Econometrica, Vol. 59, No. 3 (May, 1991), 817–858

HETEROSKEDASTICITY AND AUTOCORRELATION CONSISTENT COVARIANCE MATRIX ESTIMATION

BY DONALD W. K. ANDREWS[1]

This paper is concerned with the estimation of covariance matrices in the presence of heteroskedasticity and autocorrelation of unknown forms. Currently available estimators that are designed for this context depend upon the choice of a lag truncation parameter and a weighting scheme. Results in the literature provide a condition on the growth rate of the lag truncation parameter as $T \to \infty$ that is sufficient for consistency. No results are available, however, regarding the choice of lag truncation parameter for a fixed sample size, regarding data-dependent automatic lag truncation parameters, or regarding the choice of weighting scheme. In consequence, available estimators are not entirely operational and the relative merits of the estimators are unknown.

This paper addresses these problems. The asymptotic truncated mean squared errors of estimators in a given class are determined and compared. Asymptotically optimal kernel/weighting scheme and bandwidth/lag truncation parameters are obtained using an asymptotic truncated mean squared error criterion. Using these results, data-dependent automatic bandwidth/lag truncation parameters are introduced. The finite sample properties of the estimators are analyzed via Monte Carlo simulation.

KEYWORDS: Asymptotic mean squared error, autocorrelation, covariance matrix estimator, heteroskedasticity, kernel estimator, spectral density.

1. INTRODUCTION

THIS PAPER CONSIDERS heteroskedasticity and autocorrelation consistent (HAC) estimation of covariance matrices of parameter estimators in linear and nonlinear models. A prime example is the estimation of the covariance matrix of the least squares (LS) estimator in a linear regression model with heteroskedastic, temporally dependent errors of unknown form. Other examples include covariance matrix estimation of LS estimators of nonlinear regression and unit root models and of two and three stage LS and generalized method of moments estimators of nonlinear simultaneous equations models.

The paper has several objectives. The first is to analyze and compare the properties of several HAC estimators that have been proposed in the literature; see Levine (1983), White (1984, pp. 147–161), White and Domowitz (1984), Gallant (1987, pp. 533, 551, 573), Newey and West (1987), and Gallant and White (1988, pp. 97–103). Currently the consistency of such estimators has been established, but their relative merits are unknown.

The second objective is to make existing estimators operational by determining suitable values for the lag truncation or bandwidth parameters that are used

[1] I thank Chris Monahan for excellent research assistance. I also thank two referees, a co-editor, Jean-Marie Dufour, Whitney K. Newey, Peter C. B. Phillips, and Ken Vetzal for helpful comments. Lastly, I thank the California Institute of Technology and the University of British Columbia for their hospitality while part of this research was carried out, and the Alfred P. Sloan Foundation and the National Science Foundation for research support provided through a Research Fellowship and Grant Nos. SES-8618617 and SES-8821021 respectively.

The programming for the Monte Carlo results of Section 9 was done by Chris Monahan. The computations were carried out on the twenty-plus IBM-AT PC's at the Yale University Statistics Laboratory using the GAUSS normal random number generator.

to define the estimators. At present, no guidance is available regarding the choice of these parameters for a given finite sample situation. This is a serious problem, because the performance of these estimators can depend greatly on this choice.

The third objective of the paper is to obtain an optimal estimator out of a class of kernel estimators that contains the HAC estimators that have been proposed in the literature. An optimal estimator, called a *quadratic spectral* (QS) estimator, is obtained using an asymptotic truncated mean squared error (MSE) optimality criterion.

The fourth objective of the paper is to investigate the finite sample performance of kernel HAC estimators. Monte Carlo simulation is used. Different kernels and bandwidth parameters are compared. In addition, kernel estimators are compared with standard parametric covariance matrix estimators.

The class of kernel HAC estimators considered here includes estimators that give some weight to all $T-1$ lags of the sample autocovariance function. Such estimators have not been considered previously. As it turns out, the optimal estimator is of this form.

