Robust Semiparametric Estimation in the Presence of Heterogeneity of Unknown Form

Hodgson, Douglas J.

Working Paper No. 416
April 1996

Robust SemiparametricEstimation in the Presence of Heterogeneity of Unknown Form

Douglas J. Hodgson

Rochester Center for Economic Research
Working Paper No. 416

April 1996

Robust Semiparametric Estimation in the Presence of Heterogeneity of Unknown Form

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

April 26, 1996

Abstract

We show that semiparametric adaptive maximum likelihood estimators have desirable robustness properties when the innovations in a location parameter model are uncorrelated but not necessarily independent. We show that such estimators have asymptotic covariance matrices equal to the inverse of the Fisher information of the unconditional distribution of the data in the presence of general forms of heterogeneity, including conditional dependence in even moments. This result is important because it establishes that adaptive estimators can significantly improve upon standard techniques even when the independence assumption underlying the adaptive estimator is violated. It should be of great interest to empirical researchers because it implies that highly robust estimation is possible without the necessity of specifying a model of the higher order dependence that may be present in the data. A discussion of the semiparametric optimality properties of this estimator is included. We present Monte Carlo simulation results that illustrate the excellent performance of the adaptive

For their helpful comments, I would like to thank Bill Brown, Oliver Linton, and participants in workshops at Rochester and Cornell.

estimator for several varieties of conditional heteroskedasticity, including ARCH, Markov-switching, and threshold models.

JEL C22 - Time series models

1. INTRODUCTION

Econometric time series models generally assume that the stochastic process generating an observed sample can be reduced to a sequence of uncorrelated innovations. Efficient estimation of the parameters of a model is then carried out under the further assumption that these innovations are iid Gaussian. However, the assumptions of independence, of identical distributions, and of normality are all highly dubious for many forms of economic time series data.

It is by now a well-established empirical regularity that the innovations in many series, especially asset prices, have unconditional densities whose tails are considerably heavier than those of a normal. Given the poor efficiency properties of Gaussian pseudo-MLE's in such situations, considerable research effort has been devoted to the formulation of estimation strategies that will optimally account for this non-Gaussianity.

The technique of adaptive maximum likelihood estimation is emerging as a highly attractive alternative in this regard. If we are willing to assume that the innovations to a model are iid, then we can often compute an estimator that is asymptotically efficient, in the sense of having asymptotic covariance matrix equal to the inverse of the asymptotic information, even if we do not know the distribution from which these innovations are drawn. Adaptive estimation has been shown to be a possibility in many important econometric contexts, including location models (Stone (1975) and Beran (1974)), linear regressions (Bickel (1982)), non-linear models (Manski (1984)), ARMA models (Kreiss (1987a)), stationary time series regressions (Steigerwald (1992)), ARCH models (Linton (1993)), and cointegrated models (Jeganathan (1995) and Hodgson (1995a,b)).

It is natural to ask whether adaptive estimators retain their desirable robustness properties when the iid assumption on the innovations fails. The presence of various forms of conditional heterogeneity seems to be the rule for most economic data. Particular emphasis has been placed in recent years on the existence of dependence in second (and, to a lesser extent, fourth) moments, especially in financial time series, and on the fact that such dependence will induce non-Gaussianity in the unconditional density of the data even if a correct model of the

dependence implies Gaussian conditional densities. This has been a major theme of the ARCH literature.

In the context of a basic location parameter model such as that considered by Stone (1975), we derive an interesting and important robustness result for adaptive estimators that are computed assuming the data are iid when in fact they may be neither independently nor identically distributed. Under the crucial assumption that the density of each innovation conditional on the past is symmetric about zero, we find that the adaptive estimator has a normal asymptotic distribution with covariance matrix equal to the inverse of the information of the density of the unconditional distribution of the observations.[1] Our result allows for various forms of conditional heterogeneity.

The significance of this result is that the asymptotic efficiency gains obtained by the adaptive estimator relative to the Gaussian pseudo-MLE (the sample mean in the location model) are quantitatively identical whether the data are iid or not . In either case, the efficiency gain is the ratio of the variance of the unconditional distribution to the inverse of its information. The adaptive estimator is not necessarily fully efficient, since in the non-iid case full efficiency would require correct specification of the dependence and heterogeneity present in the data. However, in cases where our ability to achieve such a specification is highly dubious (i.e., the standard case in econometrics), this robustness property of adaptive estimators increases their appeal considerably.

As an illustration of the usefulness of our results, consider a situation in which we would like to estimate the location of a sequence of uncorrelated random variables for which we suspect the presence of unconditional non-normality and dependence in second (and possibly higher) moments. The sample mean is clearly an undesirable estimator. It fails to take account of either the non-normality or the dependence. Alternatively, we could assume that the data are generated by a Gaussian ARCH process and compute the corresponding MLE. If our assumptions are correct, then we have achieved full efficiency. However, there are several ways in which our assumptions may be incorrect, all of which are likely to be relevant in any given empirical application. First, that we have chosen correctly

[1] Our results actually allow for a limited degree of nonstationarity in the data, so that different observations are permitted to have different unconditional distributions. Hence, it is not strictly correct to speak in terms of the information of the unconditional distribution of the data (except in the stationary case); we should actually speak in terms of the information of the "average asymptotic unconditional distribution". However, we shall often abbreviate the latter expression to the former except in cases where our argument requires the more precise vocabulary.

from the long menu of possible ARCH specifications is unlikely. Second, even if we have somehow managed to correctly model the second moment dependence in the data, there is plenty of empirical evidence to suggest that the resulting standardized residuals are neither normally nor independently nor identically distributed. Conditional non-Gaussianity and higher-order dependence, in fourth moments, for example, have almost assumed the status of stylized facts in financial data. The virtue of computing an adaptive estimator based on the uncorrelated innovations is that we are not required to in any way specify the dependence or heterogeneity that may be present in the data, and yet we get an estimate that is robust to the presence of such factors, adapts itself optimally to the unconditional distribution implied by the presence of such factors, and delivers estimates considerably more efficient than the sample mean.

In Section 2, we present the model, our assumptions, and an analysis of the pseudo-MLE we would compute if we knew the unconditional distribution of the data (F, say) and assumed that the data were iid from this distribution (what Levine (1983) terms a "partial likelihood"). It turns out that this partial MLE is consistent and asymptotically normal with a variance equal to the inverse of the Fisher information of F. In Section 3, we show that a one-step iterative estimator constructed assuming that F is known is asymptotically equivalent to the partial MLE. We then show that an adaptive estimator very similar to that of Stone (1975), modified as in Bickel (1982) and Kreiss (1987a), is asymptotically equivalent to this iterative estimator and so to the partial MLE. In Section 4, we briefly consider the semiparametric optimality properties of the estimator when the dependence and heterogeneity in the data generating process are treated as unknown infinite-dimensional nuisance parameters. We conclude that our estimator does not attain the semiparametric efficiency bound and discuss the possibility of formulating one that would. In Section 5, we report Monte Carlo simulation results that illustrate the finite-sample efficiency gains possible with the adaptive estimator for a number of conditionally heterogeneous processes, including ARCH, Markov-switching, and threshold models. Section 6 discusses the applicability of our main result to more general models than the location model and Section 7 concludes.

2. THE PARTIAL MLE

Suppose that the observed sample $y_t \in R$, $t = 1, ..., n$, is generated by the following location model:

$$y_t = \theta_0 + \varepsilon_t, \tag{1}$$

where the zero-mean innovations ε_t are uncorrelated and have distribution functions F_t with the property that $F^n = n^{-1}\sum_{t=1}^{n} F_t \Rightarrow F$. We use the symbol $\Rightarrow$ to denote weak convergence of probability measures (cf. Billingsley (1968)). Associated with this model is the sequence of probability measures $\{P_{\theta_0,n}\}$. All convergence statements in the paper are under this sequence unless otherwise indicated. We assume that F has a density $f = F'$ which is symmetric about zero, twice continuously differentiable, and has finite, positive information. We define the σ-field $\Omega_{t-1} = \sigma(\varepsilon_{t-1}, \varepsilon_{t-2}, ...)$ and assume that, for every t, the innovation ε_t has a conditional density $g_t(\varepsilon\,|\Omega_{t-1})$ which is symmetric about zero in ε, has finite, positive information, and is absolutely continuous in ε. We also assume the existence of an unobservable initial vector ε_0, which we use to determine the conditional distribution of the sample and whose density we denote by $f_0(\varepsilon_0; \theta)$. We assume that

$$f_0(\varepsilon_0; \theta_1) - f_0(\varepsilon_0; \theta_2) = o_p(1)$$

in $P_{\theta_1,n}$ whenever $\theta_1, \theta_2 \in \Theta$ and $\theta_1 - \theta_2 = o_p(1)$.

If the data are stationary, then F is just the unconditional distribution of the innovations. However, this formulation allows for various forms of non-stationarity, including deterministic heteroskedasticity. Assuming the innovations are independent, for example, if an asymptotic proportion of the sample of α has a normal distribution with variance σ_1^2 (Φ_1, say) and the remaining sample is normal with variance σ_2^2 (with cdf of Φ_2, say), then it will follow that $F = \alpha\Phi_1 + (1-\alpha)\Phi_2$.

