Rochester Center for

Economic Research

Semiparametric Efficient Estimation in Time Series

Hodgson, Douglas J.

Working Paper No. 442
August 1997

University of

Rochester

Semiparametric Efficient Estimation in Time Series

Douglas J. Hodgson

Rochester Center for Economic Research
Working Paper No. 442

August 1997

Semiparametric Efficient Estimation in Time Series

Douglas J. Hodgson*
University of Rochester
Department of Economics
Harkness Hall
Rochester, NY 14627
dshn@troi.cc.rochester.edu
Phone: (716) 275-5782
Fax: (716) 256-2309

July 14, 1997

Abstract

We obtain semiparametric efficiency bounds for estimation of a location parameter in a time series model where the innovations are stationary and ergodic conditionally symmetric martingale differences but otherwise possess general dependence and distributions of unknown form. We then describe an iterative estimator that achieves this bound when the conditional density functions of the sample are known. Finally, we develop a "semi-adaptive" estimator that achieves the bound when these densities are unknown by the investigator. This estimator employs nonparametric kernel estimates of the densities. We show that this estimator has robustness properties in the presence of a certain degree of nonstationarity. We extend the method to the estimation of time series regression models and report some Monte Carlo results.

I am grateful to Bill Brown, conversations with whom directly motivated this research. I also acknowledge helpful comments from Oliver Linton, Allan Gregory and participants in workshops at Princeton, Queen's, Rochester, Texas A&M, Rice, Toronto, UBC, and Guelph.

1. INTRODUCTION

A central concern in the estimation of regression models with time series data has for decades been the appropriate modeling of the autocorrelation present in the regression disturbances. Autocorrelation must be accounted for in the calculation of OLS standard errors and must be correctly modeled in order to obtain asymptotic efficiency through the application of generalized least squares (GLS). Heterogeneity in second moments has also received much attention over the years, with the recent burgeoning of the ARCH literature (stemming from Engle (1982) and surveyed by Bollerslev, Chou, and Kroner (1992) and Bera and Higgins (1993)) also bringing dependence in second moments to the fore. In fact, econometricians are paying increasing recognition to the fact that deviations of the disturbances in a time series regression from the assumptions of the classical model can be due to quite general forms of dependence, heterogeneity, and non-Gaussianity.

The presence of all three of these factors is a typical feature of much economic time series data, financial asset price series in particular. The presence of any one of them causes OLS to lose the asymptotic efficiency property that it possesses in the special case of iid Gaussian disturbances and hence motivates the development of new estimation procedures better suited to handle the peculiarities of the data. We have already mentioned GLS, which improves upon OLS in that it accounts for the autocorrelation and heteroskedasticity that may exist in the data. The variety of ARCH and GARCH models that have been developed allow for a rich array of approaches to modeling second moment dependence that may be present in the residuals, and lead to maximum likelihood regression estimators that outperform OLS or GLS when such dependence occurs. The presence of non-Gaussianity in economic data, even after accounting for ARCH effects, has occasioned the employment of non-Gaussian likelihoods such as the Student's t or mixtures of normals (see, for example, Baillie and Bollerslev (1989a) and Lye, Martin, and Teo (1996)), and has led several investigators (including Linton (1993) and Drost and Klaassen (1996)) to explore the application of adaptive estimation, in which the use of nonparametric kernel techniques allows asymptotically efficient estimation even in the absence of knowledge of the shape of the likelihood function.

Adaptive estimation is unique among the strategies listed in the preceding paragraph in that, rather than requiring the specification of a parametric model of the departure of the disturbance process from the canonical model, it treats this departure as being due to some unknown infinite-dimension nuisance parameter. Of course, elements of the disturbance process other than the density of its innovations can and have been treated nonparametrically. In fact, it is possible to think of the dependence, heterogeneity, and non-Gaussianity in the data generating process (DGP) of the regression disturbances as being due to a single infinite-dimensional nuisance parameter, and to explore the consequences of different estimation procedures under such circumstances. These issues have received a great deal of attention in recent years, constituting the basic subject matter of the influential monographs of Bickel, Klaassen, Ritov, and Wellner (1993) and White (1996).

This paper is concerned with the problem of semiparametrically efficiently estimating the parameters of location models in which the errors may follow a stationary and ergodic time series process with dependence and distributional properties that are assumed to be unknown to the investigator. The work of Pagan and Schwert (1990) illustrates the difficulties involved in trying to find an appropriate model of conditional dependence. An immediate precedent to the work reported here is Hodgson (1996), in which the properties of an adaptive estimator of the parameter of a location model, similar to that developed by Stone (1975) as modified by Bickel (1982) and Kreiss (1987), are analyzed when the iid assumption on the error process fails to hold. Hodgson (1996) finds that the adaptive estimator has a desirable robustness property when the error process is weakly dependent and is symmetrically distributed conditional on its past trajectory - viz., that it is consistent and asymptotically normally distributed with an asymptotic variance equal to the inverse of the Fisher information of the unconditional density of the process, when the process is stationary, and equal to the inverse of the Fisher information of the "average asymptotic" unconditional density (in the sense of Bierens (1984) and Pötscher and Prucha (1989)), allowing for a certain degree of heterogeneity in the process. This result shows that, under conditional symmetry, the adaptive estimator is robust in the sense that its asymptotic distribution depends only on the (average asymptotic) unconditional distribution of the process, regardless of the nature of the dependence or heterogeneity that may be generating this unconditional distribution. However, Hodgson (1996) found that when the distributions and dependence are treated as unknown nuisance parameters, this robustness property of the Stone (1975) type estimators does not imply that they are optimal according to the criterion of semiparametric efficiency. The present paper derives estimators that are optimal in this framework, and illustrates their practical applicability through a brief Monte Carlo simulation exercise. Our semiparametric efficiency results assume stationarity in the data. Although we are unable to derive similar optimality results when we relax the stationarity assumption, we show that our estimator has an important robustness property in the presence of certain forms of nonstationarity.

In Section 2 we introduce the location model under consideration and derive the semiparametric efficiency bound. The latter is just the limit as the sample grows to infinity of the expected outer product of the *efficient score* of the parameter of interest. The efficient score is that part of the score of the sample with respect to the parameter of interest that is orthogonal to the score of the sample with respect to the nuisance parameter. For a location model with dependent, conditionally symmetric innovations, the efficient score has a simple and intuitive form. It is the sum over the sample of the scores of the *conditional* densities. The implication is that, in order to formulate semiparametrically efficient estimators of the location parameter, we must nonparametrically estimate the density of the innovation process *conditional* on its past. This is in contrast to the basic Stone-type estimator analyzed by Hodgson (1996), which only utilizes nonparametric estimates of the *unconditional* density of the innovations. The semiparametric efficiency bound is then obtained as the limit of the average over the sample of the first element along the diagonal of the information matrix of the joint density of the current and all past innovations. For example, as shown in Hodgson (1996), if the data are stationary and m^{th}-order Markov, then the semiparametric efficiency bound is just the first element along the diagonal of the joint density of an innovation and its first m lagged values. In Section 3, we derive estimators that achieve the semiparametric efficiency bound. They take the form of one-step Newton-Raphson iterations in which we begin with some $\sqrt{n}$-consistent preliminary estimator and adjust it by a term involving a nonparametric estimator of the efficient score of the model premultiplied by the inverse of the outer product of this score estimator. Section 4 contains an analysis of the robustness properties of our estimator when the data are nonstationary. In Section 5 we extend the analysis to consider the estimation of the parameters of a linear regression model with ARMA errors. In Section 6 we report the results of a Monte Carlo simulation study and Section 7 contains concluding comments.

2. THE MODEL AND EFFICIENCY BOUND

Suppose we observe the scalar time series $\{y_t\}$ for $t = 1, ..., n$, and that the data are a realization of a stochastic process on a complete probability space $(\Omega, \mathcal{F}, P)$

and are generated according to the following model:

$$y_t = \theta + \varepsilon_t, \tag{1}$$

where the innovation sequence $\{\varepsilon_t\}$ is a stationary and ergodic martingale difference sequence that is also $\varphi-$ or $\alpha-$mixing. We assume that ε_t has unconditional density $f(\varepsilon)$ and symmetric conditional density $g_t(\varepsilon\,|\varepsilon^{t-1})$ (i.e. $g_t(\varepsilon\,|\varepsilon^{t-1}) = g_t(-\varepsilon\,|\varepsilon^{t-1})$) for every $t = 1, ..., n$, where $\varepsilon^{t-1} = (\varepsilon_{t-1,...,}\varepsilon_1)'$, and where we define $g_1(\varepsilon\,|\varepsilon^0) \equiv f(\varepsilon)$. We furthermore assume that $g_t(\varepsilon\,|\varepsilon^{t-1})$ is twice differentiable in ε and ε^{t-1} for every $t = 1, ..., n$. We denote the joint density of ε^t by $f^t(\varepsilon^t)$. Denote the true value of θ by θ_0, which belongs to the interior of the compact metric space (Θ, ρ). The probability measure of the sample of size n at the parameter value θ is denoted by $P_{\theta,n}$. Our objective in this paper is to derive semiparametric efficient estimators of the location θ, where the dependence structure of the data and the density functions are treated together as a single unknown infinite-dimensional nuisance parameter.

Our first objective is to obtain an expression for the semiparametric efficiency bound for our estimation problem.[1] To this end, we introduce the nuisance parameter η, which characterizes the (unknown) dependence and distributional properties present in the data generating process. We can then represent the conditional density at t as being a function of this parameter, writing $g_t(\varepsilon\,|\varepsilon^{t-1}) = g_t(\varepsilon\,|\varepsilon^{t-1};\eta)$ for every $t = 1, ..., n$. The log-likelihood of the sample is

$$\log \mathcal{L}_n(\theta;\eta) = \sum_{t=1}^{n} \log g_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right),$$

where $\varepsilon_t(\theta) = y_t - \theta$. The first derivative of $\log \mathcal{L}_n(\theta;\eta)$ with respect to θ is

$$\frac{\partial \log \mathcal{L}_n(\theta;\eta)}{\partial\theta} = -\sum_{t=1}^{n} \frac{\partial g_t(\varepsilon_t(\theta)\,|\varepsilon^{t-1}(\theta);\eta)/\partial\varepsilon_t}{g_t(\varepsilon_t(\theta)\,|\varepsilon^{t-1}(\theta);\eta)} - \sum_{t=1}^{n}\sum_{j=1}^{t-1} \frac{\partial g_t(\varepsilon_t(\theta)\,|\varepsilon^{t-1}(\theta);\eta)/\partial\varepsilon_{t-j}}{g_t(\varepsilon_t(\theta)\,|\varepsilon^{t-1}(\theta);\eta)},$$

which we can rewrite as

$$S_{n\theta}(\theta;\eta) = -\sum_{t=1}^{n} s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right) - \sum_{t=1}^{n}\sum_{j=1}^{t-1} s_{t,j}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right),$$

where $S_{n\theta}(\theta;\eta) = \frac{\partial \log \mathcal{L}_n(\theta;\eta)}{\partial\theta}$, $s_t(\varepsilon_t(\theta)\,|\varepsilon^{t-1}(\theta);\eta) = \frac{\partial g_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)/\partial\varepsilon_t}{g_t(\varepsilon_t(\theta)|\varepsilon^{t-1}(\theta);\eta)}$, and $s_{t,j}(\varepsilon_t(\theta);|\varepsilon^{t-1}(\theta);\eta) = \frac{\partial g_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)/\partial\varepsilon_{t-j}}{g_t(\varepsilon_t(\theta)|\varepsilon^{t-1}(\theta);\eta)}$. It follows from our symmetry

[1] The concept of semiparametric efficiency bounds is discussed in detail by Bickel, Klaassen, Ritov, and Wellner (1993) and Newey (1990).

assumption on $g_t\left(\varepsilon\left|\varepsilon^{t-1}\right.\right)$ that $s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right) = -s_t\left(-\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$ for every $t = 1, ..., n$ and $s_{t,j}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right) = s_{t,j}\left(-\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$ for every $j = 1, ..., t-1$ and $t = 1, .., n$. The first derivative of $\log \mathcal{L}_n(\theta;\eta)$ with respect to η is

$$\frac{\partial \log \mathcal{L}_n(\theta;\eta)}{\partial\eta} = \sum_{t=1}^{n} \frac{\partial g_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)/\partial\eta}{g_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)}$$

or

$$S_{n\eta}(\theta;\eta) = \sum_{t=1}^{n} s_{t,\eta}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right),$$

where $S_{n\eta}(\theta;\eta) = \frac{\partial \log \mathcal{L}_n(\theta;\eta)}{\partial\eta}$, $s_{t,\eta}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right) = \frac{\partial g_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)/\partial\eta}{g_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)}$, and, by symmetry, $s_{t,\eta}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right) = s_{t,\eta}\left(-\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$ for every $t = 1, ..., n$.