The consistency of kernel HAC estimators is established under weaker conditions on the growth rate of the lag truncation/bandwidth parameter S_T than is available elsewhere. Instead of requiring $S_T = o(T^{1/4})$ or $O(T^{1/5})$, as in the papers referenced above, or $S_T = o(T^{1/2})$, as in Keener, Kmenta, and Weber (1987) and Kool (1988), we just require $S_T = o(T)$ as $T \to \infty$. Our results also provide rates of convergence of the estimators to the estimand.

To achieve the objectives outlined above, the general approach taken in this paper is to exploit existing results in the literature on kernel density estimation —both spectral and probability—whenever possible. For this purpose, the following references are particularly pertinent: Parzen (1957), Priestley (1962), Epanechnikov (1969), and Sheather (1986).

We note that the results of this paper are used in a recent paper by Andrews and Monahan (1990) to investigate a class of prewhitened kernel HAC covariance matrix estimators. Prewhitened estimators have not been considered previously in the literature on HAC covariance matrix estimation, but have been used for some time in the spectral density estimation literature. Prewhitened HAC estimators turn out to have some advantages over the kernel estimators considered here and elsewhere in the econometrics literature in terms of the accuracy of nominal confidence levels and significance levels of confidence intervals and test statistics formed using the HAC estimators.

The remainder of the paper is organized as follows: Section 2 describes the estimation problem of concern and introduces the class of kernel HAC estimators under study. Section 3 presents consistency, rate of convergence, and asymptotic truncated MSE results for these estimators. Section 4 establishes the optimality of the QS kernel. Section 5 determines asymptotically optimal sequences of fixed bandwidth parameters. Section 6 introduces data-dependent "automatic" bandwidth parameter estimators using a plug-in method. Section 7 establishes consistency, rate of convergence, and asymptotic truncated MSE results for kernel HAC estimators based on these automatic bandwidths.

Section 8 extends many of the results of Sections 3–7, which apply to unconditionally stationary random variables, to nonstationary random variables. Section 9 presents Monte Carlo results regarding the finite sample behavior of the estimators considered in earlier sections. Section 10 provides a summary of the results of the paper. An appendix contains proofs of results given in the paper.

Those interested primarily in the definition of the preferred HAC estimator —a HAC estimator with QS kernel and automatic bandwidth—should read Sections 2 and 6.

2. A CLASS OF ESTIMATORS

To motivate the definition of the estimand given below, consider the linear regression model and LS estimator:

$$(2.1)\qquad Y_t = X_t'\theta_0 + U_t, \qquad (t = 1,\ldots,T),$$

$$\hat{\theta} = \left(\sum_{t=1}^{T} X_t X_t'\right)^{-1} \sum_{t=1}^{T} X_t Y_t, \quad \text{and}$$

$$\operatorname{Var}\left(\sqrt{T}\left(\hat{\theta} - \theta_0\right)\right) = \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1} \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} EU_s X_s (U_t X_t)' \left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}.$$

Since X_t is observed, consistent estimation of $\operatorname{Var}(\sqrt{T}(\hat{\theta} - \theta_0))$ just requires a consistent estimator of $(1/T)\Sigma_{s=1}^{T}\Sigma_{t=1}^{T} EU_s X_s (U_t X_t)'$.

More generally, many parameter estimators $\hat{\theta}$ in nonlinear dynamic models satisfy

$$(2.2)\qquad (B_T J_T B_T')^{-1/2}\sqrt{T}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} N(\mathbf{0}, I_r) \qquad \text{as} \qquad T \to \infty, \quad \text{where}$$

$$J_T = \frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T} EV_s(\theta_0)V_t(\theta_0)',$$

B_T is a nonrandom $r \times p$ matrix, and $V_t(\theta)$ is a random p-vector for each $\theta \in \Theta \subset R^r$. Often it is easy to construct estimators $\hat{B}_T$ such that $\hat{B}_T - B_T \to^p 0$ as $T \to \infty$. The estimators $\hat{B}_T$ usually are just sample analogues of B_T with θ_0 replaced by $\hat{\theta}$. See Hansen (1982), Gallant (1987, Ch. 7), Gallant and White (1988), Andrews and Fair (1988), and Andrews (1989) for the treatment of broad classes of parameter estimators and models that satisfy these conditions.[2] Since consistent estimators of B_T exist, one can estimate the "asymptotic variance" of $\sqrt{T}(\hat{\theta} - \theta_0)$, viz., $B_T J_T B_T'$, if one has a consistent estimator of J_T. In consequence, we concentrate our attention on the estimation of J_T.

