator (7)-(10) is such that

$$\Big(\hat{\boldsymbol{\beta}}_{n\kappa} - \boldsymbol{\beta}_0\Big) = o_p\Big(1/\sqrt{n\kappa}\Big) \qquad \text{as} \qquad \kappa \to 0\,,\ n\kappa \to \infty$$

Proof. This result is a corollary of Theorem 3.5 in White (1996, p.29)

Having established consistency, one can now derive the asymptotic distribution together with the expression of the dispersion matrix.

Theorem 2. Under the same assumptions as Theorem 1, and the conditions that $\log f_\kappa(e_t(\boldsymbol{\beta}))$ is twice continuously differentiable with score (9), the matrix $\mathrm{E}\Big[\boldsymbol{\zeta}_t(\boldsymbol{\beta})\,\boldsymbol{\zeta}_t'(\boldsymbol{\beta})\Big]$ is positive definite for any $|\boldsymbol{\beta} - \boldsymbol{\beta}_0| < \epsilon$ and $\tau_f = \mathrm{E}[\psi^2(e_t)] < \infty$, then the KML estimator (7)-(10) is such that

$$\sqrt{n\kappa}\,\Big(\,\hat{\boldsymbol{\beta}}_{n\kappa} - \boldsymbol{\beta}_0\,\Big) \quad \xrightarrow{\mathrm{L}} \quad \mathrm{N}\Big\{\,\mathbf{0}\,,\, \mathrm{E}\Big[\boldsymbol{\zeta}_t(\boldsymbol{\beta}_0)\,\boldsymbol{\zeta}_t'(\boldsymbol{\beta}_0)\Big]^{-1}\tau_f^{-1}\,\Big\} \tag{12}$$

Proof. We adopt the standard method based on a Taylor expansion of the first-order condition of (7) about $\boldsymbol{\beta}_0$. Hence:

$$\mathbf{0} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\xi}_t(\hat{\boldsymbol{\beta}}_{n\kappa}) = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\xi}_t(\boldsymbol{\beta}_0) + \left[\, \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\Xi}_t(\bar{\boldsymbol{\beta}}) \right] (\,\hat{\boldsymbol{\beta}}_{n\kappa} - \boldsymbol{\beta}_0\,)$$

where $\boldsymbol{\Xi}_t(\boldsymbol{\beta}) = \partial\,\boldsymbol{\xi}_t(\boldsymbol{\beta})/\partial\,\boldsymbol{\beta}'$ is the matrix of second derivatives and $\bar{\boldsymbol{\beta}} \in (\hat{\boldsymbol{\beta}}_{n\kappa}, \boldsymbol{\beta}_0)$ is an intermediate point. Introducing the coefficient κ, and using the outer product

of scores for computing the Hessian: $n^{-1}\sum_t \boldsymbol{\Xi}_t = -\,n^{-1}\sum_t \boldsymbol{\xi}_t\,\boldsymbol{\xi}_t' + o_p(1)$, one has

$$\sqrt{n\kappa}\,(\,\hat{\boldsymbol{\beta}}_{n\kappa} - \boldsymbol{\beta}_0\,) = \left(\frac{1}{n\kappa}\sum_{t=1}^{n}\bar{\boldsymbol{\xi}}_t\,\bar{\boldsymbol{\xi}}_t'\right)^{-1}\frac{1}{\sqrt{n\kappa}}\sum_{t=1}^{n}\boldsymbol{\xi}_t^0 + o_p(1)$$

Having $\boldsymbol{\xi}_t = \boldsymbol{\zeta}_t\,\psi(e_t)$, the term $(n\kappa)^{-1/2}\sum_t \boldsymbol{\xi}_t$ on the right hand side of the above is suitable for the application of the central limit theorem for stationary martingale difference processes (see White, 1996). Finally, by the consistency of $\bar{\boldsymbol{\beta}}$, the ergodic theorem and the conditional independence of $\boldsymbol{\zeta}_t, \psi_t$ one has

$$\left(\frac{1}{n\kappa}\sum_{t=1}^{n}\bar{\boldsymbol{\xi}}_t\,\bar{\boldsymbol{\xi}}_t'\right) \quad \xrightarrow{\text{P}} \quad \mathrm{E}_0\big(\boldsymbol{\xi}_t\,\boldsymbol{\xi}_t'\big) = \mathrm{E}_0\big(\boldsymbol{\zeta}_t\,\boldsymbol{\zeta}_t'\,\psi_e^2\big) = \mathrm{E}_0\big(\boldsymbol{\zeta}_t\,\boldsymbol{\zeta}_t'\big)\,\mathrm{E}\big(\psi_e^2\big)$$

as $\kappa \to 0$, $n\kappa \to \infty$, where $\mathrm{E}_0(\cdot)$ means that the variables are evaluated at $\boldsymbol{\beta}_0$. The result (12) then follows by the continuous mapping theorem.