Our ultimate objective is to analyze the properties of an adaptive estimator of $\theta_0 \in \Theta$, similar to that of Stone (1975), derived under the assumption that the $\{\varepsilon_t\}$ are iid from an unknown symmetric density. The first step is to analyze the behaviour of the "partial MLE" (cf. Levine (1983)) that we would compute if we knew F and assumed that the innovations were iid draws from its density. The partial MLE is

$$\widehat{\theta}_n = \arg\max_{\theta\in\Theta} n^{-1}\sum_{t=1}^{n} \ln f(y_t - \theta), \tag{2}$$

which also satisfies

$$n^{-1}\sum_{t=1}^{n}\psi\left(y_t-\widehat{\theta}_n\right)=0, \tag{3}$$

where $\psi = f'/f$ is the (negative of the) score of f. Our objective in this section is to show that $\widehat{\theta}_n$ is consistent and asymptotically normal with variance equal to the inverse of $I_f = \int \frac{(f')^2}{f}(x)dx$, the Fisher information of f. This result is stated in Theorem 2 below. We shall then see in Section 3 that the adaptive estimator is asymptotically equivalent to the partial MLE.

We first prove that $\widehat{\theta}_n$ is consistent. To this end, we must show that the criterion function on the right-hand side of (2) satisfies a uniform law of large numbers (ULLN), for which purpose we appeal to a result of Potscher and Prucha (1989), who generalize a ULLN of Bierens (1984). We begin with a statement of our assumptions.

ASSUMPTION 1: (Θ,ρ) *is a compact metric space.*

In order to state our next two assumptions, we must introduce some notation. Let $D(y) = \sup\{|\ln f(y-\theta)| : \theta\in\Theta\}$, $q^*(y,\theta,\tau) = \sup\{\ln f(y-\theta') : \rho(\theta,\theta')<\tau\}$, and $q_*(y,\theta,\tau) = \inf\{\ln f(y-\theta') : \rho(\theta,\theta')<\tau\}$, where $\tau>0$.

ASSUMPTION 2: $\sup_n n^{-1}\sum_{t=1}^{n} E\left[D(y_t)^{1+\delta}\right]<\infty$ *for some* $\delta>0$.

ASSUMPTION 3: *For all* $\theta\in\Theta$ *there exists a sequence of positive numbers* $\tau_i = \tau_i(\theta)$, $\tau_i\to 0$, *such that for each* τ_i *the random variables* $q^*(y_t,\theta,\tau_i)$ *and* $q_*(y_t,\theta,\tau_i)$ *satisfy a strong law of large numbers, i.e., as* $n\to\infty$,

$$\begin{aligned} n^{-1}\textstyle\sum_{t=1}^{n}\left[q^*(y_t,\theta,\tau_i)-Eq^*(y_t,\theta,\tau_i)\right] &= o_{a.s.}(1),\\ n^{-1}\textstyle\sum_{t=1}^{n}\left[q_*(y_t,\theta,\tau_i)-Eq_*(y_t,\theta,\tau_i)\right] &= o_{a.s.}(1). \end{aligned}$$

The pointwise SLLN's in Assumption 3 are implied by more primitive conditions we can place on the data, such as that they satisfy certain $\phi-$ or $\alpha-$mixing

conditions. In fact, we shall make mixing assumptions in Section 3 in order to obtain desirable properties for our nonparametric kernel estimates.

These assumptions allow us to state the following lemma, due to Potscher and Prucha (1989, Theorem 2):

LEMMA 1: *Under Assumptions 1-3,* $\int \ln f(y-\theta)dF(y-\theta_0)$ *exists, is finite, is continuous on* Θ, *and*

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{t=1}^{n}[\ln f(y_t-\theta)-E\ln f(y_t-\theta)]\right| = o_{a.s.}(1),$$

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{t=1}^{n}\ln f(y_t-\theta)-\int \ln f(y-\theta)dF(y-\theta_0)\right| = o_{a.s.}(1), \text{ and}$$

$$\left\{n^{-1}\sum_{t=1}^{n}E\ln f(y_t-\theta)\right\} \text{ is equicontinuous on } \Theta.$$

With this Lemma, and with an assumption of identifiability on θ_0, we can prove the consistency of the partial MLE $\widehat{\theta}_n$. We shall employ the following identification criterion, due to Domowitz and White (1982, Definition 2.1):

DEFINITION: *Suppose that* $n^{-1}\sum_{t=1}^{n}E\ln f(y_t-\theta)$ *has a maximum at* θ_0 *for every* n=1,2,.... *Let* $\vartheta_n(\varepsilon)$ *be an open sphere centered at* θ_0 *with fixed radius* $\varepsilon>0$. *For each* n=1,2,..., *define the neighbourhood* $\Xi_n=\vartheta_n(\varepsilon)\cap\Theta$, *such that its complement in* Θ, Ξ_n^c, *is compact. The maximizer* θ_0 *is said to be* identifiably unique *if and only if*

$$\liminf_n\left[\min_{\theta\in\Xi_n^c}\left(n^{-1}\sum_{t=1}^{n}E\ln f(y_t-\theta_0)-n^{-1}\sum_{t=1}^{n}E\ln f(y_t-\theta)\right)\right]>0,$$

for any fixed $\varepsilon>0$.

We then obtain the following consistency result, which follows from Domowitz and White (1982, Theorem 2.2):

THEOREM 1: *Under the conditions of Lemma 1, and assuming that* θ_0 *is the identifiably unique maximizer of* $n^{-1}\sum_{t=1}^{n} E\ln f(y_t-\theta)$ *for every* $n=1,2,...,$ *we have*

$$\widehat{\theta}_n-\theta_0=o_{a.s.}(1).$$

Having established consistency, we must now prove asymptotic normality and justify our claim that the asymptotic variance of $\widehat{\theta}_n$ is the inverse of the information of F. We shall follow the standard method of proof, employing a mean value expansion of $n^{-1}\sum_{t=1}^{n}\psi\left(y_t-\widehat{\theta}_n\right)$ about θ_0, for which purpose we must make:

ASSUMPTION 4: Θ *is convex and* $\theta_0\in int\Theta$.

Our expansion is then

$$n^{-1}\sum_{t=1}^{n}\psi\left(y_t-\widehat{\theta}_n\right)=n^{-1}\sum_{t=1}^{n}\psi\left(y_t-\theta_0\right)+n^{-1}\sum_{t=1}^{n}\psi'\left(y_t-\overline{\theta}_n\right)\left(\widehat{\theta}_n-\theta_0\right),$$

where ψ' is the Hessian and $\overline{\theta}_n\in\left[\widehat{\theta}_n,\theta_0\right]$, from which it follows that

$$n^{1/2}\left(\widehat{\theta}_n-\theta_0\right)=-\left[n^{-1}\sum_{t=1}^{n}\psi'\left(y_t-\overline{\theta}_n\right)\right]^{-1}n^{-1/2}\sum_{t=1}^{n}\psi\left(y_t-\theta_0\right). \tag{4}$$

In analyzing the limiting behaviour of the first term on the right-hand side of (4), we again apply the ULLN of Potscher and Prucha (1989), making the following assumption:

ASSUMPTION 5: *Assumptions 2 and 3 hold, with "*$\ln f(y-\theta)$*" replaced throughout by "*$\psi'(y-\theta)$*".*

As in Lemma 1, we can then show that

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{t=1}^{n}\psi'(y_t-\theta)-\int\psi'(y-\theta)dF(y-\theta_0)\right|=o_{a.s.}(1),$$

from which it follows, using the strong consistency of $\widehat{\theta}_n$ and Theorem 2.3 of Domowitz and White (1982), that

$$\left| n^{-1} \sum_{t=1}^{n} \psi' \left(y_t - \overline{\theta}_n \right) - \int \psi'(y - \theta_0) dF(y - \theta_0) \right| = o_{a.s.}(1).$$

But $\int \psi'(y - \theta_0) dF(y - \theta_0)$ is the expectation of the Hessian of f, which is equal to the negative of the information I_f, implying that

$$- \left[n^{-1} \sum_{t=1}^{n} \psi' \left(y_t - \overline{\theta}_n \right) \right]^{-1} = I_f^{-1} + o_{a.s.}(1). \tag{5}$$

To complete our derivation of the asymptotic distribution of $\widehat{\theta}_n$, we now prove that a central limit theorem for martingale difference sequences can be applied to the second term on the right-hand side of (4), for which we first establish that $\{\psi(\varepsilon_t)\}$ are indeed martingale differences. This follows from our symmetry assumptions; since we have assumed that $f(\varepsilon)$ is symmetric about zero, it follows that $\psi(\varepsilon)$ is anti-symmetric, i.e. that $\psi(\varepsilon) = -\psi(-\varepsilon)$. Our martingale difference result then follows from the symmetry of $g_t(\varepsilon | \Omega_{t-1})$, since

$$E[\psi(\varepsilon) | \Omega_{t-1}] = \int \psi(\varepsilon) g_t(\varepsilon | \Omega_{t-1}) d\varepsilon = 0.$$