The primary ingredient of J_T is the vector $V_t(\theta)$. For LS estimation of a linear regression model, $V_t(\theta) = (Y_t - X_t'\theta)X_t$. For pseudo-ML estimation, $V_t(\theta)$ is the score function for the tth observation. For instrumental variables estimation of

[2] In particular, the estimand J_T is given by $\operatorname{Var}((1/\sqrt{N})\Sigma_{n=1}^{N} w_n)$ in Hansen (1982), by $\bar{I}_n(\lambda_n^0)$ and $\bar{S}_n(\lambda_n^0)$ in Gallant (1987, pp. 549 and 570), by B_n^0 in Gallant and White (1988, p. 100), and by $\operatorname{Var}(\sqrt{T}\bar{m}_T(\theta_0, \tau_0))$ in Andrews and Fair (1988) and Andrews (1989).

a dynamic nonlinear simultaneous equation model, $V_t(\theta)$ is the Kronecker product of the vector of model equations evaluated at θ with the instrumental variables. For unit root models, the LS estimator does not satisfy (2.2). Nevertheless, one still needs to estimate the value of an expression that has the same form as J_T with $V_t(\theta) = Y_t - Y_{t-1}$ or $V_t(\theta) = Y_t - \theta Y_{t-1}$, where $\{Y_t\}$ is the unit root process; see Phillips (1987).

By change of variables, the estimand J_T can be rewritten as

$$J_T = \sum_{j=-T+1}^{T-1} \Gamma_T(j), \qquad \text{where} \tag{2.3}$$

$$\Gamma_T(j) = \begin{cases} \dfrac{1}{T} \displaystyle\sum_{t=j+1}^{T} EV_t V_{t-j}' & \text{for } j \geqslant 0, \\ \dfrac{1}{T} \displaystyle\sum_{t=-j+1}^{T} EV_{t+j} V_t' & \text{for } j < 0, \end{cases}$$

and $V_t = V_t(\theta_0)$, $t = 1, \ldots, T$.

When $\{V_t\}$ is second order stationary, it has spectral density matrix

$$f(\lambda) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \Gamma(j) e^{-ij\lambda}, \qquad \text{where} \qquad \Gamma(j) = EV_t V_{t-j}' \tag{2.4}$$

and $i = \sqrt{-1}$. The limit as $T \to \infty$ of the estimand J_T equals 2π times the spectral density matrix evaluated at $\lambda = 0$. This fact motivates the use of spectral density estimators to estimate J_T, as noted by Hansen (1982, p. 1047) and Phillips and Ouliaris (1988) among others. Furthermore, in the second order stationary context with known θ_0, the estimators of White (1984, p. 152), Gallant (1987, p. 533), and Newey and West (1987) correspond to kernel spectral density estimators evaluated at $\lambda = 0$. The aforementioned authors have established consistency of their estimators, however, in the more general context in which $\{V_t(\theta)\}$ is non-stationary and θ_0 is estimated.

The class of estimators we consider corresponds to Parzen's (1957) class of kernel estimators of the spectral density matrix. We consider estimators of the form

$$\hat{J}_T = \hat{J}_T(S_T) = \frac{T}{T-r} \sum_{j=-T+1}^{T-1} k\left(\frac{j}{S_T}\right) \hat{\Gamma}(j), \qquad \text{where} \tag{2.5}$$

$$\hat{\Gamma}(j) = \begin{cases} \dfrac{1}{T} \displaystyle\sum_{t=j+1}^{T} \hat{V}_t \hat{V}_{t-j}' & \text{for } j \geqslant 0, \\ \dfrac{1}{T} \displaystyle\sum_{t=-j+1}^{T} \hat{V}_{t+j} \hat{V}_t' & \text{for } j < 0, \end{cases}$$

$\hat{V}_t = V_t(\hat{\theta})$, $k(\cdot)$ is a real-valued kernel in the set $\mathscr{K}_1$ defined below, and S_T is a band-width parameter. The factor $T/(T-r)$ is a small sample degrees of freedom adjustment that is introduced to offset the effect of estimation of the r-vector θ. In Sections 3–5, we consider estimators $\hat{J}_T$ for which S_T is a given

nonrandom scalar. In Sections 6 and 7, we consider "automatic" estimators $\hat{J}_T$ for which S_T is a random function of the data.