Remark 1. By comparing (6) and (12), one can state that KML is more efficient than NLS, because $\sigma_e^2 \geq 1/\tau_f$, and it better approaches the parametric ML solution. This does not mean, however, that finite-sample properties and/or the speed of convergence of the semiparametric estimator are better in *all* situations. For example, if $f(e)$ is Gaussian, then NLS coincides with the exact ML estimator and, therefore, its relative efficiency may be greater.

In the literature on adaptive estimation (e.g. Ling, 2003), the main effort was to establish the absolute (Cramer-Rao) efficiency of the proposed one-step estimators. In general, this can be achieved under the condition of *symmetry* for $f(e)$. In this paper, this constraint is not necessary, and we are mainly interested to compare the (relative) performance of the KML estimator with that of the NLS, in particular in finite samples. The simulations experiments of the next section, will show that the kernel method significantly outperforms the least squares, with the sole exception for t-student innovations. This is due to the fact that, under Gaussianity, the NLS is equivalent to the parametric ML estimator and therefore is more efficient than KML for all n. The inefficiency of KML may arise from the fact that it implicitly involves the estimation the entire noise density; furthermore, in (8) the parameters $\boldsymbol{\beta}$ are jointly estimated with the bandwidth, and in (10) the Hessian matrix is approximated by the mean outer-product of gradient.

3. Confidence Regions based on Empirical Likelihood

Given the asymptotic distribution (12), statistical inference for the regression parameters is a relatively simple task, which requires estimation of the dispersion matrix. Using (11) and (12), a natural estimator is given by

$$\hat{\boldsymbol{V}}_n = \left(\sum_{t=1}^{n} \hat{\boldsymbol{\zeta}}_t \, \hat{\boldsymbol{\zeta}}_t' \right)^{-1} \hat{\tau}_f^{-1}, \qquad \hat{\tau}_f = \frac{1}{n} \sum_{t=1}^{n} \hat{\psi}_\kappa^2(\hat{e}_t)$$

and 95% the confidence intervals become $\hat{\beta}_i \pm 1.96\sqrt{\hat{v}_{ii}}$. However, probability coverage of these intervals is not exact, because the asymptotic result (12) may not hold for finite samples. Consistently with the nonparametric nature of the point estimates developed in Section 2, more reliable confidence regions can be constructed with the empirical likelihood (EL) approach of Owen (1990, 2001).

We recall basic EL principles by following the estimating equations approach (e.g. Qin and Lawless, 1994). Let $\{\boldsymbol{z}_i\}_1^n$ be independent observations from the distribution $F(\boldsymbol{z};\boldsymbol{\theta})$, $\boldsymbol{z} \in \Re^k$, $\boldsymbol{\theta} \in \Re^m$, and suppose that information about $\boldsymbol{\theta}$ is available through m estimating functions $\boldsymbol{g}(\boldsymbol{z},\boldsymbol{\theta}) = [\, g_1(\boldsymbol{z},\boldsymbol{\theta}) \ldots g_m(\boldsymbol{z},\boldsymbol{\theta}) \,]'$, which satisfy the constraint $\mathrm{E}[\, \boldsymbol{g}(\boldsymbol{z},\boldsymbol{\theta}) \,] = \boldsymbol{0}$. In this context, maximization of the EL function $L_n(F) = \prod_i \pi_i$ (where $\pi_i = \mathrm{P}(\mathbf{z} = \boldsymbol{z}_i)$ are multinomial probabilities assigned to the observations), subject to the constraint $\sum_i \pi_i \, \boldsymbol{g}(\boldsymbol{z}_i,\boldsymbol{\theta}) = \boldsymbol{0}$, leads to the solutions $1/\pi_i = n \, [\, 1 + \boldsymbol{\lambda}' \boldsymbol{g}(\boldsymbol{z}_i,\boldsymbol{\theta}) \,]$, where $\boldsymbol{\lambda}$ are Lagrangian multipliers.

In the absence of parametric constraints, the EL function is maximized by the weights $\pi_i = 1/n$; thus, the EL ratio $R_n(F) = \prod_i n \, \pi_i$ takes the profile expression $\log R_n(\boldsymbol{\theta}) = -\sum_i \log [\, 1 + \boldsymbol{\lambda}' \boldsymbol{g}(\boldsymbol{z}_i,\boldsymbol{\theta}) \,]$. Now, the fundamental EL theorem (Owen, 1990 p.91) states that the statistic $-2 \log R_n(\boldsymbol{\theta}) \to \chi^2(m)$ in law. This can be used for inferential purposes; for example, the $(1-\alpha)$ confidence region is given by

$$C_\alpha = \left\{ \boldsymbol{\theta} \, : \, 2 \sum_{i=1}^{n} \log [\, 1 + \boldsymbol{\lambda}' \boldsymbol{g}(\boldsymbol{z}_i,\boldsymbol{\theta}) \,] \leq \chi^2_{1-\alpha}(m) \right\} \tag{13}$$

where the multipliers satisfy the constraint

$$D_\lambda = \left\{ \boldsymbol{\lambda} \, : \, \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{z}_i,\boldsymbol{\theta}) / [\, 1 + \boldsymbol{\lambda}' \boldsymbol{g}(\boldsymbol{z}_i,\boldsymbol{\theta}) \,] = \boldsymbol{0} \right\} \tag{14}$$