We shall apply a CLT given by White (1984, Corollary 5.25), for which we require the following assumption:

ASSUMPTION 6: *(a)* $E\left[\psi(\varepsilon_t)^2\right] \neq 0 \; \forall t = 1, ..., n$; *(b)* $E|\psi(\varepsilon_t)|^{2+\delta} < \infty$ *for some* $\delta > 0$ *and all* $t = 1, ..., n$; *(c) Assumptions 2 and 3 hold, with "*$\ln f(y-\theta)$*" replaced throughout by "*$\psi^2(y - \theta)$*".*

Part *(c)* of this assumption allows us to derive the following convergence result, using the ULLN of Potscher and Prucha (1989):

$$\sup_{\theta \in \Theta} \left| n^{-1} \sum_{t=1}^{n} \psi^2(y_t - \theta) - \int \psi^2(y - \theta) dF(y - \theta_0) \right| = o_{a.s.}(1),$$

from which it follows that

$$\left| n^{-1} \sum_{t=1}^{n} \psi^2(y_t - \theta_0) - \int \psi^2(y - \theta_0) dF(y - \theta_0) \right| = o_{a.s.}(1). \tag{6}$$

Combining (6) with parts *(a)* and *(b)* of Assumption 6 and using White (1984, Corollary 5.25), we have:

$$n^{-1/2} \sum_{t=1}^{n} \psi(y_t - \theta_0) \xrightarrow{d} N\left(0, \int \psi^2(y - \theta_0) dF(y - \theta_0)\right). \tag{7}$$

Recall that $\int \psi^2(y - \theta_0) dF(y - \theta_0) = I_f$, so that (4), (5), and (7) yield

THEOREM 2: *Under Assumptions 1-6, we have*

$$n^{1/2}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, I_f^{-1}\right). \tag{8}$$

We have shown that the partial MLE has an asymptotic normal distribution with variance equal to the inverse of the information of F. This is the case whether the data are iid from F or not.

3. ASYMPTOTIC EQUIVALENCE OF THE ADAPTIVE ESTIMATOR AND THE PARTIAL MLE

Stone's (1975) adaptive estimator of θ_0 (as modified by Bickel (1982) and Kreiss (1987a)) is computed by taking some $n^{1/2}$-consistent preliminary estimator θ_n^*, and adjusting it by a single Newton-Raphson iteration in which nonparametric estimates of the score and information of f, both evaluated at θ_n^*, are used. Our objective in this section is to prove that such an estimator is asymptotically equivalent to the partial MLE described in the previous section. To this end, we shall first show that an iterative estimator constructed using the correctly specified score and information of f has the same distribution as the partial MLE. We then show that an adaptive estimator that uses trimmed Gaussian kernel estimates of this score and information is asymptotically equivalent to the iterative estimator, and so to the partial MLE.

Our first step is therefore to derive the asymptotic distribution of the one-step estimator

$$\widehat{\widehat{\theta}}_n = \theta_n^* + n^{-1/2}\left[n^{-1}\sum_{t=1}^{n}\psi\left(y_t - \theta_n^*\right)^2\right]^{-1}\left[n^{-1/2}\sum_{t=1}^{n}\psi\left(y_t - \theta_n^*\right)\right], \tag{9}$$

where $(\theta_n^* - \theta_0) = O_p(n^{-1/2})$. We can then write

$$n^{1/2}\left(\widehat{\widehat{\theta}}_n - \theta_0\right) = n^{1/2}\left(\theta_n^* - \theta_0\right) + \left[n^{-1}\sum_{t=1}^{n}\psi\left(y_t - \theta_n^*\right)^2\right]^{-1}\left[n^{-1/2}\sum_{t=1}^{n}\psi\left(y_t - \theta_n^*\right)\right]$$

$$\begin{aligned} &= n^{1/2}\left(\theta_n^* - \theta_0\right) + \left[n^{-1}\sum_{t=1}^{n}\psi\left(y_t - \theta_n^*\right)^2\right]^{-1} \\ &\quad \cdot\left[n^{-1/2}\sum_{t=1}^{n}\psi\left(y_t - \theta_0\right) + n^{1/2}\left(\theta_n^* - \theta_0\right)n^{-1}\sum_{t=1}^{n}\psi'\left(y_t - \overline{\theta}_n\right)\right] \end{aligned}$$

where $\overline{\theta}_n \in [\theta_n^*, \theta_0]$ and we have used a mean-value expansion of $\sum_{t=1}^{n}\psi\left(y_t - \theta_n^*\right)$ about θ_0. We can use previous arguments to show that

$$n^{-1}\sum_{t=1}^{n}\psi\left(y_t - \theta_n^*\right)^2 = I_f + o_{a.s.}(1)$$

and

$$n^{-1}\sum_{t=1}^{n}\psi'\left(y_t - \overline{\theta}_n\right) = -I_f + o_{a.s.}(1),$$

so that we have

$$n^{1/2}\left(\widehat{\widehat{\theta}}_n - \theta_0\right) = I_f^{-1}n^{-1/2}\sum_{t=1}^{n}\psi\left(y_t - \theta_0\right) + o_p(1).$$

But it then follows that

$$n^{1/2}\left(\widehat{\widehat{\theta}}_n - \theta_0\right) \xrightarrow{d} N\left(0, I_f^{-1}\right), \tag{10}$$

so that $\widehat{\widehat{\theta}}_n$ and $\widehat{\theta}_n$ are asymptotically equivalent. It may be possible to derive (10) under weaker differentiability and boundedness conditions on F through the

use of a discretized, $\sqrt{n}$-consistent preliminary estimator θ_n^{**}. This strategy is standard in the efficient estimation literature and is due to LeCam (1960), but we have not investigated its applicability in the case at hand.

The computation of the one-step estimator $\widehat{\widehat{\theta}}_n$ given in (9) assumes that we know the functional form of the score ψ of the density f of the average asymptotic distribution F. Adaptive estimators are employed when these functional forms are unknown to the investigator. If we assume that the innovations are iid from f, then $\widehat{\widehat{\theta}}_n$ is an asymptotically efficient estimator. It has been shown (e.g., by Beran (1974) and Stone (1975)) that we can obtain adaptive estimators that are asymptotically equivalent to $\widehat{\widehat{\theta}}_n$ under iid assumptions on the data as long as f is symmetric. We describe below the construction of an adaptive estimator and then proceed to show that it is asymptotically equivalent to $\widehat{\widehat{\theta}}_n$, and so possesses the robustness properties of the partial MLE, even when the iid assumption fails.

The basic problem in adaptive estimation is to come up with some consistent preliminary estimates of the score function $\psi(\varepsilon)$ and the information I_f. The standard approach involves the use of residuals from the consistent preliminary estimate θ_n^* of the parametric component of the model to form nonparametric kernel estimates of these quantities. We show below that the adoption of such an approach produces estimates that are "adaptive", in the sense of being asymptotically equivalent to the partial MLE, even if the data are not iid. Our adaptive estimator employs the following Gaussian kernel estimators of the density f and its derivative f' :

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

where

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

is the Gaussian kernel and $\{a_n\}$ is a bandwidth sequence that converges to zero as $n \to \infty$. We define $\widehat{f}_t'(x,\theta)$ as the first derivative of $\widehat{f}_t(x,\theta)$ with respect to x,

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 $c_n \to \infty$, $e_n \to \infty$, $d_n \to 0$, $a_n c_n \to 0$, $e_n a_n^{-3} = o(n)$, and our score estimator is

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

Note that $\widehat{\psi}_t(x,\theta)$ is anti-symmetric in x by construction. We also assume that $c_n = o(n^{\delta/2})$, where $0 < \delta < \infty$ is such that $n^{-1}\sum_{t=1}^n F_t(z) - F(z) = O(n^\delta)$ for every z. Note that in the stationary case this condition on c_n is redundant since $n^{-1}\sum_{t=1}^n F_t = F$. In the case where $\{F_t\}$ consists of a periodically repeating cycle k periods long (such as in the examples of deterministic heteroskedasticity given in the preceding section and in our simulations below), we have $\delta = 1$.

We can now derive an expression for our adaptive estimator:

$$\widetilde{\theta}_n = \theta_n^* + n^{-1/2}\left[n^{-1}\sum_{t=1}^n \widehat{\psi}_t\left(\varepsilon_t\left(\theta_n^*\right),\theta_n^*\right)^2\right]^{-1}\left[n^{-1/2}\sum_{t=1}^n \widehat{\psi}_t\left(\varepsilon_t\left(\theta_n^*\right),\theta_n^*\right)\right]. \quad (11)$$

To show that $\widetilde{\theta}_n$ has the same robustness properties as the partial MLE, we must show that it is asymptotically equivalent to $\widehat{\widehat{\theta}}_n$, i.e., that

$$n^{1/2}\left(\widetilde{\theta}_n - \widehat{\widehat{\theta}}_n\right) = o_p(1), \quad (12)$$

for which it is sufficient to prove that our score and information estimators are consistent, i.e., that

$$n^{-1/2}\sum_{t=1}^n \left(\widehat{\psi}_t\left(\varepsilon_t\left(\theta_n^*\right),\theta_n^*\right) - \psi\left(\varepsilon_t\left(\theta_n^*\right)\right)\right) = o_p(1) \quad (13)$$

and

$$n^{-1}\sum_{t=1}^n \widehat{\psi}_t\left(\varepsilon_t\left(\theta_n^*\right),\theta_n^*\right)^2 = I_f + o_p(1), \quad (14)$$

respectively.