In order to determine the semiparametric efficiency bound for this model, we must first compute the efficient score, which is that component of $S_{n\theta}(\theta;\eta)$, the score with respect to the parameter of interest θ, that is orthogonal to the tangent space, i.e. the infinite-dimensional Hilbert space spanned by all functions of the form $\gamma' S_{n\eta}(\theta;\eta)$, where $S_{n\eta}(\theta;\eta)$ is the score with respect to some nuisance parameterization η and γ is any vector of real numbers with dimension equal to that of η. In addition to its symmetry in ε_t, the modified score function $\gamma' s_{t,\eta}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$ is characterized by the mean-zero property of scores. We can use these two salient features of $s_{t,\eta}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$, and hence of $\gamma' s_{t,\eta}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$, to define our tangent space, $\mathcal{T}_n$, as follows:

$$\mathcal{T}_n = \left\{ D(\varepsilon^n) : D(\cdot) = \sum_{t=1}^{n} d_t\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right),\ d_t\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right) = d_t\left(-\varepsilon_t\left|\varepsilon^{t-1}\right.\right)\ \forall t, \right.$$
$$\left. E\left[d_t\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right)\left|\varepsilon^{t-1}\right.\right] = 0\ \forall t \right\}.$$

The efficient score is the residual from the orthogonal projection of $S_{n\theta}(\theta;\eta)$ on $\mathcal{T}_n$. In our model, this orthogonal projection has a particularly convenient form. We can use our symmetry assumptions to show that the first term in $S_{n\theta}(\theta;\eta)$, viz. $-\sum_{t=1}^{n} s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$, is orthogonal to the tangent set, and the second term, $-\sum_{t=1}^{n}\sum_{j=1}^{t-1} s_{t,j}\left(\varepsilon_t(\theta);\left|\varepsilon^{t-1}(\theta);\eta\right.\right)$, actually belongs to the tangent set. Hence, the orthogonal projection of $S_{n\theta}(\theta;\eta)$ on $\mathcal{T}_n$ is just this second term, while the residual, and hence the efficient score, is just the first term. So, using $S_n(\theta)$ to denote the efficient score, we obtain the result that

$$S_n(\theta) = -\sum_{t=1}^{n} s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right),$$

where we henceforth drop η from our notation, as this explicit inclusion of a nuisance parameter was essentially a notational device which served solely to facilitate our derivation of the efficient score $S_n(\theta)$.

Under$\{P_{\theta_0,n}\}$, the semiparametric efficiency bound is then

$$\begin{aligned}\mathcal{B} &= \lim_{n\to\infty}\left(n^{-1}E\left[S_n(\theta_0)^2\right]\right)^{-1} \\ &= \lim_{n\to\infty}\left(n^{-1}\sum_{t=1}^{n}E\left[s_t\left(\varepsilon_t(\theta_0)\left|\varepsilon^{t-1}(\theta_0)\right.\right)^2\right]\right)^{-1} \\ &= \lim_{n\to\infty}\left(E\left[s_n\left(\varepsilon_n(\theta_0)\left|\varepsilon^{n-1}(\theta_0)\right.\right)^2\right]\right)^{-1},\end{aligned}$$

i.e. the inverse of the information of the conditional density of ε given the infinite past of the process. A semiparametric efficient estimator is one whose asymptotic variance equals $\mathcal{B}$. To further analyze $\mathcal{B}$, we recall that the joint density of ε^t is denoted by $f^t(\varepsilon^t)$. We can then write $g_t(\varepsilon_t|\varepsilon^{t-1}) = f^t(\varepsilon^t)/f^{t-1}(\varepsilon^{t-1})$ and $\partial g_t(\varepsilon_t|\varepsilon^{t-1})/\partial\varepsilon_t = (\partial f^t(\varepsilon^t)/\partial\varepsilon_t)/f^{t-1}(\varepsilon^{t-1})$, from which it follows that

$$\begin{aligned}s_t\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right) &= \frac{\partial g_t(\varepsilon_t|\varepsilon^{t-1})/\partial\varepsilon_t}{g_t(\varepsilon_t|\varepsilon^{t-1})} \\ &= \frac{\partial f^t(\varepsilon^t)/\partial\varepsilon_t}{f^t(\varepsilon^t)}.\end{aligned}$$

We then obtain the result that

$$\begin{aligned}E\left[s_t\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right)^2\right] &= \int s_t\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right)^2 f^t\left(\varepsilon^t\right)d\varepsilon^t \\ &= \int\frac{\left(\partial f^t(\varepsilon^t)/\partial\varepsilon_t\right)^2}{f^t(\varepsilon^t)}d\varepsilon^t,\end{aligned}$$

which is just the first element along the diagonal of the information matrix of the joint density $f^t(\varepsilon^t)$. Denote this quantity by $I_{11}(t)$. Furthermore, let us define the quantities $\mathcal{I}_n = n^{-1}\sum_{t=1}^{n} I_{11}(t)$ and $\mathcal{I} = \lim_{n\to\infty}\mathcal{I}_n$. We can then write our semiparametric efficiency bound as

$$\mathcal{B} = \mathcal{I}^{-1},$$

which is the inverse of the leading diagonal element of the information matrix of the joint density of ε and its infinite past. The preceding heuristic derivation of the semiparametric efficiency bound uses a method similar to "the semiparametric approach" of deriving bounds in iid models described in Section 3.4

of Bickel, Klaassen, Ritov, and Wellner (1993). A more rigorous derivation of the bound, similar to the "nonparametric approach" described in section 3.3 of Bickel, Klaassen, Ritov, and Wellner (1993) and using results of Ibragimov and Khas'minski (1991), is provided in Appendix A.

The asymptotic efficiency bound assumes a particularly convenient form for the special case of stationary m^{th}-order Markov processes. In this case, we can rewrite the conditional densities as

$$g_t\left(\varepsilon_t \left|\varepsilon^{t-1}\right.\right) = g\left(\varepsilon_t \left|\varepsilon_{t-1}, ..., \varepsilon_{t-m}\right.\right)$$

and the associated joint pdf of $(\varepsilon_t, ..., \varepsilon_{t-m})$ as $f(\varepsilon_t, ..., \varepsilon_{t-m})$. We then have that $\mathcal{I}_n \to \mathcal{I}$, where

$$\mathcal{I} = \int \frac{\left(\partial f\left(\varepsilon_t, ..., \varepsilon_{t-m}\right)/\partial\varepsilon_t\right)^2}{f\left(\varepsilon_t, ..., \varepsilon_{t-m}\right)} d\left(\varepsilon_t, ..., \varepsilon_{t-m}\right).$$

Our stationarity assumption is more restrictive than desired, but seems to be necessary to derive the semiparametric efficiency bound. If we were to allow for general heterogeneity, so that each observation had a possibly unique unconditional density $f_t(\varepsilon_t)$, then no finite parameterization of the model would exist (in the absence of very strong restrictions on the nature of the heterogeneity). In particular, we would have to write the nuisance parameter characterizing the heterogeneity in the form η_n, which would be time varying and would have dimensionality of the same order of magnitude as the sample size. We know of no results on semiparametric efficient estimation when the nuisance parameters are of this form.

3. ESTIMATION

We now turn our attention to the construction of estimators that will achieve this semiparametric efficiency bound $\mathcal{B}$. We begin by conducting a thought experiment in which the investigator is assumed to know the parametric structure of all the joint density functions $\{f^t(\varepsilon^t)\}_{t=1}^n$ and is interested in using the sample ε^n to construct an estimator which is asymptotically normally distributed with an asymptotic variance of $\mathcal{B}$. (Crowder (1976) analyzes fully efficient maximum likelihood estimation in this case.) This experiment will then provide us with guidance in the construction of semiparametric efficient estimators in the more realistic case where these joint densities are unknown to the investigator.

A semiparametric efficient estimator is one that sets the efficient score equal to zero, i.e., is $\widehat{\theta}_n$ such that

$$n^{-1}S_n\left(\widehat{\theta}_n\right)=-n^{-1}\sum_{t=1}^{n}s_t\left(\varepsilon_t\left(\widehat{\theta}_n\right)\left|\varepsilon^{t-1}\left(\widehat{\theta}_n\right)\right.\right)=0$$

or

$$\overline{S}_n\left(\widehat{\theta}_n\right)=n^{-1}\sum_{t=1}^{n}s_{n,t}\left(\widehat{\theta}_n\right)=0,$$

where $\overline{S}_n(\theta)=n^{-1}S_n(\theta)$ and $s_{n,t}(\theta)=-s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right)$. We first show that such an estimator does indeed achieve the efficiency bound $\mathcal{B}$, then we show how to compute it if the conditional density functions are known to the researcher, and finally we derive a semiparametric estimator that achieves the bound even if these density functions are unknown.

In order to derive the asymptotic distribution of $\widehat{\theta}_n$, we must first establish its consistency. We begin by assuming that $s_{n,t}(\theta)$ is measurable with respect to the filtration $\{\mathcal{F}_{n,t}\}$ for every $\theta\in\Theta$, where $\mathcal{F}_{n,t}\subset\mathcal{F}_{n,t+1}\subset\mathcal{F}$ for every $t=1,...,n$ and $n=1,2,....$ To obtain our consistency results, we must first state some definitions and make some assumptions (for further discussion of Definitions 1-3 and Assumptions 1-2 below, see White (1996)).

DEFINITION 1 *(White (1996, p. 352): The double array* $\{s_{n,t}(\theta)\}$ *is said to be* Lipschitz-L$_1$ a.s. on Θ *if for each* $\theta^\dagger\in\Theta$ *there exist a constant* $\delta^\dagger>0$, *functions* $L^\dagger_{n,t}:\Omega\to R^+$ *measurable-*$\mathcal{F}$ *and functions* $a^\dagger_{n,t}:R^+\to R^+$, $a^\dagger_{n,t}(\delta)\downarrow 0$ *as* $\delta\to 0$, $n,t=1,2,...$ *such that either*

(i) $\overline{a}^\dagger(\delta)\equiv\sup_n\sup_t a^\dagger_{n,t}(\delta)<\infty$ *for all* $0<\delta\leq\delta^\dagger$, $\overline{a}^\dagger(\delta)\downarrow 0$ *as* $\delta\to 0$, *and* $\left\{n^{-1}\sum_{t=1}^{n}E\left(L^\dagger_{n,t}\right)\right\}$ *is* $O(1)$; *or*

(ii) For some $p>1$, $\overline{a}^\dagger(\delta)\equiv\sup_n\left[n^{-1}\sum_{t=1}^{n}a^\dagger_{n,t}(\delta)^p\right]^{1/p}<\infty$ *for all* $0<\delta\leq\delta^\dagger$, $\overline{a}^\dagger(\delta)\downarrow 0$ *as* $\delta\to 0$, *and* $\left\{n^{-1}\sum_{t=1}^{n}\left(E\left(L^\dagger_{n,t}\right)\right)^{p/(p-1)}\right\}$ *is* $O(1)$;

and, for all θ *in* $\overline{\eta}^\dagger\left(\delta^\dagger\right)\equiv\left\{\theta\in\Theta:\rho\left(\theta,\theta^\dagger\right)\leq\delta^\dagger\right\}$, $\left|s_{n,t}(\theta)-s_{n,t}\left(\theta^\dagger\right)\right|\leq L^\dagger_{n,t}a^\dagger_{n,t}\left[\rho\left(\theta,\theta^\dagger\right)\right]$ *a.s.-P* $n,t=1,2,...$

DEFINITION 2 *(White (1996, p. 352)): For given $\theta^{\dagger} \in \Theta$ and $\delta > 0$, define the random variables $\overline{s}^{\dagger}_{n,t}(\delta) \equiv \sup_{\eta^{\dagger}(\delta)} s_{n,t}(\theta)$ and $\underline{s}^{\dagger}_{n,t}(\delta) \equiv \inf_{\eta^{\dagger}(\delta)} s_{n,t}(\theta)$, where $\eta^{\dagger}(\delta) \equiv \left\{\theta \in \Theta : \rho\left(\theta, \theta^{\dagger}\right) < \delta\right\}$. We say that $\left\{\overline{s}^{\dagger}_{n,t}(\delta)\right\}$* obeys the weak law of large numbers locally at *$\theta^{\dagger}$ if there exists $\delta^{\dagger} > 0$ (depending on $\theta^{\dagger}$) such that for all $0 < \delta < \delta^{\dagger}$, $n^{-1}\sum_{t=1}^{n}\left[\overline{s}^{\dagger}_{n,t}(\delta) - E\overline{s}^{\dagger}_{n,t}(\delta)\right] = o_p(1)$ and similarly for $\left\{\underline{s}^{\dagger}_{n,t}(\delta)\right\}$.*

ASSUMPTION 1: *The array $\{s_{n,t} : \Omega \times \Theta \rightarrow R\}$ is Lipschitz-L_1 a.s. on Θ, and $\overline{s}^{\dagger}_{n,t}(\delta)$ and $\underline{s}^{\dagger}_{n,t}(\delta)$ obey the weak law of large numbers locally at $\theta^{\dagger}$ for all $\theta^{\dagger} \in \Theta$.*

We may now state the following weak uniform law of large numbers for our efficient score function, which follows from Theorem A.2.5 of White (1996, p.353):

LEMMA 1: *Under Assumption 1, we have:*

(i) *$\widetilde{S}_n(\cdot) \equiv n^{-1}\sum_{t=1}^{n} Es_{n,t}(\cdot) : \Theta \rightarrow R$ is continuous on Θ uniformly in* n; *and*

(ii) $\overline{S}_n(\theta) - \widetilde{S}_n(\theta) = o_p(1)$ uniformly on Θ.