These coefficients can also be computed as $\boldsymbol{\lambda}(\boldsymbol{\theta}) = \arg\max \sum_i \log [\, 1 + \boldsymbol{\lambda}' \boldsymbol{g}(\boldsymbol{z}_i,\boldsymbol{\theta}) \,]$, because first order conditions on the latter provide the equation in (15).

These results were obtained under the assumption of independent sequences and could not be applied to autocorrelated data, such as time series. Kitamura (1977) and Monti (1997) have extended the basic EL theorem to stationary dependent data, by working on blocks of observations and periodogram ordinates, which are nearly independent. Both these approaches are computationally demanding and are extraneous to the point estimators developed in the previous section. Instead, we prefer to follow the *dual likelihood* approach of Mikland (1995), which extends the EL theorem to martingale difference processes.

As we have seen in the previous section, typical gradient functions $\{\, (\boldsymbol{\zeta}_t e_t),\, \boldsymbol{\xi}_t \,\}$ involved in the iterative estimators (5) and (10), are martingale difference under the true parameters $\boldsymbol{\beta}$. Moreover, the martingale $\boldsymbol{m}_n(\boldsymbol{\beta}) = n^{-1} \sum_t \boldsymbol{\xi}_t(\boldsymbol{\beta})$ (score of the kernel likelihood function), provides $m = (p + q + d + 1)$ estimating equations and $\mathrm{E}[\, \boldsymbol{\xi}_t(\boldsymbol{\beta}) \,] = \mathbf{0}$. These quantities can be directly used to construct the likelihood ratio statistic, which has a structure similar to (13) and (14):

$$\begin{aligned} -\log R_n(\boldsymbol{\beta}) &= \sum_{t=1}^{n} \log\Big[\, 1 + \boldsymbol{\lambda}' \, \boldsymbol{\xi}_t(\boldsymbol{\beta}) \,\Big] \\ \boldsymbol{\lambda} &: \sum_{t=1}^{n} \boldsymbol{\xi}_t(\boldsymbol{\beta}) \,/ \Big[\, 1 + \boldsymbol{\lambda}' \, \boldsymbol{\xi}_t(\boldsymbol{\beta}) \,\Big] = \mathbf{0} \end{aligned} \tag{15}$$

in fact, maximizing the ratio $R_n = \prod_t n\, \pi_t$ with respect to π_t, subject to the constraint $\sum_t \pi_t \, \boldsymbol{\xi}_t(\boldsymbol{\beta}) = \mathbf{0}$, provides (15). In this framework, it is worth noting that the value which maximizes (15) is just the kernel ML estimator (7).

In the Mykland's view, equation (15) is a dual likelihood, in the sense that, regarding $\boldsymbol{\beta}$ as fixed, the dual parameters $\boldsymbol{\lambda}$ become the unknown coefficients and the resulting function still shares typical features of a log-likelihood. Thus, letting

$$\ell_\beta(\boldsymbol{\lambda}) = -\log R_n(\boldsymbol{\lambda} | \, \boldsymbol{\beta}) = \sum_{t=1}^{n} \log\Big(\, 1 + \boldsymbol{\lambda}' \, \boldsymbol{\xi}_t \,\Big) \tag{16}$$

it follows that $\partial\, \ell_\beta(\boldsymbol{\lambda}) \,/ \partial\, \boldsymbol{\lambda}\,|_{\lambda=0} = \sum_t \boldsymbol{\xi}_t$ is also the score of the dual likelihood, and inference on $\boldsymbol{\beta}$ may proceed through $\boldsymbol{\lambda}$. For example, testing for $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, can be carried out by performing a likelihood ratio test with $\ell_{\beta_0}(\boldsymbol{\lambda})$ on the hypothesis $\boldsymbol{\lambda} = \mathbf{0}$. This procedure has significant advantages in terms of *accuracy* with respect to classical score-type tests (e.g. Mykland, 1995 p.403).