We shall first prove (13), for which purpose we must strengthen our assumptions on the data generating process for the innovations $\{\varepsilon_t\}$.

ASSUMPTION 7: *The innovation process* $\{\varepsilon_t\}$ *is* $\phi-$ or $\alpha-$*mixing.*

In proving (13), we first show that the convergence holds if all quantities are evaluated at θ_0, and then we shall argue that this is sufficient to establish convergence when evaluated at θ_n^*. The following lemma is proved in the Appendix.

LEMMA 2: *Under assumptions 1-7, we have*

$$n^{-1/2}\sum_{t=1}^{n}\left(\widehat{\psi}_t\left(\varepsilon_t\left(\theta_0\right),\theta_0\right)-\psi\left(\varepsilon_t\left(\theta_0\right)\right)\right)=o_p(1) \tag{15}$$

in $P_{\theta_0,n}$.

REMARK: We have used a trimmed, leave-one-out Gaussian kernel approach to obtain $\widehat{\psi}$. An alternative estimator, due to Schick (1987), employs a logistic kernel and requires neither trimming nor the omission of an observation. The estimates of the density and its derivative are

$$\widehat{f}(x,\theta)=a_n+(na_n)^{-1}\sum_{j=1}^{n}k\left(\frac{x-\varepsilon_j(\theta)}{a_n}\right)$$

and

$$\widehat{f}'(x,\theta)=n^{-1}a_n^{-2}\sum_{j=1}^{n}k'\left(\frac{x-\varepsilon_j(\theta)}{a_n}\right),$$

respectively, where $k(x)=e^{-x}(1+e^{-x})^{-2}$ and $\{a_n\}$ is a sequence of positive numbers such that

$$a_n\to 0 \text{ and } na_n^6\to\infty.$$

The score estimator is:

$$\widehat{\psi}(x,\theta)=\frac{1}{2}\left(\frac{\widehat{f}'(x,\theta)}{\widehat{f}(x,\theta)}-\frac{\widehat{f}'(-x,\theta)}{\widehat{f}(-x,\theta)}\right).$$

This estimator can be shown to satisfy (15) (at least under stationarity assumptions on ε) and so seems preferable in practice to the Gaussian kernel estimator. However, we have found in Monte Carlo experiments that its performance is significantly inferior to that of the Gaussian approach. Consequently, we recommend use of the latter based on our experience and indeed use it ourselves in the Monte Carlo study reported below.

Lemma 2 establishes that our nonparametric score estimator consistently estimates the true score when both quantities are evaluated at the true parameter value θ_0. However, recall from (13) that we seek to prove consistency of the score estimator when evaluated at the preliminary estimate θ_n^*. Now, it follows from Lemma 2 that

$$n^{-1/2}\sum_{t=1}^{n}\left(\widehat{\psi}\left(\varepsilon_t\left(\theta_n^*\right),\theta_n^*\right)-\psi\left(\theta_n^*\right)\right)=o_p(1)$$

in $\left\{P_{\theta_n^*,n}\right\}$, the sequence of probability measures associated with $\{\theta_n^*\}$. It immediately follows that (13) holds, by a property of contiguity of probability measures, due to the following lemma:

LEMMA 3: *The sequences of probability measures* $\left\{P_{\theta_n^*,n}\right\}$ *and* $\{P_{\theta_0,n}\}$ *are contiguous.*

We omit the proof, which is a straightforward application of Theorem 11 in Jeganathan (1995).

The consistency of our information matrix, i.e. (14), then follows by standard arguments. We can use a WLLN to show that

$$n^{-1}\sum_{t=1}^{n}\psi\left(\varepsilon_t\left(\theta_n^*\right)\right)^2=I_f+o_p(1)$$

in $P_{\theta_n^*}$, and so, by contiguity, in $P_{\theta_0,n}$. We then obtain (14) as a consequence of (13). This completes our argument that the adaptive estimator is asymptotically equivalent to the partial MLE.

4. OPTIMALITY THEORY

The results derived above highlight the impressive robustness properties of adaptive estimators. They are designed to be *optimal* when the innovations are iid draws from an unknown density, but also turn out to be *robust* in an important sense even when the innovations are not iid. In this section, we are interested in investigating the potential optimality properties that adaptive estimators may possess in the latter situation. They will obviously not be fully efficient, since the attainment of full efficiency would require a correct specification of the dependence and heterogeneity present in the data (the information matrix would not be block diagonal between θ and a nuisance parameter characterizing the structure of the dependence and heterogeneity; see below). However, we may enquire as to whether the adaptive estimator is optimal according to some less stringent criterion than full efficiency, one that would be more relevant to the case where we are uncertain regarding the true nature of the dependence and heterogeneity present.

We show in the first subsection that the adaptive estimator is optimal in a class of symmetric M-estimators but show in the second subsection that it does not achieve the semiparametric efficiency bound for estimation of θ when the dependence and heterogeneity in $\{\varepsilon_t\}$ is treated as an unknown infinite-dimensional nuisance parameter. We briefly discuss the computation and potential feasibility of estimators that would achieve this bound.

4.1. Symmetric M-Estimators

It is not hard to see that our partial MLE $\widehat{\theta}_n$ is optimal in the class of M-estimators that maximize symmetric, twice continuously differentiable criterion functions. Consider the M-estimator θ_n^* defined as follows:

$$\theta_n^* = \arg\max_{\theta\in\Theta} n^{-1}\sum_{t=1}^{n}\rho\left(y_t-\theta\right),$$

where $\rho\left(\cdot\right)$ is symmetric and twice continuously differentiable. This estimator also satisfies the equation

$$n^{-1}\sum_{t=1}^{n}\varphi\left(y_t-\theta_n^*\right)=0,$$

where $\varphi = \rho'$. Using earlier arguments we can show that

$$n^{1/2}(\theta_n^* - \theta_0) \xrightarrow{d} N(0, V_\rho),$$

where

$$V_\rho = \left[\int \varphi'(y - \theta_0)\, dF(y - \theta_0)\right]^{-2} \left[\int \varphi^2(y - \theta_0)\, dF(y - \theta_0)\right].$$

The partial MLE $\widehat{\theta}_n$ is optimal among all such θ_n^* because $V_\rho \geq I_f^{-1}$ for all ρ and f. This inequality holds because the asymptotic distributions of θ_n^* and $\widehat{\theta}_n$ hold independently of whether the data are iid or not. But if the data are iid, then $\widehat{\theta}_n$ is fully efficient.

4.2. Semiparametric Efficiency Bounds

Econometricians have been paying increasing attention in recent years to the question of constrained efficient estimation in semiparametric models (see Newey (1990) and the already very well-known book by Bickel, Klaassen, Ritov, and Wellner (1993)). Thinking on this subject goes back to Stein (1956), who enquired as to the conditions under which it is possible to estimate a parameter of interest as well, asymptotically, when an infinite-dimensional nuisance parameter is unknown, as it is when the latter is known. In other words, he sought conditions under which *adaptive* estimation would be possible. It turns out that a necessary condition for adaptive estimation is, roughly speaking, that the score function with respect to the parameter of interest be orthogonal to the space spanned by all score functions taken with respect to some finite parameter that correctly characterizes the component of the model that is unknown to the investigator and is hence being treated nonparametrically. This space is termed the *tangent space*. The asymptotic covariance matrix of the adaptive estimator is then equal to the expected outer product of the score with respect to the parameter of interest.

In cases where this orthogonality of scores does not hold and so adaptive, fully efficient estimation is not possible, the question arises as to whether some type of constrained efficiency is possible. The fact remains that we must estimate a parameter of interest in the presence of an unknown nuisance parameter, and so we wish to determine the best that we can possibly do conditional on our ignorance. It turns out that semiparametric efficiency bounds often can be computed for such models. In sufficiently regular cases, the lower bound on the asymptotic

covariance matrix attainable by semiparametric estimators of the parameter of interest is given by the expected outer product of the *efficient score*. The efficient score is that part of the score of the parameters of interest that is orthogonal to the tangent space. The identity of these two scores is Stein's (1956) necessary condition for adaptation. The efficient score is equal to the score of the parameters of interest subtract its orthogonal projection onto the tangent space.

We now briefly sketch the argument for why the partial MLE $\widehat{\theta}_n$ does *not* achieve the semiparametric efficiency bound for our model[2]. We then discuss the reasons why not, and describe possible estimators that would achieve it. We carry out our argument for the special case where $\{\varepsilon_t\}$ is known to be stationary and follow an m^{th}-order Markov process. A similar argument can presumably be made for the more general model.