To ensure that $\widehat{\theta}_n$ is a consistent estimator of θ_0, we must supplement the result in Lemma 1 with an identification assumption on $\widetilde{S}_n(\theta)$, to the effect that θ_0 is its unique zero. We make use of the following identification criterion, similar to one given by White (1996, p. 28).

DEFINITION 3: *Suppose that $\widetilde{S}_n(\theta)$ has a zero on Θ at θ_0 for every* n=1,2,... *Let $\nu_n(\delta)$ be an open circle in R centered at θ_0 with fixed radius $\delta > 0$. For each* n=1,2,... *define the neighbourhood $\eta_n(\delta) = \nu_n(\delta) \cap \Theta$ with compact complement $\eta^c_n(\delta)$ in Θ. The zero θ_0 is said to be* identifiably unique on Θ *if either for all $\delta > 0$ and all* n *$\eta^c_n(\delta)$ is empty, or for all $\delta > 0$*

$$\liminf_{n \rightarrow \infty}\left[\min_{\theta \in \eta^c_n(\delta)}\left|\widetilde{S}_n(\theta)\right|\right] > 0.$$

We may now state the following weak consistency result for $\widehat{\theta}_n$, using a result analogous to Theorem 3.4 of White (1996, p. 28):

THEOREM 1: *Under Assumption 1, and assuming that θ_0 is the identifiably unique zero of $\widetilde{S}_n(\theta)$, we have*

$$\widehat{\theta}_n - \theta_0 = o_p(1).$$

We now proceed to prove the asymptotic normality of $\widehat{\theta}_n$ and to derive an expression for its asymptotic variance. To this end, we introduce the notation $s_t'\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right) = \partial s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right)/\partial\varepsilon_t$, and, for every $j = 1, ..., t-1$, $s_t^j\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right) = \partial s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right)/\partial\varepsilon_{t-j}$. Before we work out the asymptotic distribution, let us consider these derivatives more carefully. We first observe that the derivatives with respect to the lagged ε, i.e. $s_t^j\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right)$, have zero expectation under $P_{\theta_0,n}$, the probability measure of the sample when evaluated at θ_0. To see this, note that

$$\begin{aligned} & E\left[s_t^j\left(\varepsilon_t(\theta_0)\left|\varepsilon^{t-1}(\theta_0)\right.\right)\right] \\ = \; & E\left[E\left[s_t^j\left(\varepsilon_t(\theta_0)\left|\varepsilon^{t-1}(\theta_0)\right.\right)\left|\varepsilon^{t-1}(\theta_0)\right.\right]\right] \end{aligned}$$

and that $E\left[s_t^j\left(\varepsilon_t(\theta_0)\left|\varepsilon^{t-1}(\theta_0)\right.\right)\left|\varepsilon^{t-1}(\theta_0)\right.\right] = 0$. We can rewrite this conditional expectation as

$$\int s_t^j\left(\varepsilon\left|\varepsilon^{t-1}\right.\right) g_t\left(\varepsilon\left|\varepsilon^{t-1}\right.\right),$$

which is equal to zero for every $t = 1, ..., n$ by the facts that $s_t^j\left(\varepsilon\left|\varepsilon^{t-1}\right.\right) = -s_t^j\left(-\varepsilon\left|\varepsilon^{t-1}\right.\right)$ and $g_t\left(\varepsilon\left|\varepsilon^{t-1}\right.\right) = g_t\left(-\varepsilon\left|\varepsilon^{t-1}\right.\right)$. As for the derivative with regard to current ε, first recall that $s_t\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right) = \frac{\partial f^t\left(\varepsilon^t\right)/\partial\varepsilon_t}{f^t(\varepsilon^t)}$, the first element of the partial derivative vector of $f^t\left(\varepsilon^t\right)$. It follows that $s_t'\left(\varepsilon_t\left|\varepsilon^{t-1}\right.\right) = \left(\frac{\partial}{\partial\varepsilon_t}\right)\frac{\partial f^t\left(\varepsilon^t\right)/\partial\varepsilon_t}{f^t(\varepsilon^t)}$, which is the first element on the diagonal of the Hessian of $f^t\left(\varepsilon^t\right)$. But the expectation of the negative of this quantity is equal to the first element on the diagonal of the information matrix of $f^t\left(\varepsilon^t\right)$, which we have denoted above by $I_{11}(t)$.

We can now investigate the asymptotic distribution of the estimator $\widehat{\theta}_n$. As usual, we employ the following mean value expansion of $S_n\left(\widehat{\theta}_n\right)$ about θ_0:

$$S_n\left(\widehat{\theta}_n\right) = S_n(\theta_0) + S_n'\left(\overline{\theta}_n\right)\left(\widehat{\theta}_n - \theta_0\right) = 0,$$

where $\overline{\theta}_n \in \left[\widehat{\theta}_n, \theta_0\right]$ and $S_n'(\theta) = \sum_{t=1}^n \frac{\partial}{\partial\theta} s_t\left(\varepsilon_t(\theta)\left|\varepsilon^{t-1}(\theta)\right.\right)$. We then have

$$n^{1/2}\left(\widehat{\theta}_n - \theta_0\right) = \left[-n^{-1}S_n'\left(\overline{\theta}_n\right)\right]^{-1} n^{-1/2}S_n(\theta_0).$$

To work out the asymptotic distribution, we consider the limiting behaviour of the two components of the RHS separately, beginning with $n^{-1/2}S_n(\theta_0)$. We can rewrite this quantity as follows:

$$n^{-1/2}S_n(\theta_0) = n^{-1/2}\sum_{t=1}^n s_{n,t}(\theta_0). \tag{2}$$

Since $\{s_{n,t}(\theta_0), \mathcal{F}_{n,t}\}$ is a martingale difference sequence, we can employ Theorem A.3.4 of White (1996, p. 357) to apply a central limit theorem to (2). In order to do so, we make the following assumption:

ASSUMPTION 2: *The sequence* $\{s_{n,t}(\theta_0), \mathcal{F}_{n,t}\}$ *satisfies the following conditions:*

(i) $E\left|s_{n,t}(\theta_0)\right|^{2+\delta} < \Delta$ *for some* $\delta > 0$, $\Delta < \infty$, t=1,...,n, n=1,2,...;
(ii) $\mathcal{I}_n > \delta' > 0$ *for almost all* n; *and*
(iii) $n^{-1}\sum_{t=1}^n s_{n,t}(\theta_0)^2 - \mathcal{I}_n = o_p(1)$.

LEMMA 2: *Under Assumption 2, we have*

$$n^{-1/2}\mathcal{I}_n^{-1/2}S_n(\theta_0) \xrightarrow{d} N(0,1).$$

We complete our analysis of the asymptotic distribution of $\widehat{\theta}_n$ by considering the limiting behaviour of the conditional Hessian

$$S_n'\left(\overline{\theta}_n\right) = n^{-1}\sum_{t=1}^n \frac{\partial}{\partial\theta} s_t\left(\varepsilon_t\left(\overline{\theta}_n\right)\left|\varepsilon^{t-1}\left(\overline{\theta}_n\right)\right.\right)$$

$$= n^{-1}\sum_{t=1}^n s_t'\left(\varepsilon_t\left(\overline{\theta}_n\right)\left|\varepsilon^{t-1}\left(\overline{\theta}_n\right)\right.\right) + n^{-1}\sum_{j=1}^m\sum_{t=1}^n s_t^j\left(\varepsilon_t\left(\overline{\theta}_n\right)\left|\varepsilon^{t-1}\left(\overline{\theta}_n\right)\right.\right).$$

Using the facts that $\overline{\theta}_n - \theta_0 = o_p(1)$, $n^{-1}\sum_{t=1}^{n} E\left[s_t'\left(\varepsilon_t\left(\theta_0\right)|\varepsilon^{t-1}\left(\theta_0\right)\right)\right] = -\mathcal{I}_n$, and, for every $j = 1, ..., t-1$, $E\left[s_t^j\left(\varepsilon_t\left(\theta_0\right)|\varepsilon^{t-1}\left(\theta_0\right)\right)\right] = 0$, it follows from our stationarity and ergodicity assumption that $-S_n'\left(\overline{\theta}_n\right)^{-1} - \mathcal{B} = o_p(1)$, which, combined with Lemma 2, yields the following convergence result:

THEOREM 2: *Under our assumptions, we have*

$$\mathcal{B}^{-1/2}n^{1/2}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0,1\right)$$

so that $\widehat{\theta}_n$ achieves the semiparametric efficiency bound.

We now describe an iterative estimator, constructed under the assumption that the innovation densities are known, that achieves the semiparametric efficiency bound $\mathcal{B}$. We assume the existence of some discretized $n^{1/2}$-consistent preliminary estimator θ_n^* (the discretized sample mean, for example), and construct the following iterative estimator:

$$\theta_n^+ = \theta_n^* + \left[n^{-1}\sum_{t=1}^{n} s_t\left(\varepsilon_t\left(\theta_n^*\right)|\varepsilon^{t-1}\left(\theta_n^*\right)\right)^2\right]^{-1}\left[n^{-1}\sum_{t=1}^{n} s_t\left(\varepsilon_t\left(\theta_n^*\right)|\varepsilon^{t-1}\left(\theta_n^*\right)\right)\right]. \quad (3)$$

It is easily shown, under our differentiability assumptions and using a mean-value expansion of $n^{-1}\sum_{t=1}^{n} s_t\left(\varepsilon_t\left(\theta_n^*\right)|\varepsilon^{t-1}\left(\theta_n^*\right)\right)$ about θ_0, that θ_n^+ constructed in this manner achieves the semiparametric efficiency bound. In particular, we have the following result, which is proved in Appendix B:

THEOREM 3: *Under our assumptions, the estimator θ_n^+ given by (3) achieves the semiparametric efficiency bound $\mathcal{B}$, so that*

$$\mathcal{B}^{-1/2}n^{1/2}\left(\theta_n^+ - \theta_0\right) \xrightarrow{d} N\left(0,1\right).$$

In practice, of course, the conditional densities g_t and their scores s_t are unknown to the investigator. To overcome this difficulty, we now investigate the possibility of using nonparametric kernel methods to estimate this density and its score, and so formulate an estimator that achieves the semiparametric efficiency bound in absence of knowledge of the functional form of g_t and s_t. The basic structure of our semiparametric efficient estimator is the same as that of $\widehat{\theta}_n$ as given in (3), except that in place of the unknown scores $\{s_t(\varepsilon_t(\theta_n^*)|\varepsilon^{t-1}(\theta_n^*))\}$ we employ nonparametric estimators $\{\widehat{s}_t(\varepsilon_t(\theta_n^*)|\varepsilon^{t-1}(\theta_n^*))\}$. To derive such an estimator, recall that

$$s_t\left(\varepsilon_t(\theta_n^*)\middle|\varepsilon^{t-1}(\theta_n^*)\right) = \frac{\partial f^t(\varepsilon^t(\theta_n^*))/\partial\varepsilon_t}{f^t(\varepsilon^t(\theta_n^*))},$$

so our problem reduces to that of obtaining multivariate kernel estimators of the scaled sum of the joint t+1-dimensional densities $f^t(\varepsilon^t(\theta_n^*))$ and of the first elements of their partial derivative vectors, viz. $\partial f^t(\varepsilon^t(\theta_n^*))/\partial\varepsilon_t$. In practice, we employ an $m+1$-dimensional Gaussian kernel estimator, where the kernel is given by

$$\pi_{a_n}(x) = \left(a_n\sqrt{2\pi}\right)^{-(m+1)} \exp\left(\frac{-x'x}{2a_n^2}\right),$$

where $\{a_n\}$ is a bandwidth sequence that converges to zero as $n \to \infty$. Our density estimator is

$$\widehat{f}_t(x,\theta) = (n-1)^{-1}\sum_{\substack{j=1\\ j\neq t}}^{n} \pi_{a_n}(x - \widetilde{\varepsilon}_j(\theta)),$$

where $\widetilde{\varepsilon}_j(\theta) = (\varepsilon_j(\theta), \varepsilon_{j-1}(\theta), ..., \varepsilon_{j-m}(\theta))'$. The use of such an estimator can, in principle, be justified either by the assumption that the data follow an m^{th}-order Markov process, so that $g_t(\varepsilon_t|\varepsilon^{t-1}) = g_t(\varepsilon_t|\varepsilon_{t-1}, ..., \varepsilon_{t-m})$, or by thinking of m as a truncation parameter that increases to infinity with the sample size. In the context of nonparametric estimation of a conditional variance function, Pagan and Hong (1991) consider a similar situation, conjecturing that the latter approach may be possible but not being aware of any results in the nonparametric literature regarding the estimation of infinite-dimensional functions. Due to a similar lack of awareness, we proceed under the Markov assumption here, leaving the more general situation for further investigation.