Now, the extension of the EL theorem operated in the Mykland's approach consists of proving that the dual likelihood ratio statistic (16) has the same asymptotic properties as the classical score-type statistics. As before, consider first and second derivatives of $\ell_\beta(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ evaluated at the point $\mathbf{0}$, and construct the quadratic score (Wald) statistic

$$W_n(\boldsymbol{\beta}) = -\dot{\ell}_\beta(\mathbf{0})' \, \ddot{\ell}_\beta(\mathbf{0})^{-1} \dot{\ell}_\beta(\mathbf{0}) = \sum_{t=1}^{n} \boldsymbol{\xi}_t' \left(\sum_{t=1}^{n} \boldsymbol{\xi}_t \boldsymbol{\xi}_t' \right)^{-1} \sum_{t=1}^{n} \boldsymbol{\xi}_t$$

Now, supposing that W_n is tight and that $\boldsymbol{\xi}_t$ is asymptotically negligible with respect to the smallest eigenvalue ρ_n of the matrix $-\ddot{\ell}_\beta(\mathbf{0})$ (namely $\sup_t \|\boldsymbol{\xi}_t\|/\rho_n = o_p(1)$), it can be shown that the dual likelihood ratio statistics satisfies

$$\sup_{\lambda} \ell_\beta(\boldsymbol{\lambda}) = \frac{1}{2} W_n(\boldsymbol{\beta}) + o_p(1)$$

where the supremum is taken in a neighborhood of zero where $L_\beta(\boldsymbol{\lambda})$ is nonnegative (see Mykland, 1995, p.407). Finally, applying the central limit theorem for martingale sequences it can be shown that $W_n(\boldsymbol{\beta}) \to \chi^2(m)$ in law, and the extension of EL theorem follows from the definition in (16).

This result legitimates the use of equations (13)-(14), with $\boldsymbol{g}_i$ replaced by $\boldsymbol{\xi}_t$, in building EL confidence regions for the parameters of the model (1). The dual nature of the approach can be appreciated in the computational aspects:

Step 1. Select a grid of values of $\boldsymbol{\beta}$ in the parameter space $\boldsymbol{B}$ of the model (1) and generate the corresponding series $\boldsymbol{\xi}_t(\boldsymbol{\beta})$ with the formula (9).

Step 2. Solve the problem $\boldsymbol{\lambda} = \arg\max \sum_t \log(1 + \boldsymbol{\lambda}' \boldsymbol{\xi}_t)$, for each series $\{\boldsymbol{\xi}_t\}$.

Step 3. Select the pairs $\{\boldsymbol{\xi}_t^\alpha, \boldsymbol{\lambda}^\alpha\}$ which satisfy $2\sum_t \log(1 + \boldsymbol{\lambda}' \boldsymbol{\xi}_t) \leq \chi^2_{1-\alpha}(m)$.

The confidence region (13) is then given by the set of parameters corresponding to these pairs, namely $C_\alpha = \{\boldsymbol{\beta} : \boldsymbol{\xi}_t(\boldsymbol{\beta}) = \boldsymbol{\xi}_t^\alpha, \boldsymbol{\lambda}(\boldsymbol{\beta}) = \boldsymbol{\lambda}^\alpha\}$. This region is "centred" on the point

$$\hat{\boldsymbol{\beta}}_n = \arg\max \; -\sum_{t=1}^{n} \log\left[1 + \boldsymbol{\lambda}(\boldsymbol{\beta})' \boldsymbol{\xi}_t(\boldsymbol{\beta})\right] \tag{17}$$

which represents the maximum EL estimator and coincides with (7) by first order conditions. Finally, it should be noted that Bartlett corrections of the nominal confidence level of the region C_α can be performed as in Monti (1997).

4. Numerical Studies

4.1. Simulations

To illustrate the methods of the previous sections, we perform simulation experiments and applications to real data. We test point estimators on a first order ARMAX system with chi-square, uniform, t-student, and normal mixture innovations

$$\begin{aligned} y_t &= .5\, y_{t-1} + .5\, x_{t-1} + .5\, e_{t-1} + e_t\,, \qquad x_t = .5\, x_{t-1} + u_t \\ f(e) &= \chi^2(5),\ \mathrm{U}(-5,5),\ t(5),\ \frac{1}{2}\Big[\,\mathrm{N}(-3,2^2) + \mathrm{N}(2,1)\,\Big]\,, \qquad f(u) = \mathrm{N}(0,2^2) \end{aligned} \tag{18}$$

all densities were centred, and $\{e_t, u_t\}$ are mutually independent. We compared the performance of the NLS estimator (3), with those of the KML methods (8) and (7); the latter was implemented with the heuristic design $\hat{\kappa}_n^* = \hat{\sigma}_e / n^{1/5}$, where $\hat{\sigma}_e^2$ is the sample variance of NLS residuals. As is known (e.g. Wand and Jones, 1995 p.60), this solution minimizes the integrated mean squared error of the kernel density : $\int \mathrm{E}[\hat{f}_\kappa(e) - f(e)]^2 \mathrm{d}e$, when both $f(e), K(\cdot)$ are Gaussian.