Under our stationary Markov assumption, we can write the density of ε_t conditional on the past as

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

Suppose we can imagine some parametric specification of the conditional density that contains the full correct parametric specification. Such a specification is referred to in the literature as a *parametric submodel.* We shall write the conditional density for a parametric submodel parameterized by the vector η as $g\left(\varepsilon_t \left| \eta; \varepsilon_{t-1}, ..., \varepsilon_{t-m} \right.\right)$. Now assume that we observe initial conditions $\underline{\varepsilon}_0 = (\varepsilon_{1-m}, ..., \varepsilon_0)$ whose density for this parametric submodel is $f_0\left(\underline{\varepsilon}_0 \left| \eta \right.\right)$. The log-likelihood for a sample of size n is

$$\ln L = \sum_{t=1}^{n} \ln g\left(y_t - \theta \left| \eta; y_{t-1} - \theta, ..., y_{t-m} - \theta \right.\right).$$

Recall our symmetry assumption that

$$g\left(\varepsilon_t \left| \eta; \varepsilon_{t-1}, ..., \varepsilon_{t-m} \right.\right) = g\left(-\varepsilon_t \left| \eta; \varepsilon_{t-1}, ..., \varepsilon_{t-m} \right.\right).$$

Our first step in computing the efficient score for our model is to derive expressions for the scores with repect to the parameter of interest θ and the nuisance

[2]A fully general and rigorous analysis of semiparametric efficiency bounds for our model and estimators that would achieve them is beyond the scope of the present paper and forms the basis for the next one (Hodgson (1996)).

parameter η. These scores can be written respectively as

$$S_\theta = -\sum_{t=1}^{n} \frac{g_\varepsilon}{g}\left(\varepsilon_t \left| \eta; \varepsilon^{t-m}\right.\right) - \sum_{t=1}^{n}\sum_{j=1}^{m} \frac{g_j}{g}\left(\varepsilon_t \left| \eta; \varepsilon^{t-m}\right.\right)$$

$$= -\sum_{t=1}^{n} s_\varepsilon\left(\varepsilon_t\right) - \sum_{t=1}^{n}\sum_{j=1}^{m} s_j\left(\varepsilon_t\right)$$

and

$$\begin{aligned} S_\eta &= \sum_{t=1}^{n} \frac{g_\eta}{g}\left(\varepsilon_t \left| \eta; \varepsilon^{t-m}\right.\right) \\ &= \sum_{t=1}^{n} s_\eta\left(\varepsilon_t\right), \end{aligned}$$

where $\varepsilon^{t-m} = \varepsilon_{t-1}, ..., \varepsilon_{t-m}$, $g_\varepsilon = \partial g/\partial \varepsilon_t$, $g_j = \partial g/\partial \varepsilon_{t-j}$ $\forall j = 1, ..., m$, $g_\eta = \partial g/\partial \eta$, $s_\varepsilon\left(\varepsilon_t\right) = \frac{g_\varepsilon}{g}\left(\varepsilon_t \left| \eta; \varepsilon^{t-m}\right.\right)$, $s_j\left(\varepsilon_t\right) = \frac{g_j}{g}\left(\varepsilon_t \left| \eta; \varepsilon^{t-m}\right.\right)$ $\forall j = 1, ..., m$, and $s_\eta\left(\varepsilon_t\right) = \frac{g_\eta}{g}\left(\varepsilon_t \left| \eta; \varepsilon^{t-m}\right.\right)$. Note that our symmetry assumption implies that $s_\varepsilon\left(\varepsilon\right) = -s_\varepsilon\left(-\varepsilon\right)$, $s_j\left(\varepsilon\right) = s_j\left(-\varepsilon\right)$ $\forall j = 1, ..., m$, and $s_\eta\left(\varepsilon\right) = s_\eta\left(-\varepsilon\right)$.

We follow Newey (1990) in deriving the tangent set $\mathcal{T}_n$ for our model, which can be roughly defined as the space spanned by scores of the form S_η for all parametric submodels. Before stating the tangent set for our model, we make a few observations about S_η. In addition to their symmetry property, we can show that the scores of the conditional densities for the individual observations, $s_\eta\left(\varepsilon_t\right)$, satisfy the zero mean property of score functions, so that $E\left[s_\eta\left(\varepsilon_t\right)\left|\varepsilon^{t-m}\right.\right] = 0$. It obviously follows that $E\left[s_\eta\left(\varepsilon_t\right) s_\eta\left(\varepsilon_{t-j}\right)\right] = 0\ \forall j \neq 0$. We can use these properties to derive the following representation of the tangent set:

$$\mathcal{T}_n = \left\{ D\left(\underline{\varepsilon}_0, \varepsilon_1, ..., \varepsilon_n\right) : D\left(\cdot\right) = \sum_{t=1}^{n} d\left(\varepsilon_t \left| \varepsilon^{t-m}\right.\right),\ d\left(\varepsilon_t \left| \varepsilon^{t-m}\right.\right) = d\left(-\varepsilon_t \left| \varepsilon^{t-m}\right.\right)\ \forall t,\right.$$

$$\left. E\left[d\left(\varepsilon_t \left| \varepsilon^{t-m}\right.\right) \left| \varepsilon^{t-m}\right.\right] = 0\ \forall t \right\}.$$

The efficient score is the residual from the orthogonal projection of S_θ 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_θ, viz. $-\sum_{t=1}^{n} s_\varepsilon\left(\varepsilon_t\right)$, is orthogonal to the tangent set, and that the second term, $-\sum_{t=1}^{n}\sum_{j=1}^{m} s_j\left(\varepsilon_t\right)$, actually belongs to the tangent set. Hence, the orthogonal projection of S_θ 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 to denote the efficient score, we obtain the result that

$$S = -\sum_{t=1}^{n} s_{\varepsilon}(\varepsilon_t).$$

The semiparametric asymptotic efficiency bound is

$$\begin{aligned} B &= p\lim n^{-1} E\left[S^2\right]^{-1} \\ &= E\left[s_{\varepsilon}^2(\varepsilon_t)\right]^{-1}, \end{aligned}$$

which is the inverse of the unconditional expectation of the square of the score of the conditional density $g(\varepsilon_t|\varepsilon^{t-m})$. The question arises as to whether this bound is equal to the asymptotic variance of our partial MLE, I_f^{-1}, which is the inverse of the unconditional expectation of the square of the score of the *unconditional* density $f(\varepsilon_t)$. It is easy to show that these two quantities are unequal, so that the partial MLE (and hence the "adaptive" estimator) is not semiparametrically efficient. We shall discuss the reason for this inefficiency and suggest possible estimation strategies to attain B below, but first we analyze B more carefully and show that it is not equal to I_f^{-1}. To this end, we introduce some new notation. Let $\widetilde{f}(\varepsilon_t, \varepsilon^{t-m})$ denote the joint density of ε_t and ε^{t-m}, $\widetilde{f}_m(\varepsilon^{t-m})$ denote the marginal density of ε^{t-m}, and $f(\varepsilon_t)$ denote the marginal of ε_t, as before. Since $g(\varepsilon_t|\varepsilon^{t-m}) = \widetilde{f}(\varepsilon_t, \varepsilon^{t-m})/\widetilde{f}_m(\varepsilon^{t-m})$, it follows that $\partial g(\varepsilon_t|\varepsilon^{t-m})/\partial\varepsilon_t = \left(\partial\widetilde{f}(\varepsilon_t, \varepsilon^{t-m})/\partial\varepsilon_t\right)/\widetilde{f}_m(\varepsilon^{t-m})$ and hence that $s_{\varepsilon}(\varepsilon_t) = \frac{\partial g(\varepsilon_t|\varepsilon^{t-m})/\partial\varepsilon_t}{g(\varepsilon_t|\varepsilon^{t-m})} = \frac{\partial\widetilde{f}(\varepsilon_t,\varepsilon^{t-m})/\partial\varepsilon_t}{\widetilde{f}(\varepsilon_t,\varepsilon^{t-m})}$. We then have the following expression for the semiparametric efficiency bound:

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

In words, B is equal to the inverse of the first element on the diagonal of the information matrix of the joint density $\widetilde{f}(\varepsilon_t, \varepsilon^{t-m})$. This is not the same as I_f^{-1},

which is the inverse of the information of the marginal density of ε_t, because the diagonal elements of the information matrix of a multivariate density are not the same as the informations of the respective marginal univariate densities. (Think of the multivariate Gaussian density. The information is the inverse of the covariance matrix, but we know that the reciprocals of the diagonal elements of a matrix generally only equal the diagonal elements of the inverse of the matrix if the matrix is diagonal.)

Our finding, that the semiparametric efficiency bound is equal to the asymptotic variance of an estimator the maximizes the sum of the *conditional* scores of the sample, rather than the *unconditional* scores (as in the case of the partial MLE), is actually quite intuitive. The reason is that these unconditional scores are in principal nonparametrically consistently estimable. The sample provides us with information regarding the distribution of ε_t conditional on the past, even if we have no knowledge of the parametric structure of this conditional dependence, and it seems very plausible that the partial MLE, which ignores this information, can be improved upon. Our finding that the adaptive estimator is asymptotically equivalent to the partial MLE hinges on our proof that the score function of the unconditional density $f(\varepsilon)$ can be consistently nonparametrically estimated. But there is no reason why we cannot also consistently nonparametrically estimate the score of the conditional density $g(\varepsilon_t | \varepsilon^{t-m})$. In fact, the estimation of such a score is carried out by Jeganathan (1995) in a different context. The only practical limitation on such a procedure would be the curse of dimensionality problem which would limit the empirical feasibility of such an approach to cases where m is small. Further investigation of this point shall be carried out elsewhere (Hodgson (1996)).