Define the first element of the vector of partial first derivatives (with respect to x) of $\widehat{f}_t(x,\theta)$ by $\widehat{f}_t'(x,\theta)$, and further define

$$q_t(x,\theta)=\begin{cases}\frac{\widehat{f}_t'(x,\theta)}{\widehat{f}_t(x,\theta)} & if \begin{cases}\widehat{f}_t(x,\theta)\geq d_n\\ |x|\leq e_n\\ \left|\widehat{f}_t'(x,\theta)\right|\leq c_n\widehat{f}_t(x,\theta)\end{cases}\\ 0 & otherwise,\end{cases}$$

where $c_n\longrightarrow\infty$, $e_n\longrightarrow\infty$, $d_n\longrightarrow 0$, $a_nc_n\longrightarrow 0$, $e_na_n^{-(m+3)}=o(n)$, and our score estimator is

$$\widehat{s}_t(x,\theta)=\frac{1}{2}\left(q_t(x,\theta)-q_t(-x,\theta)\right).$$

Our information estimator is

$$\widehat{I}(\theta)=n^{-1}\sum_{t=1}^{n}\widehat{s}_t^2\left(\widetilde{\varepsilon}_t(\theta),\theta\right). \tag{4}$$

We therefore have the semiparametric efficient estimator

$$\widetilde{\theta}_n=\theta_n^*+n^{-1/2}\widehat{I}\left(\theta_n^*\right)^{-1}\left[n^{-1/2}\sum_{t=1}^{n}\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_n^*\right),\theta_n^*\right)\right]. \tag{5}$$

The following theorem is proved in Appendix B:

THEOREM 4: *Under our assumptions, and further assuming that the innovation process $\{\varepsilon_t\}$ follows an $m^{th}-$order Markov process, the estimator $\widetilde{\theta}_n$ given by (5) is semiparametric efficient, so that*

$$\mathcal{B}^{-1/2}n^{1/2}\left(\widetilde{\theta}_n-\theta_0\right)\xrightarrow{d}N(0,1).$$

Equivalently, $\widetilde{\theta}_n$ is asymptotically equivalent to θ_n^+, as given in (3), so that the following convergence result holds:

$$n^{1/2}\left(\widetilde{\theta}_n-\theta_n^+\right)=o_p(1). \tag{6}$$

Theorem 4 is undesirably restrictive due to the assumption that the innovations follow an m^{th}-order Markov process. As mentioned earlier, it would be desirable to drop this assumption and to show that the estimator as constructed above is asymptotically semiparametric efficient in the more general case by using an argument under which the truncation parameter m goes to infinity with sample size, although we do not know if this is possible. However, in the following section we discuss an important and desirable robustness property which $\widetilde{\theta}_n$ possesses even when the data are not Markov.

4. ROBUSTNESS TO HETEROGENEITY AND NON-MARKOV DATA

We have so far assumed that the innovation process $\{\varepsilon_t\}$ is stationary and ergodic, and have indicated in Section 2 that our derivation of semiparametric efficiency bounds relies heavily on this assumption. Our derivation of a semiparametric efficient estimator in Section 3 relies on the further assumption that the data are m^{th}-order Markov. However, these assumptions may well both fail in many empirical applications, so that it is desirable to investigate the robustness properties of our estimator in the presence of such failure. In Hodgson (1996), it is shown that a Stone (1975)-type estimator, which only nonparametrically estimates the unconditional density of the innovation process, has important robustness properties in the presence of certain types of heterogeneity and dependence. In the present section, we extend this result to the case of the semiparametric efficient estimator, which estimates the conditional density of the innovation process.

Allowing for the possible presence of heterogeneity in the data generating process, we now denote the density function of the $m+1$-dimensional vector $\widetilde{\varepsilon}_t = (\varepsilon_t, ..., \varepsilon_{t-m})'$ by f_t, with associated cdf of F_t. Note that neither stationarity nor Markovicity are being assumed here. The cdf F_t is the *marginal* distribution of the vector $\widetilde{\varepsilon}_t$; the existence of such a marginal cdf clearly does not imply that ε_t is independent of ε_{t-j} when $j > m$. The t subscript indicates that the marginal distribution can be changing over time, allowing the possibility of nonstationarity in the process. However, following Bierens (1984) and Pötscher and Prucha (1989), we shall restrict the degree of heterogeneity in the process by assuming a type of "average asymptotic stationarity", in the sense that $n^{-1}\sum_{t=1}^{n} F_t \Rightarrow F$, where $\Rightarrow$ denotes weak convergence of probability measures (cf. Billingsley (1968)). Furthermore, we assume that the average asymptotic unconditional distribution F has a pdf $f = F'$ that is symmetric about zero, twice continuously differentiable, and has finite, positive definite information. We assume that the data are $\varphi-$ or $\alpha-$mixing, but need not be Markov.

We first analyze the behaviour of the "partial" semiparametric efficient estimator $\widehat{\theta}_n$ that we would compute if we knew the unconditional density function f and proceeded as if the data were stationary and m^{th}-order Markov. We would then select $\widehat{\theta}_n$ so as to set the "partial" efficient score equal to zero:

$$n^{-1} \sum_{t=1}^{n} s\left(\widetilde{\varepsilon}_t\left(\widehat{\theta}_n\right)\right) = 0, \tag{7}$$

where $s\left(\widetilde{\varepsilon}_t\left(\theta\right)\right) = \frac{\partial f\left(\widetilde{\varepsilon}_t(\theta)\right)/\partial \varepsilon_t}{f\left(\widetilde{\varepsilon}_t(\theta)\right)}$ is the first element of the score vector of the joint density $f\left(\widetilde{\varepsilon}_t\left(\theta\right)\right)$. The following Lemma can then be established (the proof is in Appendix B):

LEMMA 3: *Under the assumptions of this section, the estimator $\widehat{\theta}_n$ as defined in (7) has the asymptotic distribution*

$$n^{1/2}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}^{-1}\right),$$

where now $\mathcal{I} = \int \frac{\left(\partial f\left(\widetilde{\varepsilon}_t\right)/\partial \varepsilon_t\right)^2}{f\left(\widetilde{\varepsilon}_t\right)} d\widetilde{\varepsilon}_t$. Furthermore, the same asymptotic distribution is obtained by the following one-step iterative estimator:

$$\theta_n^{+} = \theta_n^{*} + \left[n^{-1} \sum_{t=1}^{n} s\left(\widetilde{\varepsilon}_t\left(\theta_n^{*}\right)\right)^2\right]^{-1} \left[n^{-1} \sum_{t=1}^{n} s\left(\widetilde{\varepsilon}_t\left(\theta_n^{*}\right)\right)\right]. \tag{8}$$

We can then derive the following robustness result for our semiparametric estimator $\widetilde{\theta}_n$, a proof of which can also be found in Appendix B:

THEOREM 5: *Under the assumptions of this section, and assuming that $c_n = o\left(n^{\delta/2}\right)$ (recall the definition of $q_t\left(x, \theta\right)$), where $0 < \delta < \infty$ is such that $n^{-1}\sum_{t=1}^{n} F_t(z) - F(z) = O(n^{\delta})$ for every z, the estimator $\widetilde{\theta}_n$, the construction of which is given in (5), has the following asymptotic distribution:*

$$n^{1/2}\left(\widetilde{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}^{-1}\right). \tag{9}$$

Equivalently, we have

$$n^{1/2}\left(\widetilde{\theta}_n - \theta_n^{+}\right) = o_p(1),$$

where the construction of θ_n^+ is given by (8). The following consistency result also holds:

$$\widehat{I}\left(\theta_n^*\right) - \mathcal{I} = o_p(1), \tag{10}$$

where the construction of $\widehat{I}\left(\theta_n^*\right)$ is given by (4).

Note that the condition on c_n in Theorem 5 is redundant in the stationary case since $n^{-1}\sum_{t=1}^n F_t = F$ for every n. In the case where $\{F_t\}$ consists of a periodically repeating cycle k periods long (such as periodic heteroskedasticity which might occur, for example, in daily stock market data if certian days of the week or month have higher variance than others), we have $\delta = 1$.

We conclude from (9) that the estimator $\widetilde{\theta}_n$, although being semiparametric efficient only under stationarity and Markov assumptions on the data generating process, still has the important robustness property that in the presence of heterogeneity and more general dependence, it is asymptotically normal with variance equal to the inverse of the first element on the diagonal of the information matrix of the average asymptotic unconditional joint density $f\left(\widetilde{\varepsilon}\right)$. We know from (10) that our estimator of the asymptotic variance, and hence of the standard errors, is robust to the presence of heterogeneity and non-Markovicity. What this means is that, as far as usefulness for conducting inference is concerned, it doesn't matter whether or not the stationary, Markov assumptions of the preceding section hold - the asymptotic distribution of $\widetilde{\theta}_n$ will be unchanged and the standard errors will still be consistently estimated. This result generalizes that of Hodgson (1996), and has the desirable implication that inference conducted using the estimator $\widetilde{\theta}_n$ and its estimated asymptotic standard errors is robust to heterogeneity and general dependence in the data. The notion of heterogeneity we have used is of course somewhat restrictive, since it assumes a certain asymptotic form of stationarity. It would be nice to allow for more general forms of nonstationarity, but we have yet to conduct investigations upon these lines.

5. LINEAR REGRESSION WITH ARMA ERRORS

We have so far couched our analysis entirely in terms of the simple location parameter model given by (1). We have so restricted ourselves primarily to facilitate the exposition of the main ideas involved in the derivation of semiparametric efficient estimators in the presence of dependence of unknown form, but these ideas are applicable to a much broader range of interesting econometric models, including linear and non-linear regressions, ARMA models, cointegrating regressions, error correction models, etc. In this section, we briefly sketch the theory for the linear regression model whose errors follow a stationary and invertible ARMA process with symmetric martingale difference sequence innovations and show how to semiparametrically efficiently estimate the regression parameters.

We consider the following model:

$$\begin{aligned} y_t &= \alpha + x_t'\beta + u_t, \\ u_t &= \sum_{j=1}^{p} a_j u_{t-j} + \varepsilon_t + \sum_{j=1}^{q} b_j \varepsilon_{t-j}, \end{aligned}$$

where α is the intercept, x_t and β are k-vectors, and we observe the sequence $\{y_t, x_t\}$ for $t = 1, ..., n$. We assume that the ARMA parameters satisfy the usual conditions of stationarity, invertibility, and identifability, that $\{x_t\}_{t=1}^{n}$ is predetermined for β and that $E\left[x_t x_t'\right] = M_t$ is finite and positive-definite for every t and $n^{-1}\sum_{t=1}^{n}\left(x_t x_t' - M_t\right) = o_p(1)$. We furthermore define the σ-field $\Omega_{t-1} = \sigma\left(x_t, x_{t-1}, ...; \varepsilon_{t-1}, ...\right)$ and maintain our assumption that $\{\varepsilon_t\}$ is a stationary m^{th}-order Markov process with symmetric conditional density.[2] Since ε_t is independent of $\{x_j\}_{j=1}^{t}$, we can write the conditional density of ε_t as $g\left(\varepsilon_t \left|\Omega_{t-1}\right.\right) = g\left(\varepsilon_t \left|\varepsilon_{t-1}, ..., \varepsilon_{t-m}\right.\right)$ with the symmetry property that $g\left(\varepsilon_t \left|\varepsilon_{t-1}, ..., \varepsilon_{t-m}\right.\right) = g\left(-\varepsilon_t \left|\varepsilon_{t-1}, ..., \varepsilon_{t-m}\right.\right)$.

We can then write the log-likelihood of the sequence $\{y_t\}$, conditional on the regressors $\{x_t\}$, and assuming the existence of initial conditions, as follows:

$$\mathcal{L}\left(\theta, \eta\right) = \sum_{t=1}^{n} \log g\left(\varepsilon_t\left(\theta\right) \left|\varepsilon_{t-1}\left(\theta\right), ..., \varepsilon_{t-m}\left(\theta\right); \eta\right.\right),$$

where $\theta = \left(\alpha, \beta, a_1, ..., a_p; b_1, ..., b_q\right)'$ and we once again use η to denote a parameterization which includes the true (unknown) model of the conditional density function g. Following Kreiss (1987), we have

$$\varepsilon_t\left(\theta\right) = \sum_{k=1}^{t} \gamma_{k-1}\left(\theta\right)\left\{u_{t+1-k}\left(\theta\right) - \sum_{j=1}^{p} a_j u_{t+1-k-j}\left(\theta\right)\right\} + \sum_{s=0}^{q-1} \varepsilon_{-s}\left(\sum_{k=0}^{s} \gamma_{j+s-k}\left(\theta\right) b_k\right),$$

where the constants $\{\gamma_k\left(\theta\right)\}_{k=0}^{\infty}$ are such that $\sum_{k=0}^{\infty} \gamma_k\left(\theta\right) z^k = \left(1 + b_1 z + \cdots + b_q z^q\right)^{-1}$ and $\gamma_s\left(\theta\right) + b_1\gamma_{s-1}\left(\theta\right) + \cdots + b_q\gamma_{s-q}\left(\theta\right) = 0\ \forall s \geq 1$, with $\gamma_s\left(\theta\right) = 0\ \forall s \leq 0$ and

[2] An adaptive estimator of this model has been derived by Steigerwald (1992) under an iid assumption on the sequence $\{\varepsilon_t\}$.