In the experiment, the sample size and the number of replications were $n = 200$ and $m = 1000$ respectively. Summary statistics, such as root mean squared errors $s_\beta = \Big[m^{-1} \sum_i (\hat{\beta}_i - \beta_0)^2\Big]^{1/2}$ and mean biases $d_\beta = \Big(m^{-1} \sum_i \hat{\beta}_i - \beta_0\Big)$ are reported in Table 1. As a general result, we can note that KML methods significantly outperform the NLS one, especially when the noise density is asymmetric, bimodal or flat (i.e. non-Gaussian). This conclusion is true for both indices s, d, with the sole exception for the index s_β in the third experiment. Specifically, in the presence of t-student innovations, NLS tends to have the smaller MSE, and (8) is preferable to (7). On average, however, the two KML methods perform similarly, and this leads to prefer the solution (7) with the design $\kappa_n^* = \sigma_e / n^{1/5}$, both in view of the computational simplicity and by the normal distribution of its estimates.

The better performance of NLS in the case of bell-shaped densities is a consequence of the fact that, under Gaussianity, it is equivalent to the full ML estimator for any n (whereas it is only true asymptotically for KML). This conclusion cannot, however, be extended to any symmetric $f(e)$, as the results of the simulation with Uniform innovations show. In general, the more the distance from the Normal dis-

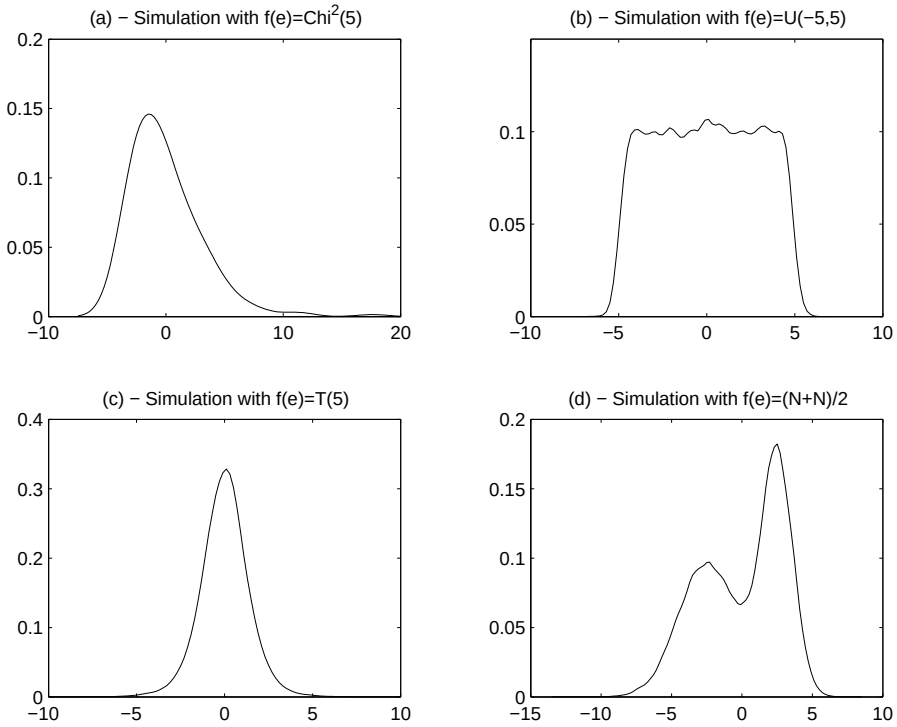
tribution, the more the gain of efficiency of KML over NLS. In view of this situation, a sensible estimation strategy is to perform tests of Gaussianity to least-squares residuals, before applying the kernel method.

Table 1. Results of the simulation experiment applied to the system (18). s_β are root mean squared errors, d_β are mean biases, and $\bar{s}, |\bar{d}|$ are their averages over the 3 parameters. $\bar{\kappa}$ is the mean value of the bandwidths, where in (7) they are estimated as $\hat{\sigma}_e/n^{1/5}$. Finally, (N+N)/2 means mixture of normal densities.

$f(e)$	$\hat{\boldsymbol{\beta}}_n$	s_ϕ	s_δ	s_θ	$\bar{s}$	d_ϕ	d_δ	d_θ	$\lvert\bar{d}\rvert$	$\bar{\kappa}$
$\chi^2(5)$	(3)	.0742	.1077	.0762	.086	-.0154	-.0018	.0096	.009	.
”	(7)	.0600	.0860	.0610	**.069**	-.0093	.0024	.0051	.006	1.1437
”	(8)	.0648	.0959	.0671	.076	-.0078	.0019	.0023	**.004**	.7975
U(-5,5)	(3)	.0721	.1044	.0775	.085	-.0114	-.0033	.0008	.005	.
”	(7)	.0436	.0648	.0469	.052	-.0040	-.0003	-.0004	.002	1.0522
”	(8)	.0436	.0574	.0458	**.049**	.0011	.0009	-.0036	**.002**	.2369
$t(5)$	(3)	.0548	.0469	.0742	**.059**	-.0087	-.0017	.0042	.005	.
”	(7)	.0735	.0624	.1039	.079	-.0042	-.0006	-.0028	.003	.4677
”	(8)	.0592	.0490	.0877	.064	-.0060	-.0007	.0001	**.002**	.5766
(N+N)/2	(3)	.0735	.1030	.0794	.085	-.0112	-.0055	.0002	.006	.
”	(7)	.0412	.0574	.0447	**.048**	-.0024	-.0012	-.0003	**.001**	1.0753
”	(8)	.0510	.0714	.0539	.059	-.0008	.0009	-.0028	.001	.4786