5. EXAMPLES AND SIMULATIONS

In this section we illustrate the results derived above and evaluate the finite sample performance of the adaptive estimator for three examples of stationary, uncorrelated stochastic processes with second-moment dependence, viz., the ARCH, threshold, and Markov switching models. We report Monte Carlo simulation results for these three models, as well as for iid and deterministically heteroskedastic models. We begin by introducing the three models of conditional heteroskedasticity, then we describe the simulation experiment, in which we compare the finite-sample behaviour of the sample mean and adaptive estimator.

5.1. 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 do not know the unconditional distribution of the ARCH(1) process, so a derivation of the information of this distribution, and hence of the asymptotic variance of the adaptive estimator and its efficiency gain over the sample mean, is not possible. However, our Monte Carlo results do illustrate the gains possible for particular parameter settings and sample sizes.

5.2. 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. Each point at which a jump occurs is a threshold value.[3] 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.

[3] Since the ARCH conditional variance is a continuous function of the lagged absolute value of the process, we conjecture that an ARCH model can be arbitrarily well approximated by a threshold model as the set of threshold values becomes very fine. This fact may allow us to approximate the information of an ARCH process arbitrarily well since we know the unconditional distribution, and hence the information, of a threshold model. See below.

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.

5.3. 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.[4]

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

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:

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

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 $\gamma = (1-q)/(2-p-q)$.

5.4. Simulation Results

The results of our simulation study comparing the finite sample performances of the sample mean and the adaptive estimator for the models described above are presented in Tables 1-4, while Table 5 reports similar results for a deterministically heteroskedastic data generating process and Table 6 for iid data. In all cases we report bias and mean squared error (MSE) statistics for the sample mean and the adaptive estimator (the latter for three different bandwidth settings), as well as computing the ratio between the MSE's of the two estimators. We generated data sets of length 100 and 250 (with an additional startup observation for the ARCH, Markov, and threshold data sets), and in all cases used 1000 iterations. The parameters of each data generating process were selected such that the unconditional variance was three.

Table 1: Simulation Results for ARCH(1) Model: MSE and Bias of Adaptive Estimator (n=100)

$\alpha_0=0.3 \quad \alpha_1=0.9$			
Sample Mean: MSE = 2.80×10^{-2} Bias = 4.97×10^{-3}			
Bandwidth	MSE(Adaptive)	Bias(Adaptive)	MSE(Ad)/MSE(SM)
0.65	1.82×10^{-2}	-7.19×10^{-4}	.65
0.73	1.74×10^{-2}	-6.98×10^{-4}	.62
0.80	1.68×10^{-2}	-1.21×10^{-3}	.60
$\alpha_0=0.6 \quad \alpha_1=0.8$			
Sample Mean: MSE = 3.13×10^{-2} Bias = 4.64×10^{-3}			
0.65	2.22×10^{-2}	-8.60×10^{-4}	.71
0.73	2.08×10^{-2}	-1.01×10^{-3}	.66
0.80	2.01×10^{-2}	-8.56×10^{-4}	.64
$\alpha_0=0.9 \quad \alpha_1=0.7$			
Sample Mean: MSE = 3.01×10^{-2} Bias = 2.94×10^{-3}			
0.65	2.63×10^{-2}	-1.30×10^{-3}	.87
0.73	2.52×10^{-2}	-1.53×10^{-3}	.84
0.80	2.46×10^{-2}	-1.51×10^{-3}	.82

Notes: (a) For each parameter setting, 1000 iterations were carried out.
(b) The Silverman (1986) rule-of-thumb bandwidth is 0.73.

Table 2: Simulation Results for ARCH(1) Model: MSE and Bias of Adaptive Estimator (n=250)

$\alpha_0=0.3 \quad \alpha_1=0.9$			
Sample Mean: MSE = 9.74×10^{-3} Bias = -3.41×10^{-3}			
Bandwidth	MSE(Adaptive)	Bias(Adaptive)	MSE(Ad)/MSE(SM)
0.50	4.58×10^{-3}	-2.13×10^{-3}	.47
0.60	4.29×10^{-3}	-2.39×10^{-3}	.44
0.70	4.33×10^{-3}	-1.86×10^{-3}	.44
$\alpha_0=0.6 \quad \alpha_1=0.8$			
Sample Mean: MSE = 1.18×10^{-2} Bias = -4.21×10^{-3}			
0.50	8.22×10^{-3}	-2.86×10^{-3}	.70
0.60	8.02×10^{-3}	-3.26×10^{-3}	.68
0.70	7.59×10^{-3}	-2.85×10^{-3}	.64
$\alpha_0=0.9 \quad \alpha_1=0.7$			
Sample Mean: MSE = 1.22×10^{-2} Bias = -4.41×10^{-3}			
0.50	1.01×10^{-2}	-3.91×10^{-3}	.83
0.60	9.70×10^{-3}	-3.39×10^{-3}	.80
0.70	9.44×10^{-3}	-3.31×10^{-3}	.77

Notes: (a) For each parameter setting, 1000 iterations were carried out.
(b) The Silverman (1986) rule-of-thumb bandwidth is 0.61.

Table 3: Simulation Results for Threshold Model: MSE and Bias of Adaptive Estimator

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

n=100			
Sample Mean: MSE = 2.92×10^{-2} Bias = -4.87×10^{-4}			
Bandwidth	MSE(Adaptive)	Bias(Adaptive)	MSE(Ad)/MSE(SM)
0.65	1.14×10^{-2}	-1.55×10^{-3}	.39
0.73	1.30×10^{-2}	-5.55×10^{-4}	.45
0.80	1.45×10^{-2}	-1.65×10^{-3}	.50
n=250			
Sample Mean: MSE = 1.21×10^{-2} Bias = -3.19×10^{-3}			
0.50	2.98×10^{-3}	-6.27×10^{-4}	.25
0.60	4.33×10^{-3}	-2.89×10^{-4}	.36
0.70	6.72×10^{-3}	-5.30×10^{-4}	.56

Notes: (a) For each parameter setting, 1000 iterations were carried out.
(b) The Silverman (1986) rule-of-thumb bandwidths are 0.73 and 0.61, respectively.

Table 4: Simulation Results for Switching Model: MSE and Bias of Adaptive Estimator

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

n=100			
Sample Mean: MSE = 2.98×10^{-2} Bias = 1.34×10^{-3}			
Bandwidth	MSE(Adaptive)	Bias(Adaptive)	MSE(Ad)/MSE(SM)
0.65	1.26×10^{-2}	-1.53×10^{-3}	.42
0.73	1.55×10^{-2}	-8.81×10^{-4}	.52
0.80	1.92×10^{-2}	-8.99×10^{-4}	.64
n=250			
Sample Mean: MSE = 1.18×10^{-2} Bias = -2.89×10^{-3}			
0.50	3.31×10^{-3}	-4.95×10^{-4}	.28
0.60	4.60×10^{-3}	-1.15×10^{-3}	.39
0.70	6.98×10^{-3}	-2.02×10^{-3}	.59

Notes: (a) For each parameter setting, 1000 iterations were carried out.
(b) The Silverman (1986) rule-of-thumb bandwidths are 0.73 and 0.61, respectively.

Table 5: Simulation Results for Deterministic Heteroskedasticity: MSE and Bias of Adaptive Estimator

n=100			
Sample Mean: MSE = 2.91×10^{-2} Bias = -3.31×10^{-3}			
Bandwidth	MSE(Adaptive)	Bias(Adaptive)	MSE(Ad)/MSE(SM)
0.65	1.30×10^{-2}	-1.13×10^{-3}	.45
0.73	1.57×10^{-2}	-1.41×10^{-3}	.54
0.80	1.92×10^{-2}	-1.90×10^{-3}	.66
n=250			
Sample Mean: MSE = 1.19×10^{-2} Bias = -5.83×10^{-2}			
0.50	3.23×10^{-3}	7.22×10^{-4}	.27
0.60	4.67×10^{-3}	2.72×10^{-3}	.39
0.70	7.38×10^{-3}	2.14×10^{-3}	.62

Notes: (a) For each parameter setting, 1000 iterations were carried out.
(b) The Silverman (1986) rule-of-thumb bandwidths are 0.73 and 0.61, respectively.
(c) All observations are N(0,1/3), except for every tenth observation, which is N(0,27).