$\gamma_0(\theta) = 1$. The following notation will facilitate our analysis of the scores for the model:

$$Z_{t-1}(\theta) = \sum_{k=0}^{t-1} \gamma_k(\theta) \left(u_{t-1-k}(\theta), ..., u_{t-p-k}(\theta); \varepsilon_{t-1-k}(\theta), ..., \varepsilon_{t-q-k}(\theta)\right)',$$

$$\overline{\gamma}_t(\theta) = \left(\sum_{j=0}^{t-1} \gamma_j(\theta)\right)\left(1 - \sum_{j=1}^{p} a_j\right),$$

$$\Gamma_{t-1}(\theta) = \sum_{j=0}^{t-1} \gamma_j(\theta)\left(x_{t-j} - \sum_{k=1}^{p} a_j x_{t-j-k}\right),$$

and

$$H_{t-1}(\theta) = \left(\overline{\gamma}_t(\theta), \Gamma'_{t-1}(\theta), Z'_{t-1}(\theta)\right)'.$$

We can then write the score with respect to the parameter θ as follows:

$$\begin{aligned} \frac{\partial \log \mathcal{L}_n(\theta,\eta)}{\partial\theta} &= -\sum_{t=1}^{n} H_{t-1} \frac{\partial g\left(\varepsilon_t(\theta) \,|\varepsilon_{t-1}(\theta), ..., \varepsilon_{t-m}(\theta); \theta, \eta\right)/\partial\varepsilon_t}{g\left(\varepsilon_t(\theta) \,|\varepsilon_{t-1}(\theta), ..., \varepsilon_{t-m}(\theta); \theta, \eta\right)} \\ & -\sum_{t=1}^{n}\sum_{j=1}^{t-1} H_{t-1-j} \frac{\partial g\left(\varepsilon_t(\theta) \,|\varepsilon_{t-1}(\theta), ..., \varepsilon_{t-m}(\theta); \theta, \eta\right)/\partial\varepsilon_{t-j}}{g\left(\varepsilon_t(\theta) \,|\varepsilon_{t-1}(\theta), ..., \varepsilon_{t-m}(\theta); \theta, \eta\right)}, \end{aligned}$$

and the score with respect to the nuisance parameter η as

$$\frac{\partial \log \mathcal{L}_n(\theta,\eta)}{\partial\eta} = \sum_{t=1}^{n} \frac{\partial g\left(\varepsilon_t(\theta) \,|\varepsilon_{t-1}(\theta), ..., \varepsilon_{t-m}(\theta); \theta, \eta\right)/\partial\eta}{g\left(\varepsilon_t(\theta) \,|\varepsilon_{t-1}(\theta), ..., \varepsilon_{t-m}(\theta); \theta, \eta\right)}.$$

We can use our symmetry assumptions and arguments similar to those in Section 2 to show that the efficient score for the regression model is

$$S_n(\theta) = -\sum_{t=1}^{n} H_{t-1}(\theta)\, s\left(\widetilde{\varepsilon}_t, \theta\right),$$

where, as earlier, we have $s\left(\widetilde{\varepsilon}_t, \theta\right) = \frac{\partial g(\varepsilon_t(\theta)|\varepsilon_{t-1}(\theta),...,\varepsilon_{t-m}(\theta);\theta)/\partial\varepsilon_t}{g(\varepsilon_t(\theta)|\varepsilon_{t-1}(\theta),...,\varepsilon_{t-m}(\theta);\theta)} = \frac{\partial f\left(\widetilde{\varepsilon}_t(\theta)\right)/\partial\varepsilon_t}{f\left(\widetilde{\varepsilon}_t(\theta)\right)}$. The semiparametric efficency bound is now given by

$$\begin{aligned} & \left[\lim_{n\to\infty} E\left[n^{-1}\sum_{t=1}^{n} H_{t-1} s\left(\widetilde{\varepsilon}_t, \theta\right)\right]^2\right]^{-1} \\ = & \left[\lim_{n\to\infty} E\left[H_{n-1}(\theta_0)\, H_{n-1}(\theta_0)'\, s\left(\widetilde{\varepsilon}_t, \theta_0\right)^2\right]\right]^{-1}. \end{aligned}$$

We can use argumentation similar to that given above to show that the following semiparametric iterative estimator achieves this bound:

$$\widetilde{\theta}_n = \theta_n^* + \left[n^{-1}\sum_{t=1}^{n} H_{t-1}\left(\theta_n^*\right) H'_{t-1}\left(\theta_n^*\right) \widehat{s}\left(\widetilde{\varepsilon}_t\left(\theta_n^*\right)\right)^2\right]^{-1} \left[n^{-1}\sum_{t=1}^{n} H_{t-1}\widehat{s}\left(\widetilde{\varepsilon}_t\left(\theta_n^*\right)\right)\right].$$

The preliminary estimator θ_n^* is assumed to be discretized and $n^{1/2}$-consistent - the discretized Gaussian pseudo-MLE is an obvious and convenient choice.

6. MONTE CARLO RESULTS

The simulations reported in this section compare the finite-sample mean-squared error (MSE) performances of the sample mean, the Stone (1975)-type "adaptive" estimator, and our "semi-adaptive" semiparametric efficient estimator, for the basic location parameter model given in (1). We consider a variety of models of the second-moment dependence in the data, and carry out MSE comparisons for sample sizes of 100, 250, 500, 750, and 1000. In all cases we conduct 1000 iterations and report the MSE figures for each of the three estimators. To implement the two semiparametric estimators, we use the sample mean as the preliminary estimator, employ the Silverman (1986) rule-of-thumb bandwidth, and set the trimming parameter as in Hodgson (1995). In the case of the semiparametric efficient estimator, we always set $m = 1$, so that we are only accounting for dependence in the data at one lag (although some of the DGP's we consider are not Markov processes). In this case, we are nonparametrically estimating a bivariate density (i.e. that of ε_t and ε_{t-1}), and so use the appropriate rule-of-thumb bandwidth. The results presented here are of fairly limited scope. We intend to carry out more extensive simulation work both of the estimator developed in this paper and of other related estimators in future research (Hodgson (1997)). We now describe our particular data-generating processes.

6.1. The ARCH Model

The model we consider here is the basic Gaussian ARCH(1) process introduced by Engle (1982), in which the conditional variance is a linear function of the lagged square of the process. Hence, a large absolute value of the realization of the process in one period increases the probability of large absolute values in subsequent periods, leading to the long-recognized phenomenon in economic and financial time series of volatility clustering. Formally, we have an ARCH(1) model if the innovations $\{\varepsilon_t\}$ in (1) are generated as follows:

$$\begin{aligned} \varepsilon_t &= u_t\sqrt{h_t} \\ u_t &\sim iidN(0,1) \\ h_t &= \alpha_0 + \alpha_1\varepsilon_{t-1}^2. \end{aligned}$$

We assume that $|\alpha_1| < 1$ and $\alpha_0 > 0$. In this formulation, h_t denotes the conditional variance of the process. The unconditional variance is $\sigma_\varepsilon^2 = \alpha_0/(1-\alpha_1)$, which is also the asymptotic variance of the properly scaled and centered sample mean. Unfortunately, we know neither the unconditional distribution of the ARCH(1) process nor the joint distribution of ε_t and ε_{t-1}, so that we are unable to derive the asymptotic efficency gains possible through the employment of the Stone (1975) or semiparametric efficient estimators. Note that this model does have the first-order Markov property assumed in our construction of the semiparametric efficient estimator.

6.2. The Threshold Model

This model is a simplified version of the ARCH model, in which the conditional variance is an increasing step function of the lagged absolute value (and hence of the lagged square) of the process. It also implies first-order Markov data. Each point at which a jump occurs is a threshold value. In our Monte Carlo exercises, we will consider a simple case where there is one threshold value, so that the conditional variance can assume one of only two possible values, the smaller one when the lagged absolute value of the series is below the threshold value and the larger one otherwise. The conditional distribution is always Gaussian.

Our model is formalized as follows:

$$\begin{aligned} \varepsilon_t &= u_t\sqrt{h_t} \\ u_t &\sim iidN(0,1) \\ h_t &= \begin{cases} \sigma_A^2 & if\ |\varepsilon_{t-1}| < \alpha \\ \sigma_B^2 & otherwise, \end{cases} \end{aligned}$$

where $\sigma_B^2 > \sigma_A^2$ and α is the threshold value. An appealing feature of this model is that we know its unconditional distribution and so can calculate its information and hence the efficiency gain of the adaptive estimator over the sample mean. This unconditional is a mixture of two normals, with variances of σ_B^2 and σ_A^2, where the probability of a low variance draw equals the unconditional probability that $|\varepsilon_{t-1}| < \alpha$, which we denote by γ. So the unconditional density of ε is

$$f(\varepsilon) = \gamma N\left(0, \sigma_A^2\right) + (1-\gamma) N\left(0, \sigma_B^2\right),$$

where

$$\begin{aligned} \gamma &= prob\left(|\varepsilon_{t-1}| < \alpha\right) = p_B / \left(1 - p_A + p_B\right) \\ p_A &= prob\left(|\varepsilon_t| < \alpha \middle| h_t = \sigma_A^2\right) = \Phi\left(\alpha/\sigma_A\right) - \Phi\left(-\alpha/\sigma_A\right) \\ p_A &= prob\left(|\varepsilon_t| < \alpha \middle| h_t = \sigma_B^2\right) = \Phi\left(\alpha/\sigma_B\right) - \Phi\left(-\alpha/\sigma_B\right), \end{aligned}$$

and $\Phi(\cdot)$ is the standard Gaussian cdf. We can therefore compute the asymptotic efficiency gains possible for the Stone (1975) estimator over the sample mean. It should be possible to work out the joint distribution of ε_t and ε_{t-1} and compute the asymptotic efficiency gain of the semiparametric efficient estimator. This point will be pursued in Hodgson (1997).

6.3. The Markov Switching Model

This model shares with the threshold model the property that the conditional distribution of the process is Gaussian with a variance belonging to a finite set of possible values, and, like both models described above, it implies volatility clustering. However, it differs from both these models in that the conditional variance is not determined by lagged values of the series but rather follows an "exogenous" Markov process in which the probability of a high variance state is higher if the previous state was also high variance than if it was not.[3] Ironically (given its name), it also differs from these models in that it implies that $\{\varepsilon_t\}$ is not a Markov process.

We again consider the simplest case, in which there are only two states, and formalize our model as follows:

$$\begin{aligned} \varepsilon_t &= u_t \sqrt{h_t} \\ u_t &\sim iidN(0,1), \end{aligned}$$

and h_t follows a Markov process characterized by the following transition probabilities:

$$prob\left(h_t = \sigma_A^2 \middle| h_{t-1} = \sigma_A^2\right) = p$$

[3] See Hamilton (1989) for more on Markov switching models.

$$\begin{aligned} prob\left(h_t = \sigma_B^2 \middle| h_{t-1} = \sigma_A^2\right) &= 1-p \\ prob\left(h_t = \sigma_A^2 \middle| h_{t-1} = \sigma_B^2\right) &= 1-q \\ prob\left(h_t = \sigma_B^2 \middle| h_{t-1} = \sigma_B^2\right) &= q, \end{aligned}$$

where we assume that $\sigma_B^2 > \sigma_A^2$ and $p > 1-q$.

As with the threshold model, we know that the unconditional distribution of the Markov switching model is a mixture of Gaussian random variables with respective variances of σ_A^2 and σ_B^2. We have

$$f(\varepsilon) = \gamma N\left(0, \sigma_A^2\right) + (1-\gamma)\, N\left(0, \sigma_B^2\right),$$

where $\gamma = (1-q)\,/\,(2-p-q)$. Similarly, we should be able to work out the joint distribution of ε_t and ε_{t-1}.

6.4. Simulation Results

Tables 1-3 report the results of our Monte Carlo simulations for the threshold, ARCH, and Markov switching models, respectively. For the threshold model, we employ the parameter settings $\sigma_A^2 = 1/3$, $\sigma_B^2 = 27$, and $\alpha = 1.32$, in the ARCH model we set $\alpha_0 = 0.15$ and $\alpha_1 = 0.95$, while in the Markov model we have $\sigma_A^2 = 1/3$, $\sigma_B^2 = 27$, $p = 0.92$ and $q = 0.30$. As mentioned above, we cannot compute the asymptotic efficiency gains possible through use of the Stone (1975) or semiparametric efficient estimators in the ARCH models, but for the other two models we can compute the efficiency gain of the former estimator over the sample mean and should be able to do likewise for the latter. The parameters for the threshold and Markov models are chosen so that $\gamma = 0.9$, so that the unconditional distribution of these two models is the same as that of the iid data used in simulation studies by Hsieh and Manski (1987) and Hodgson (1995). The ratio of the asymptotic variances of the Stone (1975) estimator and the sample mean in this case is 0.13. In each of Tables 1 to 3, we report the MSE figures for the sample mean (SM), Stone estimator (ST), and our semiparametric efficient estimator (SE), with the ratios of the two latter estimators' MSE's with respect to that of the sample mean being reported in the final two columns of each table.