The class of kernels $\mathscr{K}_1$ is given by

$$\mathscr{K}_1 = \Big\{ k(\cdot): R \to [-1,1] \,|\, k(0) = 1,\ k(x) = k(-x)\ \forall x \in R, \int_{-\infty}^{\infty} k^2(x)\,dx < \infty,\ k(\cdot) \text{ is continuous at 0 and at all but a finite number of other points} \Big\}. \tag{2.6}$$

The conditions $k(0) = 1$ and $k(\cdot)$ is continuous at 0 reflect the fact that for j small relative to T one wants the weight given to $\hat{\Gamma}(j)$ to be close to one.

Examples of kernels in $\mathscr{K}_1$ include the following:

(2.7)

Truncated: $$k_{TR}(x) = \begin{cases} 1 & \text{for } |x| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

Bartlett: $$k_{BT}(x) = \begin{cases} 1 - |x| & \text{for } |x| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

Parzen: $$k_{PR}(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{for } 0 \leqslant |x| \leqslant 1/2, \\ 2(1 - |x|)^3 & \text{for } 1/2 \leqslant |x| \leqslant 1, \\ 0 & \text{otherwise} \end{cases}$$

Tukey-Hanning: $$k_{TH}(x) = \begin{cases} (1 + \cos(\pi x))/2 & \text{for } |x| \leqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

Quadratic Spectral: $$k_{QS}(x) = \frac{25}{12\pi^2 x^2}\left(\frac{\sin(6\pi x/5)}{6\pi x/5} - \cos(6\pi x/5) \right).$$

The estimators $\hat{J}_T$ corresponding to the truncated, Bartlett, and Parzen kernels are the estimators proposed by White (1984, p. 152), Newey and West (1987), and Gallant (1987, p. 533) respectively. The Tukey-Hanning and QS kernels have not been considered in the literature concerning HAC estimation. The Tukey-Hanning kernel is popular in the spectral density estimation literature, however, and the QS kernel has been considered in the spectral and probability density estimation literature by Priestley (1962) and Epanechnikov (1969) respectively.

If $k(x) = 0$ for $|x| > 1$ (and $k(x) \neq 0$ for some $|x|$ arbitrarily close to 1), then S_T is referred to as the *lag truncation parameter*, because lags of order $j > S_T$ receive zero weight.[3] Since some kernels in $\mathscr{K}_1$ are nonzero for arbitrarily large

[3] The lag truncation parameters of White (1984, p. 152), White and Domowitz (1984), Newey and West (1987), Gallant (1987, pp. 533, 551, 573), and Gallant and White (1988, p. 97), viz., l, l, m, $l(n)$, and m_n respectively, are equal to $S_T - 1$ in our notation when S_T is an integer. The aforementioned authors consider only integer-valued lag truncation parameters, but there is no reason to restrict the estimators in this way and our formulae below for optimal S_T values yield real-valued parameters.

For example, Newey and West (1987) define their weights as $1 - j/(m+1)$ for $j \leqslant m$ and 0 otherwise, where m is an integer. In our notation, their weights are $1 - j/S_T$ for $j \leqslant S_T$ and 0 otherwise, where S_T is real-valued. If S_T is an integer, then these weights are equivalent when $S_T = m + 1$.