Table 6: Simulation Results for i.i.d. Data: MSE and Bias of Adaptive Estimator

n=100			
Sample Mean: MSE = 3.13×10^{-2} Bias = 2.77×10^{-3}			
Bandwidth	MSE(Adaptive)	Bias(Adaptive)	MSE(Ad)/MSE(SM)
0.65	1.36×10^{-2}	-1.54×10^{-3}	.43
0.73	1.54×10^{-2}	-2.68×10^{-3}	.49
0.80	1.82×10^{-2}	-3.26×10^{-3}	.58
n=250			
Sample Mean: MSE = 1.23×10^{-2} Bias = -2.19×10^{-3}			
0.50	3.27×10^{-3}	-3.20×10^{-3}	.27
0.60	4.73×10^{-3}	-2.29×10^{-3}	.38
0.70	7.23×10^{-3}	-3.29×10^{-3}	.59

Notes: (a) For each parameter setting, 1000 iterations were carried out.
(b) The Silverman (1986) rule-of-thumb bandwidths are 0.73 and 0.61, respectively.
(c) The data are iid 0.9N(0,1/3)+0.1N(0,27).

In computing the adaptive estimators, we employed trimming parameter settings of $d_n = \exp(-112)$ and $e_n = c_n = 15$ in all cases. Each parameter therefore trims at approximately eight standard deviations from the origin. We did not experiment with varying these parameters, as previous simulation studies (Hsieh and Manski (1987) and Hodgson (1995c)) find that the performance of adaptive estimators is less sensitive to such variation than to variation in bandwidths. We evaluate this latter sensitivity by employing three bandwidth settings for each data set, with the median bandwidth approximately equalling Silverman's (1986) plug-in bandwidth for density estimation problems. We acknowledge that this is a very crude and suboptimal approach to bandwidth selection, but our results illustrate the point that adaptive estimators, even when suboptimally implemented, still perform very well.

We consider three parameter settings for the ARCH process, viz. $(\alpha_0, \alpha_1) =$ (0.3,0.9), (0.6,0.8), and (0.9,0.7). We chose the values of α_1 fairly close to unity to accord with the stylized facts of economic data and because it is in this range, as emphasized by Engle (1982), that the sample mean becomes highly inefficient. The values of α_0 were chosen to obtain an unconditional variance of three, as mentioned above. The simulation results for the ARCH models, reported in Tables 1 and 2, indicate the finite sample efficiency gains obtainable by the adaptive estimator. We cannot tell how closely these gains come to the asymptotic efficiency gains because, as discussed above, we do not know the latter in the absence of knowledge of the unconditional distribution of the ARCH process. However, we *can* obtain a fairly close numerical approximation to the efficiency gain obtainable over the sample mean by the fully efficient MLE of the correctly specified ARCH model, using an analytical result of Engle (1982, p. 999). For our parameter settings, the respective maximum possible efficiency gains are approximately 0.13, 0.25, and 0.37. Using the Silverman bandwidth, the adaptive estimator's efficiency gains for the respective ARCH parameter settings are 0.62, 0.66, and 0.84 when n=100, and 0.44, 0.68, and 0.80 when n=250. These gains are quite respectable. It would be interesting to compare them with the finite sample gains obtained by the MLE's of correctly and incorrectly specified ARCH models.

The parameter values of each of the four remaining models have been chosen such that the unconditional distribution (or, in the case of the nonstationary deterministically heteroskedastic model, the "average" asymptotic distribution)

is the following variance-contaminated mixture of normals:

$$f(\varepsilon) = 0.9N(0, 1/3) + 0.1N(0, 27).$$

In both the threshold and switching models, we set $\sigma_A^2 = 1/3$ and $\sigma_B^2 = 27$. We set $\alpha = 1.32$ in the threshold model and $(p, q) = (0.92, 0.30)$ in the switching model. The deterministically heteroskedastic model is implemented by having $\varepsilon_t \sim N(0, 1/3)$ except when t is a multiple of ten, in which case $\varepsilon_t \sim N(0, 27)$. The unconditional distribution $f(\varepsilon)$ is also employed in simulations by Hodgson (1995c) and Hsieh and Manski (1987), the latter authors scaling down the variances to produce a distribution with unit variance. In all four cases, the asymptotic efficiency ratio of the partial MLE over the sample mean is the same, and equals

$$\left(\sigma_\varepsilon^2 I_f\right)^{-1} \cong 0.13.$$

We would therefore expect similar numbers in each of the Tables 3-6, which we do obtain. The efficiency ratio is in the .4-.65 range, depending on bandwidth choice, when the sample size is 100, while for samples of 250, it is in the .25-.6 range. We can see that for good bandwidth choices, the adaptive estimator is coming quite close to its asymptotic effciency ratio for small to moderate samples, and is still improving substantially upon the sample mean even for poor bandwidth choices.

6. EXTENSIONS TO OTHER MODELS

Our main result, that the adaptive estimator is asymptotically normal with variance equal to the inverse of the information of the unconditional distribution, can be extended to models more general than the location case. The key point to recognize is that our result relies on the fact that the score sequence of the partial likelihood (the "partial score" sequence) is a martingale difference sequence, a fact implied by our assumption of conditional symmetry. Now consider a general time series model with innovations $\{\varepsilon_t\}$ that are assumed to be iid but are conditionally or unconditionally heterogeneous as described above. Then, under assumptions implying that the partial scores of this model are martingale differences, it will be a straightforward extension of the results derived above to show that the partial MLE and the adaptive estimator will be asymptotically normal (or mixed normal in a cointegrated model) with covariance matrix equal to the inverse of the

asymptotic information matrix based on the unconditional distribution F of the innovations.

The assumption of conditionally symmetric innovations will imply this martingale difference property for many important models that have been shown to be adaptively estimable. These include Bickel's (1982) linear regression model, Kreiss' (1987a) ARMA model, Steigerwald's (1992) time series regression model, and Linton's (1993) ARCH model[5]. Asymmetry is permissable for the errors in a linear regression model (cf. Bickel (1982)), as long as the regressors are exogenous and are martingale differences when unconditionally demeaned. The result also carries over to multivariate models with innovations that have conditional densities symmetric about zero, including adaptive estimators of multivariate location models, multiple-equation regression models such as SUR and simultaneous equations, and error correction models (Hodgson (1995b)).

The result can presumably also be extended to apply to adaptive estimators in multivariate models in which an elliptical symmetry assumption is employed to reduce the nonparametric kernel density estimation problem to a univariate one, hence averting the curse of dimensionality. Bickel (1982) has applied such an approach to the multivariate location problem and Hodgson, Linton, and Gozalo (1996) have applied it to stationary and cointegrating SUR models. The basic requirement here would seem to be that the unconditional densities are all elliptically symmetric about zero with identical correlation structures.

In models for which the partial score sequence is not a martingale difference sequence, such as the location model with asymmetric conditional densities, our result will not hold. The partial MLE will still be asymptotically normally distributed, but its asymptotic variance will not have the simple interpretation of being the information of an unconditional density. Rather, it will have the "sandwich" structure typical of M-estimators in misspecified models (cf. White (1982) and Levine (1983)). It will have the following asymptotic distribution:

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

where $V_\psi = \sum_{t=1}^{\infty}\sum_{j=1}^{\infty} E\left[\psi\left(\varepsilon_t\right)\psi\left(\varepsilon_j\right)\right]$. If we know ψ, we can estimate V_ψ using one of the standard long-run covariance matrix procedures. If we do not know ψ, we can use the nonparametric score estimates $\widehat{\psi}_t$ from the adaptive estimation

[5]Lee and Hansen (1994) analyze ARCH models in which the innovation, scaled by its conditional variance, may still possess temporal dependence.

procedure in our calculation of a long-run variance estimate. Note that if we are concerned about the presence of asymmetry, we will not want to use the adaptive estimation procedure outlined in Section 3, since it employs a kernel score estimator that is anti-symmetric by construction. Rather, we will want to compute an adaptive estimator like that used by Kreiss (1987b) for autoregressions with asymmetric errors, and which is similar in structure to the estimator of Stone (1975). Following this latter approach, it is possible in applications to gauge the importance of asymmetry to one's results by comparing the estimates of V_ψ and I_f. These estimates could be used in the construction of a formal test analogous to the information matrix test of White (1982). We shall defer a complete analysis of the case of non-martingale difference scores and its application to the various models described above to a later paper.

7. CONCLUSIONS

We have shown analytically and illustrated through simulations that the technique of adaptive estimation, which delivers fully asymptotically efficient estimates in the iid location parameter model, and in many other econometric models that can be reduced to a series of iid innovations, also delivers estimates that are robust to many forms of conditional heterogeneity. We have shown that in a location parameter model with mixing innovations whose conditional densities are symmetric about zero, the adaptive estimator has an asymptotic normal distribution with variance equal to the inverse of the information of the unconditional distribution of the data. Our Monte Carlo results indicate that the finite sample efficiency gains of the adaptive estimator over the sample mean are substantial for many types of non-iid data. The results can be extended to many models of interest in time series econometrics, provided that the partial score sequence is a martingale difference sequence.