TABLES 1-3: MONTE CARLO RESULTS

Table 1: Threshold Model - MSE

$(\sigma_A^2 = 1/3; \sigma_B^2 = 27; \alpha = 1.32)$

n	MSE(SM)	MSE(ST)	MSE(SE)	ST/SM	SE/SM
100	2.92×10^{-2}	1.30×10^{-2}	1.98×10^{-2}	.45	.68
250	1.21×10^{-2}	4.52×10^{-3}	4.79×10^{-3}	.37	.40
500	6.06×10^{-3}	1.95×10^{-3}	1.84×10^{-3}	.32	.30
750	4.01×10^{-3}	1.31×10^{-3}	1.03×10^{-3}	.33	.26
1000	2.79×10^{-3}	9.16×10^{-4}	7.08×10^{-4}	.33	.25

Table 2: ARCH(1) Model - MSE

$(\alpha_0 = 0.15; \alpha_1 = 0.95)$

n	MSE(SM)	MSE(ST)	MSE(SE)	ST/SM	SE/SM
100	1.74×10^{-2}	1.00×10^{-2}	1.34×10^{-2}	.57	.77
250	7.25×10^{-3}	2.72×10^{-3}	3.66×10^{-3}	.38	.50
500	2.84×10^{-3}	1.14×10^{-3}	1.17×10^{-3}	.40	.41
750	3.67×10^{-3}	1.41×10^{-3}	1.99×10^{-3}	.38	.54
1000	2.47×10^{-3}	6.25×10^{-4}	1.04×10^{-4}	.25	.42

Table 3: Markov Switching Model - MSE

$(\sigma_A^2 = 1/3; \sigma_B^2 = 27; p = 0.92; q = 0.30)$

n	MSE(SM)	MSE(ST)	MSE(SE)	ST/SM	SE/SM
100	3.43×10^{-2}	1.53×10^{-2}	1.74×10^{-2}	.45	.51
250	1.30×10^{-2}	5.09×10^{-3}	4.65×10^{-3}	.39	.36
500	6.00×10^{-3}	2.02×10^{-3}	1.64×10^{-3}	.34	.27
750	4.27×10^{-3}	1.18×10^{-3}	9.94×10^{-4}	.28	.23
1000	3.26×10^{-3}	7.89×10^{-4}	7.17×10^{-4}	.24	.22

Notes on Tables 1-3:

(a) 1000 iterations were used for each model.

(b) SM = sample mean; ST = Stone (1975) pseudo-adaptive estimator; SE = semiparametric efficient estimator with *m*=1.

(c) A/B = ratio of MSE's of estimators A and B, respectively.

(d) The Silverman (1986) rule-of-thumb bandwidths were used in the computation of ST and SE.

The results show that, in general, the two semiparametric estimators considered improve substantially upon the sample mean for all sample sizes, with the degree of improvement increasing in sample size. Although SE is asymptotically superior to ST, we would expect that in small samples it may not perform as well because of the problems induced by the nonparametric kernel estimation of a multivariate density. The primary question we address in these simulations is therefore that of the sort of sample size that is required for the asymptotic efficiency gains of SE to be realized in practice. The three tables suggest that the break-even point occurs at a sample size of approximately 500, although for the Markov switching model SE already beats ST at n=250. For the threshold and Markov models, we find that for sample sizes of 750 and 1000, SE reduces the MSE of ST by nearly 25%. The ARCH results are rather odd. For samples of 750 and 1000, SE deteriorates relative to ST, even though at n=500 the estimators have equal performance. These results must be considered highly suspect. Notice from the third and fourth rows of Table 2 that all three estimators' MSE's *increase* when the sample is increased from 500 to 750. This strange result may be occurring because the ARCH parameter $\alpha_1 = 0.95$ is so close to the boundary of the set within which the process has finite variance, the data are behaving in a manner similar to data generated by an infinite-variance process. As mentioned above, more extensive analysis of the practical properties of our estimator will be reported in Hodgson (1997). Nevertheless, the fundamental message of Tables 1-3 is that SE always improves upon SM, and for sample sizes of 500 and over is capable of consistently improving upon ST as well.

7. CONCLUSIONS

We have investigated the problem of semiparametric efficient estimation in time series models and found that, in the presence of dependence of unknown form, the semiparametric efficiency bound is achieved by an estimator that utilizes nonparametric estimates of the score of the density of the martingale difference innovation process conditional on its past. The derivation of the efficiency bound and the semiparametric kernel estimator proposed in the paper rely on the assumption that the data are stationary and that the conditional densities are symmetric about zero. The kernel estimator depends on the further assumption that the data are m^{th}-order Markov, but its asymptotic distribution is invariant to the failure of this assumption, as well as to certain departures of the data from the stationarity assumption, as are the estimated standard errors. We develop the estimation theory for the location model of Stone (1975) and the time series regression model with ARMA errors of Steigerwald (1992), with extensions to many other models of interest in time series econometrics being possible.

The methodology is implemented through a brief simulation study which indicates that the semiparametric efficient estimator improves considerably upon the Gaussian pseudo-MLE in terms of mean-square error for samples as small as 100 and has "caught up" with the pseudo-adaptive estimator, which relies on kernel estimates of the score of the unconditional density of the data, by the time the sample size has reached 500, for all models considered. For larger samples, the semiparametric efficient estimator surpasses the pseudo-adaptive estimator for two of the three models considered.

The work reported in this paper undoubtedly has close relationships with previous work. Pagan and Hong (1990) and Pagan and Schwert (1990) have considered nonparametric estimation of conditional variance models. If all the dependence present in a time series occurs through the second moment, then this approach could presumably be used to compute estimators in location models that achieve the semiparametric efficiency bound, provided that the parametric family to which the conditional density belongs is well specified. On the other hand, if we are willing to put our trust in the model of conditional variance that we use, but are uncertain about the parametric family to which the conditional density belongs, then we can estimate the density nonparametrically following Engle and Gonzalez-Rivera (1991), Linton (1993), or Drost and Klaassen (1996). To date we know of no other methods that nonparametrically treat both the conditional dependence and the non-Gaussianity present in the data, although it may be possible to show that the seminonparametric estimation strategy of Gallant and Tauchen (1989) is applicable to our problem and, under certain conditions, produces semiparametric efficient estimators asymptotically equivalent to that proposed here. Similarly, an estimator recently developed by Kuersteiner (1996) is related to the problem considered here. We intend to investigate the semiparametric efficiency properties of these estimators and their practical implications in Hodgson (1997).

Appendix A: A derivation of the bound using results of Ibragimov and Khasminski (1991)

To simplify exposition, we shall only state the proof for the case of first-order Markov data. A proof for the more general model would follow similar lines. We begin by defining the following vector of unknown parameters: $\nu = (\theta, g(x_1 | x_2))$, where ν is an element of the parameter space $\Xi = \Theta \times \Gamma$, where the properties of Θ are stated in the text and Γ is the space of symmetric conditional density functions $g(x_1 | x_2)$ satisfying the properties described in the text. The observed sample of size n is generated according to the sequence of probability measures $P_{\nu,n}$, where $P_{\nu,n}$ is the same as $P_{\theta,n}$ as described in the text, but with the infinite-dimensional parameter $g(x_1 | x_2)$ now included in the parameter vector.

We begin our derivation of the semiparametric efficiency bound by showing that the family of probability measures $\{P_{\nu,n}; \nu \in \Xi\}$ is locally asymptotically normal (LAN) according to the definition of Ibragimov and Khas'minski (1991, p.1682). We define the Hilbert space $\mathbf{H} = \overline{\mathbf{H}}_1 + \overline{\mathbf{H}}_2$, where $\mathbf{H}_1$ is the Hilbert space containing functions of the form

$$h_1(x_1, x_2) = \frac{-kg_1(x_1 | x_2)}{\sqrt{g(x_1 | x_2) / f(x_2)}} - \frac{kg_2(x_1 | x_2)}{\sqrt{g(x_1 | x_2) / f(x_2)}}$$

(where $f(\cdot)$ is the unconditional density of the innovations ε and k is a constant), and $\mathbf{H}_2$ is the set of all bounded, integrable functions $h_2(x_1, x_2)$ having the properties that $h_2(x_1, x_2) = h_2(-x_1, x_2)$ $a.s.$ x_2 and $\int h_2(x_1, x_2) \sqrt{g(x_1 | x_2)} dx_1 = 0$ $a.s.$ x_2. We define the norm $\|\cdot\|_H$ on $\mathbf{H}$ as follows: for every $h \in \mathbf{H}$, $\|h\|_H^2 = \int (h_1(x_1, x_2) f(x_2) + h_2(x_1, x_2))^2 dx_1 dx_2$ (it can be shown that this is indeed a norm).

We define the sequence of linear operators $\{A_n\}$, which map $\mathbf{H}$ into $R \times L_2$, as follows:

$$A_n(h) = n^{-1/2} \begin{bmatrix} I_{11}^{-1} \int \left(\frac{-g_1(x_1|x_2)}{\sqrt{g(x_1|x_2)/f(x_2)}} \right) h_1(x_1, x_2) dx_1 dx_2 \\ \sqrt{g(x_1 | x_2) / f(x_2)} h_2(x_1, x_2) \end{bmatrix}, \tag{11}$$

where $I_{11} = \int \frac{g_1(x_1|x_2)^2 f(x_2)}{g(x_1|x_2)} dx_1 dx_2$. Note that, for $h \in \mathbf{H}$, we have

$$\nu + A_n(h) = \nu + n^{-1/2} \begin{bmatrix} k \\ \sqrt{g(x_1 | x_2) / f(x_2)} h_2(x_1, x_2) \end{bmatrix}$$

We must check that conditions 1-3 on p.1682 of Ibragimov and Khas'minski are satisfied. The first condition, that $\lim_{n \to \infty} \|A_n(h)\| = 0 \; \forall h \in \mathbf{H}$, clearly holds for any norm on $R \times L_2$. Condition 2 will follow if we can show that, for every $h \in \mathbf{H}$, there exists n sufficiently large that $\nu + A_n(h) \in \Xi$. We first note that, for n sufficiently large, the result $\theta + n^{-1/2}k \in \Theta$ holds, since $\theta \in int\Theta$. We also claim that the function $g(x_1|x_2) + \sqrt{g(x_1|x_2)/f(x_2)}h_2(x_1, x_2)$ is a density function symmetric about zero in x_1. This follows from the symmetry of $g(x_1|x_2)$ and $h_2(x_1, x_2)$ and from the fact that $\int \sqrt{g(x_1|x_2)/f(x_2)}h_2(x_1, x_2)\,dx_1 = 0$.

To complete our verification of the LAN conditions, we must analyze the asymproric behaviour of the likelihood ratio $\Lambda_n(\nu + A_n(h), \nu) = \frac{dP_{\nu+A_n(h),n}}{dP_{\nu,n}}$. Under standard assumptions, we can ignore the initial conditions asymptotically and base our analysis on the following approximation of the likelihood ratio:

$$\Lambda_n(\nu + A_n(h), \nu) = \prod_{t=1}^{n} \left\{ \frac{g\left(y_t - \theta - n^{-1/2}k \,\middle|\, y_{t-1} - \theta - n^{-1/2}k\right)}{g(y_t - \theta\,|y_{t-1} - \theta)} + \frac{n^{-1/2}\sqrt{\frac{g\left(y_t - \theta - n^{-1/2}k \middle| y_{t-1} - \theta - n^{-1/2}k\right)}{f\left(y_{t-1} - \theta - n^{-1/2}k\right)}}h_2\left(y_t - \theta - n^{-1/2}k, y_{t-1} - \theta - n^{-1/2}k\right)}{g(y_t - \theta\,|y_{t-1} - \theta)} \right\}. \tag{12}$$

We can expand the numerator of the first component on the RHS of (12) to obtain

$$\begin{aligned} g\left(y_t - \theta - n^{-1/2}k \,\middle|\, y_{t-1} - \theta - n^{-1/2}k\right) &= g_t(\theta) - n^{-1/2}k\left(g_{1t}(\theta) + g_{2t}(\theta)\right) \\ &+ (2n)^{-1}k^2\left(g_{11t}(\theta) + g_{21t}(\theta) + g_{12t}(\theta) + g_{22t}(\theta)\right) + O_p\left(n^{-3/2}\right), \end{aligned} \tag{13}$$

where we write $g_t(\theta) = g(y_t - \theta\,|y_{t-1} - \theta)$, etc., to prevent notational clutter. Using the notation

$$z_t(\theta) = \sqrt{\frac{g(y_t - \theta\,|y_{t-1} - \theta)}{f(y_{t-1} - \theta)}}h_2(y_t - \theta, y_{t-1} - \theta),$$

we can write the numerator of the second component on the RHS of (12) as

$$n^{-1/2}z_t\left(\theta + n^{-1/2}k\right) = n^{-1/2}z_t(\theta) - n^{-1}k\left(z_{1t}(\theta) + z_{2t}(\theta)\right) + O_p\left(n^{-3/2}\right), \tag{14}$$

where z_{1t} and z_{2t} denote the derivatives of z_t with respect to ε_t and ε_{t-1}, respectively. Substituting (13) and (14) into (12), we obtain

$$\Lambda_n(\nu + A_n(h), \nu) = \prod_{t=1}^{n} \left\{ 1 - n^{-1/2}k\left(\frac{g_{1t}(\theta)}{g_t(\theta)} + \frac{g_{2t}(\theta)}{g_t(\theta)} - \frac{z_t(\theta)}{kg_t(\theta)}\right) \right.$$

$$+\left(2n\right)^{-1}k^{2}\left(\frac{g_{11t}\left(\theta\right)+g_{21t}\left(\theta\right)+g_{12t}\left(\theta\right)+g_{22t}\left(\theta\right)}{g_{t}\left(\theta\right)}-\frac{2\left(z_{1t}\left(\theta\right)+z_{2t}\left(\theta\right)\right)}{kg_{t}\left(\theta\right)}\right) \tag{15}$$

$$+O_{p}\left(n^{-3/2}\right)\Big\}.$$

We claim that (15) can be used to show that

$$\Lambda_{n}\left(\nu+A_{n}\left(h\right),\nu\right)=\exp\left\{-\frac{k}{n^{1/2}}\sum_{t=1}^{n}\left(\frac{g_{1t}\left(\theta\right)}{g_{t}\left(\theta\right)}+\frac{g_{2t}\left(\theta\right)}{g_{t}\left(\theta\right)}-\frac{z_{t}\left(\theta\right)}{kg_{t}\left(\theta\right)}\right)\right.$$

$$\left.-\frac{k^{2}}{2n}\left(\sum_{t=1}^{n}\left(\frac{g_{1t}\left(\theta\right)}{g_{t}\left(\theta\right)}+\frac{g_{2t}\left(\theta\right)}{g_{t}\left(\theta\right)}-\frac{z_{t}\left(\theta\right)}{kg_{t}\left(\theta\right)}\right)\right)^{2}+o_{p}(1)\right\}. \tag{16}$$