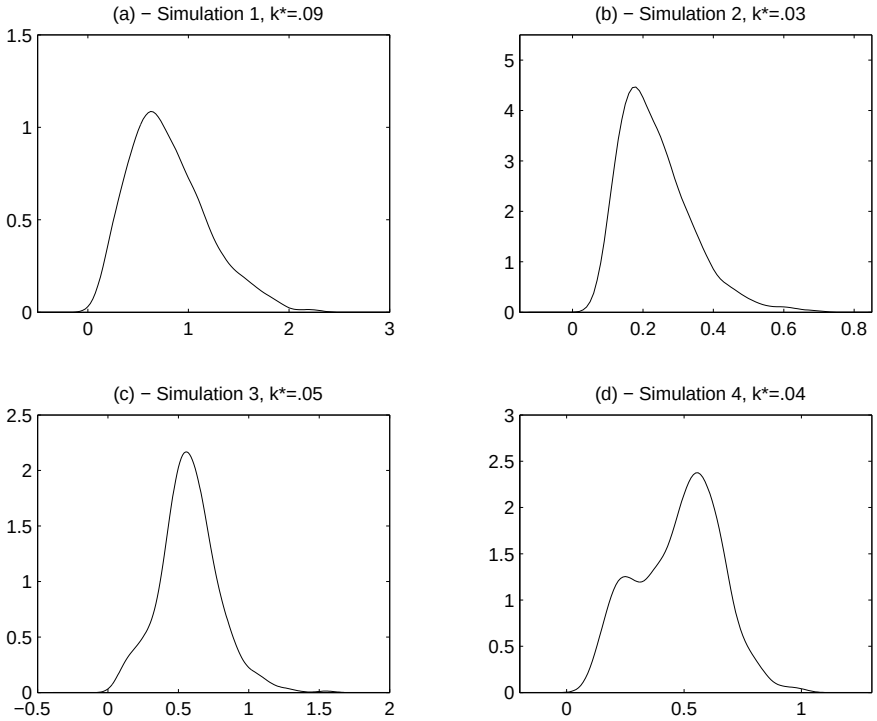
For the sake of completeness, we also estimated the kernel densities of the innovations yielded by the simulations of Table 1. At the ith replication, the residuals $\hat{e}_{it}$ were generated with the coefficients $\hat{\phi}_i$, $\hat{\delta}_i$, $\hat{\theta}_i$, then $\hat{f}_{i\kappa}(e)$ was computed with the bandwidth $\hat{\kappa}_i$. This was carried out only for the method (8), in order to asses the goodness of the direct bandwidth estimation. Mean values $\bar{f}(e) = m^{-1}\sum_{i=1}^{m}\hat{f}_{i\hat{\kappa}}(e)$ over the first $m = 100$ replications are displayed in Figure 1. As we can see, they reproduce well the underlying functions.

Figure 1. Mean values of the kernel densities of the residuals of the simulations of Table 1, obtained with the method (8).



Finally, we investigate the sample distributions of the parameter estimates in order to check the validity of the result (12) and to get insight on the behavior of the bandwidths. In general, the regression coefficients $\hat{\phi}_i$, $\hat{\delta}_i$, $\hat{\theta}_i$ have a normal distribution, whereas those of $\hat{\kappa}_i$ are close to normality only in the case of the heuristic design $\hat{\sigma}_e/n^{1/5}$. Instead, when the bandwidths are estimated with the method (8), they tend to follow either a χ^2 distribution or a density which is related to the underlying $f(e)$. While the first occurrence is theoretically supported by the analysis of Chiu (1990), and by the fact that the role of κ in (8) is similar to a variance, the second situation is unexpected, and largely unexplored. Figure 2 displays these results by means of kernel densities computed on the estimates (8) of Table 1 with the smoothing coefficient $\hat{\kappa}^*_m = \hat{\sigma}_\kappa/m^{1/5}$.

Figure 2. Kernel densities of the bandwidth estimates $\hat{\kappa}_i$ obtained with the method (8) in the simulation experiments of Table 1.