8. APPENDIX

PROOF OF LEMMA 2: We shall establish mean square consistency, i.e. that

$$n^{-1}E\left[\sum_{t=1}^{n}\left(\widehat{\psi}_t\left(\varepsilon_t\left(\theta_0\right),\theta_0\right)-\psi\left(\varepsilon_t\left(\theta_0\right)\right)\right)\right]^2\rightarrow 0.$$

Because $\psi(\varepsilon) = -\psi(-\varepsilon)$ and $\widehat{\psi}_t(\varepsilon, \theta_0) = -\widehat{\psi}_t(-\varepsilon, \theta_0)$ for every t, it follows that $\left\{\widehat{\psi}_t(\varepsilon_t(\theta_0), \theta_0) - \psi(\varepsilon_t(\theta_0))\right\}$ is a martingale difference sequence, so our problem reduces to showing

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

which will follow from our proof that

$$E\left[\widehat{\psi}_t(\varepsilon_t(\theta_0), \theta_0) - \psi(\varepsilon_t(\theta_0))\right]^2 \to 0 \quad \forall t = 1, ..., n, \tag{16}$$

i.e., that

$$\int \left\{q_t(\varepsilon) - \frac{f'}{f}(\varepsilon)\right\}^2 f(\varepsilon)d\varepsilon \to 0, \tag{17}$$

where we have simplified notation by writing $q_t(\varepsilon, \theta_0) = q_t(\varepsilon)$.

Our proof of (17) is similar to the proof of Lemma 4.1 in Bickel (1982) and proceeds in three steps, the first of which is to prove

$$\int \left\{q_t(\varepsilon) - \frac{\overline{f}'_n}{\overline{f}_n}(\varepsilon)\right\}^2 \overline{f}_n(\varepsilon)d\varepsilon \to 0, \tag{18}$$

where $\overline{f}_n = n^{-1}\sum_{t=1}^{n} f_t(\varepsilon)$. But to prove (18), we must verify that

$$\int \left\{q_t(\varepsilon) - \frac{\overline{f}'_{na}}{\overline{f}_{na}}(\varepsilon)\right\}^2 \overline{f}_{na}(\varepsilon)d\varepsilon \to 0, \tag{19}$$

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

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

Denoting by $g^{(\nu)}$ the ν^{th} partial derivative of a function g, for $\nu = 0, 1$, we have

$$E\left[\widehat{f}_t^{(\nu)}(x)\right] = (n-1)^{-1}\sum_{\substack{j=1 \\ j\neq t}}^{n} E\left[\pi_{a_n}^{(\nu)}(x - \varepsilon_j)\right]$$

$$= (n-1)^{-1} \sum_{\substack{j=1 \\ j \neq t}}^{n} \int \pi_{a_n}^{(\nu)}(x-y) F_j(dy)$$

$$= (n-1)^{-1} \sum_{\substack{j=1 \\ j \neq t}}^{n} f_{ja}^{(\nu)}(x) \equiv \overline{f}_{ta}^{(\nu)}(x).$$

We are now interested in deriving a bound for the variance of $\widehat{f}_t^{(\nu)}(x)$, $\nu = 0, 1$, which will enable us to apply the argument of Bickel (1982, Lemma 6.1) to prove (19). We have

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

$$= E\left[(n-1)^{-1} \sum_{\substack{j=1 \\ j \neq t}}^{n} \left(\pi_{a_n}^{(\nu)}(x - \varepsilon_j) - f_{ja}^{(\nu)}(x)\right)\right]^2$$

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

where $z_j = \pi_{a_n}^{(\nu)}(x - \varepsilon_j) - f_{ja}^{(\nu)}(x)$. It follows that

$$V_{nt}^{(\nu)}(x) = (n-1)^{-2} E\left[\sum_{j=1}^{n} z_j - z_t\right]^2 \tag{20}$$

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

Now we note that because $\{\varepsilon_t\}$ is mixing, so is $\{z_t\}$. Define $\gamma_{t,\tau} = E[z_t z_{t-\tau}]$ and $\rho_{t,\tau} = \gamma_{t,\tau}/\gamma_{t,0}$. Our mixing assumption implies that $|\rho_{t,\tau}| \to 0$ as $|\tau| \to \infty$ for every t (White and Domowitz (1984)). From (20), we have

$$V_{nt}^{(\nu)}(x) = (n-1)^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_{i,i-j} - 2(n-1)^{-2} \sum_{i=1}^{n} \gamma_{i,i-t} + (n-1)^{-2} \gamma_{t,0}$$

$$= (n-1)^{-2}\sum_{i=1}^{n}\gamma_{i,0}\sum_{j=1}^{n}\rho_{i,i-j} - 2(n-1)^{-2}\sum_{i=1}^{n}\gamma_{i,0}\rho_{i,i-t} + (n-1)^{-2}\gamma_{t,0}$$

$$\leq \left[(n-1)^{-1}\sum_{i=1}^{n}\gamma_{i,0}\right]\left[(n-1)^{-1}\max_{i\in[1,\ldots,n]}\sum_{j=1}^{n}|\rho_{i,i-j}|\right.$$

$$\left.-2\,(n-1)^{-1}\max_{i\in[1,\ldots,n]}|\rho_{i,i-t}| + (n-1)^{-1}\frac{\gamma_{t,0}}{\sum_{i=1}^{n}\gamma_{i,0}}\right]$$

$$\leq (n-1)^{-1}\left[(n-1)^{-1}\sum_{i=1}^{n}\gamma_{i,0}\right]M$$

since there exists some $0 < M < \infty$ such that $\max_{i\in[1,\ldots,n]}\sum_{j=1}^{n}|\rho_{i,i-j}| - 2\max_{i\in[1,\ldots,n]}|\rho_{i,i-t}| + \frac{\gamma_{t,0}}{\sum_{i=1}^{n}\gamma_{i,0}} < M$ for every n. It follows that

$$V_{nt}^{(\nu)}(x) \leq \frac{M}{n-1}\left[(n-1)^{-1}\sum_{i=1}^{n}\gamma_{i,0}\right]$$

$$\leq \frac{M}{n-1}\left[(n-1)^{-1}\sum_{i=1}^{n}E\left[\pi_{a_n}^{(\nu)}(x-\varepsilon_i)\right]^2\right].$$

Now,

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

for some constant κ_ν (see Stone (1975)), so

$$\frac{M}{n-1}\left[(n-1)^{-1}\sum_{i=1}^{n}E\left[\pi_{a_n}^{(\nu)}(x-\varepsilon_i)\right]^2\right]$$

$$\leq \frac{M\kappa_\nu}{(n-1)a_n^{2\nu+1}}\left[(n-1)^{-1}\sum_{i=1}^{n}E\left[\pi_{a_n}(x-\varepsilon_i)\right]\right] = \frac{M\kappa_\nu}{(n-1)a_n^{2\nu+1}}\overline{f}_{na}(x),$$

yielding

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

where $\tau_\nu = M\kappa_\nu$. The proof of (19) now follows the same lines as that of Lemma 6.1 in Bickel (1982). To complete our proof of (18), we can use Lemmas 6.2 and

6.3 of Bickel (1982) to obtain the following two convergence results:

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

and

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

This establishes (18) and completes the first step of our proof of (17). The second step is to show

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

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

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

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

■

References

[1] Beran, R. 1974. Asymptotically efficient rank estimates in location models. *Annals of Statistics* 2:63-74.

[2] Bickel, P.J. 1975. On adaptive estimation. *Annals of Statistics* 10:647-671.

[3] 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.

[4] Billingsley, P. 1968. *Convergence of Probability Measures.* New York; John Wiley & Sons.

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

[6] Domowitz, I. and White, H. 1982. Misspecified models with dependent observations. *Journal of Econometrics* 20:35-58.

[7] Engle, R.F. 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50:987-1007.

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

[9] Hodgson, D.J. 1995a. Adaptive estimation of cointegrating regressions with ARMA errors. RCER Working Paper #408, University of Rochester.

[10] Hodgson, D.J. 1995b. Adaptive estimation of error correction models. RCER Working Paper #410, University of Rochester.

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

[12] Hodgson, D.J. 1996. Semiparametric efficient estimation in Markov models. In progress.

[13] Hodgson, D.J., Linton, O., and Gozalo, P. 1996. Estimation in time series regression models with elliptically symmetric errors. Unpublished manuscript.

[14] Jeganathan, P. 1995. Some aspects of asymptotic theory with applications to time series models. *Econometric Theory* 11:818-887.

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

[16] Kreiss, J.-P. 1987b. On adaptive estimation in autoregressive models when there are nuisance functions. *Statistics & Decisions* 5:59-76.

[17] LeCam, L. 1960. Locally asymptotically normal families of distributions. *University of California Publications in Statistics* 3:37-98.

[18] Lee, S.-W. and Hansen, B.E. 1994. Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory* 10:29-52.

[19] Levine, D. 1983. A remark on serial correlation in maximum likelihood. *Journal of Econometrics* 23:337-342.

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

[21] Manski, C.F. 1984. Adaptive estimation of non-linear regression models. *Econometric Reviews* 3:145-210 (including commentary).

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

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

[24] Schick, A. 1987. A note on the construction of asymptotically linear estimators. *Journal of Statistical Planning and Inference* 16:89-105.

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

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

[27] Stein, C. 1956. Efficient nonparametric testing and estimation. *Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability.*

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

[29] White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50:1-26.

[30] White, H. 1984. *Asymptotic Theory for Econometricians.* Orlando; Academic Press.

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