To prove this, we apply to the RHS of (16) the formula

$$\exp\left(x\right)=1+x+\frac{1}{2}x^{2}+\frac{1}{6}x^{3}+\cdots$$

and use the fact that

$$n^{-1}\sum_{t=1}^{n}\left(\frac{g_{11t}\left(\theta\right)+g_{21t}\left(\theta\right)+g_{12t}\left(\theta\right)+g_{22t}\left(\theta\right)}{g_{t}\left(\theta\right)}-\frac{2\left(z_{1t}\left(\theta\right)+z_{2t}\left(\theta\right)\right)}{kg_{t}\left(\theta\right)}\right)=o_{p}(1).$$

Defining the quantity

$$\Delta_{n}\left(h\right)=-\frac{k}{n^{1/2}}\sum_{t=1}^{n}\left(\frac{g_{1t}\left(\theta\right)}{g_{t}\left(\theta\right)}+\frac{g_{2t}\left(\theta\right)}{g_{t}\left(\theta\right)}-\frac{z_{t}\left(\theta\right)}{kg_{t}\left(\theta\right)}\right),$$

we can show that

$$\Delta_{n}\left(h\right)\xrightarrow{d}N\left(0,\left\|h\right\|_{H}^{2}\right),$$

and that

$$\frac{k^{2}}{n}\left(\sum_{t=1}^{n}\left(\frac{g_{1t}\left(\theta\right)}{g_{t}\left(\theta\right)}+\frac{g_{2t}\left(\theta\right)}{g_{t}\left(\theta\right)}-\frac{z_{t}\left(\theta\right)}{kg_{t}\left(\theta\right)}\right)\right)^{2}=\left\|h\right\|_{H}^{2}+o_{p}(1).$$

It follows that

$$\Lambda_{n}\left(\nu+A_{n}\left(h\right),\nu\right)=\exp\left\{\Delta_{n}\left(h\right)-\frac{1}{2}\left\|h\right\|_{H}^{2}+o_{p}(1)\right\},$$

so that the conditions given by Ibragimov and Khas'minski (1991) for a model to be in the LAN family are satisfied.

Ibragimov and Khas'minski (1991) derive the semiparametric efficiency bound for estimation of some parameter of interest $\delta(\nu)$ when a model belongs to the LAN family. Since our parameter of interest is just θ, $\delta(\nu)$ is the function that identifies the first element of ν. The lower bound on the covariance is just the correlation operator KK^*, where the operator K is defined as

$$K = \lim_{n\to\infty} n^{1/2} \frac{\partial \delta(\nu)}{\delta \nu'} A_n P_H,$$

where P_H is the projection operator onto the space $\mathbf{H}$. For our model, $\frac{\partial\delta(\nu)}{\delta\nu'} = [1, 0]$, so K will be the first element of the operator $\widetilde{K} = \lim_{n\to\infty} n^{1/2} A_n P_H$. From Theorem A.4.5 of Bickel, Klaassen, Ritov, and Wellner (1993, p.444), we can write $P_H = P_1 + P_2$, where $P_1$is the projection operator onto the one-dimensional space spanned by $\frac{-g_1(x_1|x_2)}{\sqrt{g(x_1|x_2)/f(x_2)}}$, and P_2 is the projection operator onto $\mathbf{H}_2$. We can then write $K = \widetilde{A}_1(P_1+P_2)$, where $\widetilde{A}_1(w) = I_{11}^{-1}\int\left(\frac{-g_1(x_1|x_2)}{\sqrt{g(x_1|x_2)/f(x_2)}}\right) w(x_1, x_2) f(x_2)\, dx_1 dx_2$, so that

$$K(w) = I_{11}^{-1}\int\left(\frac{-g_1(x_1|x_2)}{\sqrt{g(x_1|x_2)/f(x_2)}}\right) P_1 w(x_1, x_2)\, dx_1 dx_2,$$

since $\int\left(\frac{-g_1(x_1|x_2)}{\sqrt{g(x_1|x_2)/f(x_2)}}\right) h_2(x_1, x_2)\, dx_1 dx_2 = 0$ for every $h_2 \in \mathbf{H}_2$. It then follows that the semiparametric efficiency bound is $KK^* = I_{11}^{-1}$, the result obtained in the text.

Appendix B: Proofs of Theorems and Lemmas

Proof of Theorem 3: Subtracting θ_0 from both sides of (3), multiplying by $n^{1/2}$, substituting a mean value expansion of $n^{-1}\sum_{t=1}^n s_t(\varepsilon_t(\theta_n^*)|\varepsilon^{t-1}(\theta_n^*))$ about θ_0 into the right hand side, and writing $s_t(\theta) = s_t(\varepsilon_t(\theta)|\varepsilon^{t-1}(\theta))$, we obtain

$$n^{1/2}\left(\theta_n^+ - \theta_0\right) = n^{1/2}(\theta_n^* - \theta_0) + \left[n^{-1}\sum_{t=1}^n s_t(\theta_n^*)^2\right]^{-1} n^{1/2}\sum_{t=1}^n s_t(\theta_0)$$

$$+ n^{1/2}(\theta_n^* - \theta_0)\left[n^{-1}\sum_{t=1}^n s_t(\theta_n^*)^2\right]^{-1}\left[n^{-1}\sum_{t=1}^n \frac{\partial}{\partial\theta} s_t(\theta_0)\right]. \qquad (17)$$

The theorem will then follow from (17) and the facts, established in the text, that $n^{-1}\sum_{t=1}^{n} s_t(\theta_n^*)^2 = \mathcal{B}^{-1} + o_p(1)$, $n^{-1}\sum_{t=1}^{n} \frac{\partial}{\partial\theta} s_t(\theta_0) = -\mathcal{B}^{-1} + o_p(1)$, and $\mathcal{B}^{1/2} n^{1/2} \sum_{t=1}^{n} s_t(\theta_0) \xrightarrow{d} N(0,1)$.

Proof of Theorem 4: The proof follows similar lines to those of the proof of Lemma 2 in Hodgson (1996). To establish (6), it is sufficient to check the following two conditions:

$$n^{-1/2}\sum_{t=1}^{n}\left[\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_n^*\right),\theta_n^*\right) - s_t\left(\theta_n^*\right)\right] = o_p(1) \tag{18}$$

and

$$n^{-1}\sum_{t=1}^{n}\left[\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_n^*\right),\theta_n^*\right)^2 - s_t\left(\theta_n^*\right)^2\right] = o_p(1). \tag{19}$$

A consequence of our LAN result in Appendix A is that the sequences of probability measures $\{P_{\theta_0,n}\}$ and $\left\{P_{\theta_n^*,n}\right\}$ are contiguous, so that to obtain (18) it is sufficient to prove

$$n^{-1/2}\sum_{t=1}^{n}\left[\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_0\right),\theta_0\right) - s_t\left(\theta_0\right)\right] = o_p(1). \tag{20}$$

We shall establish (20) by proving the following mean-square consistency result:

$$n^{-1}E\left[\sum_{t=1}^{n}\left(\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_0\right),\theta_0\right) - s_t\left(\theta_0\right)\right)\right]^2 \to 0.$$

Under our symmetry assumptions, and due to our manner of constructing $\widehat{s}_t$, it follows that the difference $\{\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_0\right),\theta_0\right) - s_t\left(\theta_0\right)\}$ is a martingale difference sequence, so that we need only show that

$$n^{-1}\sum_{t=1}^{n}E\left[\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_0\right),\theta_0\right) - s_t\left(\theta_0\right)\right]^2 \to 0$$

which follows if we can show

$$E\left[\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_0\right),\theta_0\right) - s_t\left(\theta_0\right)\right]^2 \to 0 \quad \forall t = 1, ..., n, \tag{21}$$

i.e., that

$$\int\left\{q_t\left(\widetilde{\varepsilon}\right) - \frac{f'}{f}\left(\widetilde{\varepsilon}\right)\right\}^2 f\left(\widetilde{\varepsilon}\right)d\widetilde{\varepsilon} \to 0, \tag{22}$$

where we have written $q_t(\tilde{\varepsilon}) = q_t(\tilde{\varepsilon}(\theta_0), \theta_0)$, where $f(\tilde{\varepsilon})$ is an $m+1$-dimensional density function, and f' is the first partial of f with respect to the first element of $\tilde{\varepsilon}$.

We can now establish (22) following the argument of Bickel (1982, Lemma 4.1). We first prove that

$$\int \left\{ q_t(\tilde{\varepsilon}) - \frac{f_a'}{f_a}(\tilde{\varepsilon}) \right\}^2 f_a(\tilde{\varepsilon})\, d\tilde{\varepsilon} \to 0, \tag{23}$$

where f_a denotes the convolution of f and the $N(0, a_n^2 I_{m+1})$ density. Denoting by $g^{(\nu)}$ the ν^{th} partial derivative of a function g, for $\nu = 0, 1$ (when $\nu = 1$, $g^{(\nu)}$ refers to the first element of the vector of first partial derivatives), we have

$$E\left[\widehat{f}_t^{(\nu)}(x)\right] = E\left[\pi_{a_n}^{(\nu)}(x - \tilde{\varepsilon})\right] = \int \pi_{a_n}^{(\nu)}(x-y)\, F(dy) = f_a^{(\nu)}(x).$$

In order to apply the argument of Bickel (1982, Lemma 6.1) to prove (23), we must derive a bound for the variance of $\widehat{f}_t^{(\nu)}(x)$, $\nu = 0, 1,$. We have

$$V_{nt}^{(\nu)}(x) = var\left[\widehat{f}_t^{(\nu)}(x)\right] = E\left[\widehat{f}_t^{(\nu)}(x) - f_a^{(\nu)}(x)\right]^2$$

$$= E\left[(n-1)^{-1} \sum_{\substack{j=1 \\ j \neq t}}^{n} \left(\pi_{a_n}^{(\nu)}(x - \tilde{\varepsilon}_j) - f_a^{(\nu)}(x)\right)\right]^2$$

$$= (n-1)^{-2} E\left[\sum_{\substack{j=1 \\ j \neq t}}^{n} z_j\right]^2,$$

where $z_j = \pi_{a_n}^{(\nu)}(x - \tilde{\varepsilon}_j) - f_a^{(\nu)}(x)$. It follows that

$$V_{nt}^{(\nu)}(x) = (n-1)^{-2} E\left[\sum_{j=1}^{n} z_j - z_t\right]^2$$

$$= (n-1)^{-2} E\left[\sum_{j=1}^{n} z_j^2\right] - 2(n-1)^{-2} E\left[\sum_{j=1}^{n} z_j z_t\right] + (n-1)^{-2} E\left[z_t^2\right]. \tag{24}$$

Now we note that because $\{\varepsilon_t\}$ is mixing, so is $\{\tilde{\varepsilon}_t\}$ and also $\{z_t\}$. Define $\gamma_\tau = E\left[z_t z_{t-\tau}\right]$ and $\rho_\tau = \gamma_\tau/\gamma_0$. Our mixing assumption implies that $|\rho_\tau| \to 0$ as $\tau \to \infty$ (White and Domowitz (1984)). From (24), we have

$$\begin{aligned} V_{nt}^{(\nu)}(x) &= n(n-1)^{-2}\gamma_0 - 2(n-1)^{-2}\sum_{j=1}^{n}\gamma_{j-t} + (n-1)^{-2}\gamma_0 \\ &= \gamma_0\left[n(n-1)^{-2} - 2(n-1)^{-2}\sum_{j=1}^{n}\rho_{j-t} + (n-1)^{-2}\right] \\ &\leq \frac{M}{n-1}\gamma_0, \end{aligned}$$

since there exists some $0 < M < \infty$ such that $n(n-1)^{-1} - 2(n-1)^{-1}\sum_{j=1}^{n}\rho_{j-t} + (n-1)^{-1} < M$ for every t and n. It follows that

$$V_{nt}^{(\nu)}(x) \leq \frac{M}{n-1}E\left[\pi_{a_n}^{(\nu)}(x-\tilde{\varepsilon})\right]^2.$$

Now,

$$\left(\pi_{a_n}^{(\nu)}(x)\right)^2 \leq \frac{\kappa_\nu}{a_n^{2\nu+m+1}}\pi_{a_n}(x)$$

for some constant κ_ν (see Stone (1975) and Jeganathan (1995)). We therefore have

$$\frac{M}{n-1}E\left[\pi_{a_n}^{(\nu)}(x-\tilde{\varepsilon})\right]^2 \leq \frac{M\kappa_\nu}{(n-1)a_n^{2\nu+m+1}}E\left[\pi_{a_n}(x-\tilde{\varepsilon})\right] = \frac{M\kappa_\nu}{(n-1)a_n^{2\nu+m+1}}f_a(x),$$

yielding

$$V_{nt}^{(\nu)}(x) \leq \frac{\tau_\nu}{(n-1)a_n^{2\nu+m+1}}f_a(x),$$

where $\tau_\nu = M\kappa_\nu$. The proof of (23) now follows the same lines as that of Lemma 6.1 of Bickel (1982), while the proof of (22) and therefore (18) is completed by applying Lemmas 6.2 and 6.3 of Bickel (1982). We can apply standard methods (cf. Bickel (1982) or Kreiss (1987)) to show that (19) then follows.