4.2. Application

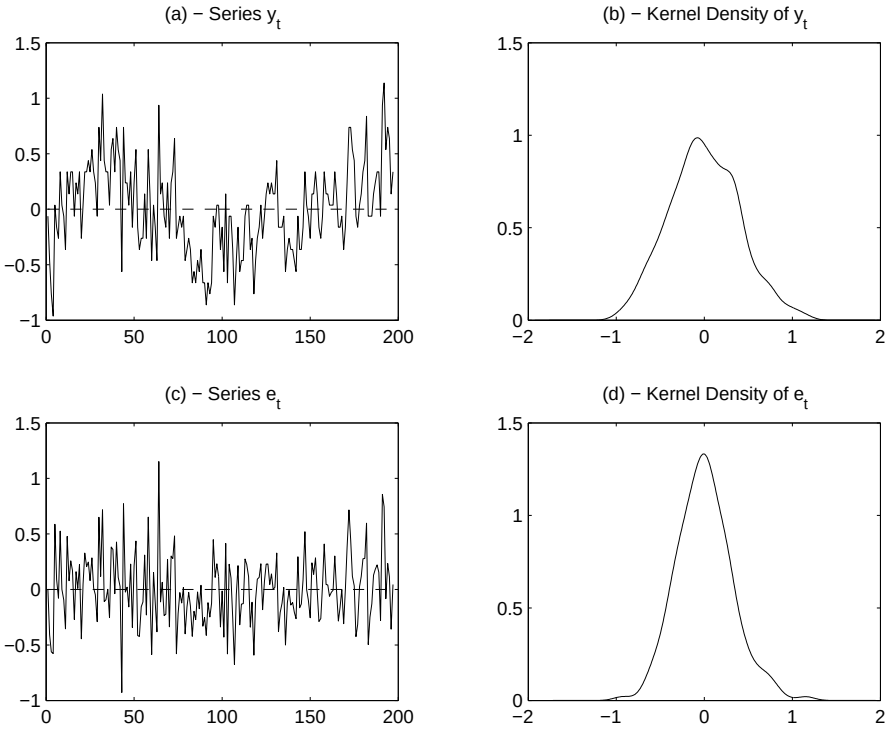
To illustrate the inferential procedure based on kernel empirical likelihood, it is useful to consider an application to a real case study. We consider Series A of Box *et al.* (1994, p.86) which consists of the uncontrolled concentration of a chemical process measured every two hours; total number of observations is n=197. Plot of the centred series $y_t = (Y_t - \bar{Y})$, together with its kernel density, are provided in Figure 3 (a,b); they show some departures from stationarity and gaussianity. We have retained the model structure identified by Box *et al.* (1994, p.186), and we have applied estimators (8) and (3), obtaining

$$\begin{aligned} \text{KML}: \quad y_t &= .905\, y_{t-1} - .565\, e_{t-1} + e_t \qquad \hat{\kappa} = .160 \\ \text{NLS}: \quad y_t &= .863\, y_{t-1} - .486\, e_{t-1} + e_t \qquad \hat{\sigma}_e^2 = .098 \end{aligned} \tag{19}$$

these estimates are very similar, and confirm the results of Box *et al.* (1994, p.196). Figure 3 (c,d) show the residuals of the first model and its kernel density; in this

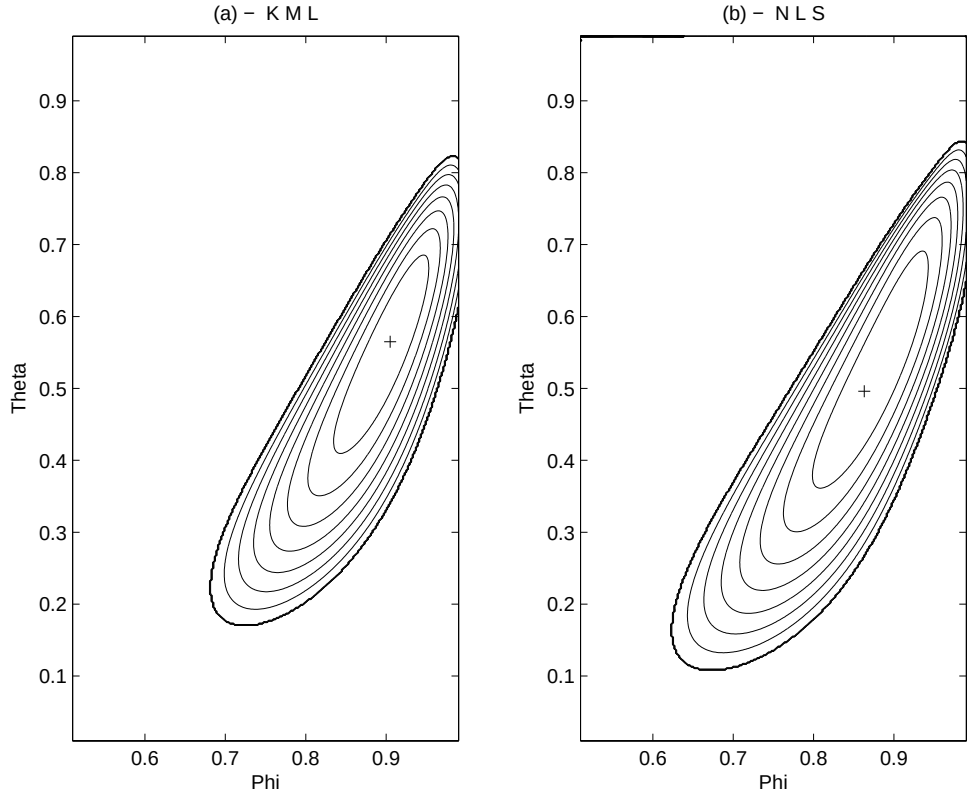
case, the hypotheses of stationarity and gaussianity are more plausible.