Proof of Lemma 3: Our first step in proving the convergence result is to establish the consistency of $\widehat{\theta}_n$. The proof follows standard lines so we shall not go into complete detail. We note that, under appropriate regularity conditions, we can apply Theorem 2 of Pötscher and Prucha (1989) to prove that

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{t=1}^{n}\left[s\left(\widetilde{\varepsilon}_t\left(\theta\right)\right)-Es\left(\widetilde{\varepsilon}_t\left(\theta\right)\right)\right]\right|=o_{a.s.}(1),$$

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{t=1}^{n}s\left(\widetilde{\varepsilon}_t\left(\theta\right)\right)-\int s\left(\widetilde{\varepsilon}\left(\theta\right)\right)dF\left(\widetilde{\varepsilon}_t\left(\theta_0\right)\right)\right|=o_{a.s.}(1),$$

and

$$\left\{n^{-1}\sum_{t=1}^{n}Es\left(\widetilde{\varepsilon}_t\left(\theta\right)\right)\right\}$$

is equicontinuous on Θ. These conditions, along with an assumption of θ_0 being the identifiably unique zero of $n^{-1}\sum_{t=1}^{n}s\left(\widetilde{\varepsilon}_t\left(\theta\right)\right)$, will yield consistency. To obtain asymptotic normality, we again use the mean-value expansion

$$n^{1/2}\left(\widehat{\theta}_n-\theta_0\right)=-\left[n^{-1}\sum_{t=1}^{n}\left(s'\left(\widetilde{\varepsilon}_t\left(\overline{\theta}_n\right)\right)+\sum_{j=1}^{m}s^j\left(\widetilde{\varepsilon}_t\left(\overline{\theta}_n\right)\right)\right)\right]^{-1}$$

$$\cdot\left[n^{-1/2}\sum_{t=1}^{n}s\left(\widetilde{\varepsilon}_t\left(\theta_0\right)\right)\right].$$

We can again apply the results of Pötscher and Prucha (1989), along with the appropriate central limit theorem for martingale difference sequences, to prove that

$$n^{-1}\sum_{t=1}^{n}s^j\left(\widetilde{\varepsilon}_t\left(\overline{\theta}_n\right)\right)=o_p(1)\quad\forall j=1,...,m,$$

$$n^{-1}\sum_{t=1}^{n}s'\left(\widetilde{\varepsilon}_t\left(\overline{\theta}_n\right)\right)=-\mathcal{I}+o_p(1),$$

and

$$n^{-1/2}\sum_{t=1}^{n}s\left(\widetilde{\varepsilon}_t\left(\theta_0\right)\right)\xrightarrow{d}N\left(0,\mathcal{I}\right),$$

from which three conditions we obtain the desired result. The proof of (8) follows the same lines as our proof of Theorem 3, only now making use of Theorem 2 of Pötscher and Prucha (1989) in the manner just described.

Proof of Theorem 5: By arguments similar to those we made in the proof of Theorem 4, all results of the Theorem will follow easily once we have established that

$$E\left[\widehat{s}_t\left(\widetilde{\varepsilon}_t\left(\theta_0\right),\theta_0\right) - s\left(\widetilde{\varepsilon}_t\left(\theta_0\right)\right)\right]^2 \to 0 \;\; \forall t = 1,...,n, \tag{25}$$

i.e., that

$$\int \left\{ q_t\left(\widetilde{\varepsilon}\right) - \frac{f'}{f}\left(\widetilde{\varepsilon}\right)\right\}^2 f\left(\widetilde{\varepsilon}\right) d\widetilde{\varepsilon} \to 0. \tag{26}$$

Our proof of (26) is similar to the proof of Lemma 4.1 in Bickel (1982) and of Lemma 2 in Hodgson (1996) and proceeds in three steps, the first of which is to prove that

$$\int \left\{ q_t\left(\widetilde{\varepsilon}\right) - \frac{\overline{f}_n'}{\overline{f}_n}\left(\widetilde{\varepsilon}\right)\right\}^2 \overline{f}_n\left(\widetilde{\varepsilon}\right) \to 0, \tag{27}$$

where $\overline{f}_n\left(\widetilde{\varepsilon}\right) = n^{-1}\sum_{t=1}^{n} f_t\left(\widetilde{\varepsilon}\right)$. But to prove (27), we must verify that

$$\int \left\{ q_t\left(\widetilde{\varepsilon}\right) - \frac{\overline{f}_{na}'}{\overline{f}_{na}}\left(\widetilde{\varepsilon}\right)\right\}^2 \overline{f}_{na}\left(\widetilde{\varepsilon}\right) \to 0, \tag{28}$$

where $\overline{f}_{na}$ denotes the convolution of $\overline{f}_n$ and the $N\left(0, a_n^2 I_{m+1}\right)$ density, as follows:

$$\overline{f}_{na} = \overline{f}_n * \pi_{a_n} = n^{-1}\sum_{t=1}^{n} f_t * \pi_{a_n} = n^{-1}\sum_{t=1}^{n} f_{ta}.$$

To prove (28), we can follow Lemma 6.1 of Bickel (1982), using an argument smilar to that in the proof of Theorem 4 above to show that

$$V_{nt}^{(\nu)}\left(x\right) \leq \frac{\tau_\nu}{(n-1)a_n^{2\nu+m+1}}\overline{f}_{na}\left(x\right).$$

To complete our proof of (27), we can use Lemmas 6.2 and 6.3 of Bickel (1982) to obtain the following two convergence results:

$$\int_{\overline{f}_n>0} \left\{ \frac{\overline{f}_{na}'}{\sqrt{\overline{f}_{na}}}\left(\widetilde{\varepsilon}\right) - \frac{\overline{f}_n'}{\sqrt{\overline{f}_n}}\left(\widetilde{\varepsilon}\right)\right\}^2 d\widetilde{\varepsilon} \to 0,$$

and

$$\int_{\overline{f}_n>0} q_t^2\left(\widetilde{\varepsilon}\right)\left(\sqrt{\overline{f}_{na}}\left(\widetilde{\varepsilon}\right) - \sqrt{\overline{f}_n}\left(\widetilde{\varepsilon}\right)\right)^2 d\widetilde{\varepsilon} \xrightarrow{p} 0.$$

This establishes (27) and completes the first step in our proof of (26). The second step is to show

$$\int_{f>0}\left\{\frac{\overline{f}'_n}{\sqrt{\overline{f}_n}}(\tilde{\varepsilon}) - \frac{f'}{\sqrt{f}}(\tilde{\varepsilon})\right\}^2 d\tilde{\varepsilon} \to 0,$$

which can be done using Lemma 6.2 of Bickel (1982). The final step is to verify

$$\int_{f>0} q_t^2(\tilde{\varepsilon})\left(\sqrt{\overline{f}_n}(\tilde{\varepsilon}) - \sqrt{f}(\tilde{\varepsilon})\right)^2 d\tilde{\varepsilon} \xrightarrow{p} 0.$$

But, recalling that $q_t^2(\tilde{\varepsilon}) < c_n^2$, the desired result will follow from our restriction on the rate of divergence of c_n.

References

[1] Baillie, R.T. and Bollerslev, T. 1989a. The message in daily exchange rates: A conditional variance tale. *Journal of Business and Economic Statistics* 7:297-305.

[2] Baillie, R.T. and Bollerslev, T. 1989b. Common stochastic trends in a system of exchange rates. *Journal of Finance* 44:167-181.

[3] Bekaert, G. and Hodrick, R.J. 1993. On biases in the measurement of foreign exchange risk premiums. *Journal of International Money and Finance* 12:115-138.

[4] Bera, A.K. and Higgins, M.L. 1993. ARCH models: Properties, estimation, and testing. *Journal of Economic Surveys* 7:305-362.

[5] Bickel, P.J. 1982. On adaptive estimation. *Annals of Statistics* 10:647-671.

[6] Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. 1993. *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore; Johns Hopkins University Press.

[7] Bierens, H. 1984. Model specification testing of time series regressions. *Journal of Econometrics* 26:323-353.

[8] Billingsley, P. 1968. *Convergence of Probability Measures.* New York; Wiley.

[9] Bollerslev, T., Chou, R.Y., and Kroner, K.F. 1992. ARCH modelling in finance: A review of the theory and empirical evidence. *Journal of Econometrics* 52:5-59.

[10] Cornell, B. 1989. The impact of data errors on measurement of the foreign exchange risk premium. *Journal of International Money and Finance* 8:147-157.

[11] Crowder, M. 1976. Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society B* 38:45-53.

[12] Drost, F.C. and Klaassen, C.A.J. 1996. Efficient estimation in semiparametric GARCH models. Forthcoming, *Journal of Econometrics.*

[13] Engle, R.F. 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50:987-1008.

[14] Fama, E.F. 1984. Forward and spot exchange rates. *Journal of Monetary Economics* 14:319-338.

[15] Gallant, A.R. and Tauchen, G. 1989. Seminonparametric estimation of conditionally constrained heterogeneous processes: Asset pricing applications. *Econometrica* 57:1091-1120.

[16] Hamilton, J.D. 1989. A new approach to economic analysis of nonstationary time series and the business cycle. *Econometrica* 57:357-384.

[17] Hodgson, D.J. 1995. Adaptive estimation of cointegrated models: Simulation evidence and an application to the forward exchange market. RCER Working Paper #409, University of Rochester.

[18] Hodgson, D.J. 1996. Robust semiparametric estimation in the presence of heterogeneity of unknown form. RCER Working Paper #416, University of Rochester.

[19] Hodgson, D.J. 1997. On the semiparametric efficiency properties of some well-known estimators in time series. Work in progress.

[20] Hsieh, D.A. and Manski, C.F. 1987. Monte Carlo evidence on adaptive maximum likelihood estimation of a regression. *Annals of Statistics* 15:541-551.

[21] Ibragimov, I.A. and Khas'minski, R.Z. 1991. Asymptotically normal families of distributions and efficient estimation. *Annals of Statistics* 19:1681-1724.

[22] Kreiss, J.-P. 1987. On adaptive estimation in stationary ARMA processes. *Annals of Statistics*15:112-133.

[23] Kuersteiner, G. 1996. Efficient IV estimation for autoregressive models with conditional heteroskedasticity. Unpublished manuscript, Yale University.

[24] Linton, O. 1993. Adaptive estimation in ARCH models. *Econometric Theory* 9:539-569.

[25] Lye, J.L., Martin, V.L., and Teo,L. 1996. Parametric distributional flexibility and conditional variance models. Unpublished manuscript, University of Melbourne.

[26] McCallum, B.T. 1994. A reconsideration of the uncovered interest parity relationship. *Journal of Monetary Economics* 33:105-132.

[27] Newey, W.K. 1990. Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5:99-135.

[28] Newey, W.K. and West, K.D. 1987. A simple positive semi-definite, heteroskedasticity and autocorrealtion consistent covariance matrix. *Econometrica* 55:703-708.

[29] Pagan, A.R. and Hong, Y.S. 1991. Nonparametric estimation and the risk premium. In Barnett, W.A., Powell, J. and Tauchen, G., *Nonparametric and Semiparametric Methods in Econometrics and Statistics,* Cambridge.

[30] Pagan, A.R. and Schwert, G.W. 1990. Alternative models for conditional stock volatility. *Journal of Econometrics* 45:267-290.

[31] Pötscher, B.M. and Prucha, I.R. 1989. A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica* 57:675-683.

[32] Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis.* London; Chapman and Hall.

[33] Steigerwald, D. 1992. Adaptive estimation in time series regression models. *Journal of Econometrics* 54:251-276.

[34] Stone, C. 1975. Adaptive maximum likelihood estimation of a location parameter. *Annals of Statistics* 3:267-284.

[35] White, H. 1996. *Estimation, Inference, and Specification Analysis.* Cambridge.

[36] White, H. and Domowitz, I. 1984. Nonlinear regression with dependent observations. *Econometrica* 52:143-161.