Figure 3. Plot of series $\{ y_t, \hat{e}_t \}$ and their kernel densities for the model (19).



Subsequently, we have constructed the EL confidence regions for the pair $(\phi, -\theta)$, at the nominal level 95%, based on the estimates in (19). The two solutions mainly differ from the structure of their scores: $\boldsymbol{\xi}_t = \boldsymbol{\zeta}_t e_t$, $\boldsymbol{\zeta}_t \psi_t$. The score of the noise density ψ_t was computed as in (11), by using the bandwidth estimate in (19), which is close to the heuristic solution $\hat{\sigma}_e / n^{1/5} = 0.21$. The results are presented in Figure 4 (a,b), and display a significant departure from the asymptotic ellipsoidal shape. The region corresponding to the KML scores is smaller than that of NLS, but its size increases as the value of κ, and we used the smaller value in (19). Finally it is interesting noting that point estimates in (19) are nearly at the center of the two regions, confirming the relationship between kernel and empirical likelihoods as described by the equation (17).

Figure 4. EL confidence regions of level 95% for the parameters of (19).



5. Conclusions

In this paper we have presented a complete inferential procedure for time-series models, based on nonparametric likelihoods. Point estimation is based on the maximization of kernel likelihood functions built on the regression residuals. Confidence regions are obtained from the empirical likelihood approach applied to the kernel scores of the point estimates. Asymptotic analysis, obtained from limit theorems of martingale difference sequences, and simulation experiments with non-Gaussian innovations, demonstrate the validity of the methods. Only in the presence of bell-shaped symmetric densities and finite sample sizes the least squares method may be more efficient. Therefore, before applying nonparametric likelihood solutions, it may be advisable to perform tests of Gaussianity on the least squares residuals. Directions for further research are represented by the development of recursive (on-line) versions of the KML estimator (10) and the asymptotic analysis of the direct bandwidth estimates in (8).

References

Azzalini A., Bowman A.W., Hardle W. (1989), On the use of nonparametric regression for model checking. *Biometrika*, **76**, 1-11.

Bickel P.J. (1982), On adaptive estimation. *Annals of Statistics*, **10**, 647-671.

Box G.E.P., Jenkins G.M., Reinsel G.C. (1994), *Time Series Analysis: Forecasting and Control* (Third Edition). Prentice-Hall: Englewood Cliffs.

Bowman A.W., Hall P., Titterington D.M. (1984), Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika*, **71**, 341-351.

Chiu S.-T. (1990), On the asymptotic distribution of bandwidth estimates. *Annals of Statistics*, **18**, 1696-1711.

Drost F.C., Gonzales-Rivera G. (1999), Efficiency comparisons of maximum-likelihood-based estimators in GARCH models. *Jour. of Econometrics*, **93**, 93-111.

Grillenzoni C. (1991), Iterative and recursive estimation of transfer functions.*Journal of Time Series Analysis*, **12**, 105-127.

Grillenzoni C. (2000), Nonparametric Regression for Nonstationary Processes. *Journal of Nonparametric Statistics*, **12**, 265-282.

Kitamura Y. (1997), Empirical likelihood methods with weakly dependent processes.*Annals of Statistics*, **25**, 2084-2102.

Koul H.L., Shick A. (1997), Efficient estimation of nonlinear autoregressive models. *Bernoulli*, **3**, 247-277.

Kreiss J.-P. (1987), On adaptive estimation of stationary ARMA models. *Annals of Statistics*, **15**, 112-133.

Ling S. (2003), Adaptive estimators and tests of ARFIMA-GARCH models. *Journal of American Statistical Association*, **98** , 955-967.

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

Mykland P.A. (1995), Dual likelihood. *Annals of Statistics*, **23**, 396-421.

Owen A.B. (1990), Empirical likelihood ratio confidence regions.*Annals of Statistics*, **18**, 90-120.

Owen A.B. (2001), *Empirical Likelihood*. Chapman and Hall: London.

Qin J., Lawless J. (1994), Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**, 300-325.

Wand M.P., Jones M.C. (1995), *Kernel Smoothing*. Chapman and Hall: London.

White H. (1996), *Estimation, Inference and Specification Analysis*. Cambridge University Press: New York.