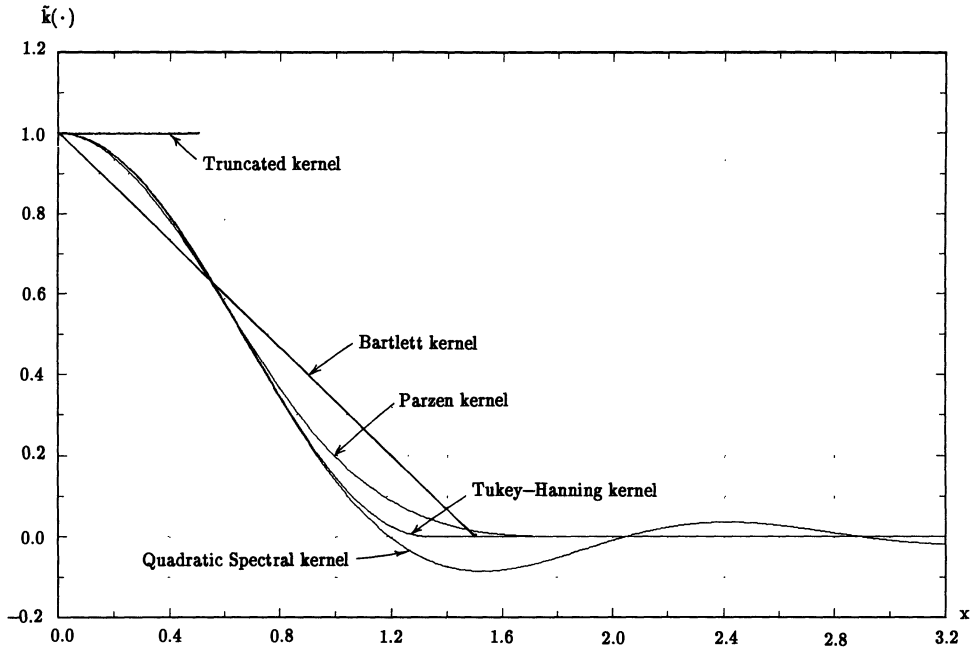

FIGURE 1.—Comparison of kernels.[a]

[a]These kernels have been renormalized as described in the text below equation (2.7).

values of x, it is not possible to normalize all kernels in $\mathscr{K}_1$ such that $k(x) = 0$ for $|x| > 1$. Thus, lag truncation parameters do not exist for all kernels in $\mathscr{K}_1$. The QS kernel is an example.

Figure 1 graphs the five kernels of (2.7), but renormalized such that each yields the same asymptotic variance of $\hat{J}_T$—only their asymptotic biases vary.[4] (The renormalization is necessary for comparative purposes in order to make any given S_T value equally suitable for each kernel.) For a given value of S_T, the figure illustrates the different weights the renormalized kernels $\tilde{k}(\cdot)$ put on the lagged covariances. For example, if $S_T = 3$, then $\tilde{k}_{BT}(1/3), \tilde{k}_{BT}(2/3), \ldots,$ are the weights the renormalized Bartlett kernel puts on $\hat{\Gamma}(1), \hat{\Gamma}(2), \ldots$.

For some results below, we consider a subset of $\mathscr{K}_1$. Let

$$\mathscr{K}_2 = \{k(\cdot) \in \mathscr{K}_1 | K(\lambda) \geq 0 \ \forall \lambda \in R\}, \qquad \text{where}$$

$$K(\lambda) = \frac{1}{2\pi} \int_{-\infty}^{\infty} k(x) e^{-ix\lambda} \, dx.$$

The function $K(\lambda)$ is referred to as the *spectral window generator* corresponding to the kernel $k(\cdot)$. The set $\mathscr{K}_2$ contains all kernels in $\mathscr{K}_1$ that necessarily

[4] By construction, a renormalized kernel $\tilde{k}(\cdot)$ satisfies $\int_{-\infty}^{\infty} \tilde{k}^2(x) \, dx = 1$. The renormalized kernels of (2.7) are given by $\tilde{k}_a(x) = k_a(c_a x)$ for $a = TR$, BT, PR, and TH, where $c_a = \int k_a^2(x) \, dx$, $c_{TR} = 2$, $c_{BT} = 2/3$, $c_{PR} = .539285$, and $c_{TH} = 3/4$. The QS kernel satisfies $\int k^2(x) \, dx = 1$, and hence, does not need to be renormalized.

generate positive semi-definite (psd) estimators in finite samples. (To see this, note that estimators of the form (2.5) are weighted averages of the periodogram matrix at different frequencies λ with weights given by $K(\lambda)$, e.g., see Priestley (1981, pp. 580–581). Since the periodogram is psd, so is an estimator $\hat{J}_T$ provided $K(\lambda) \geqslant 0\ \forall \lambda \in R$.) As emphasized by Newey and West (1987), this property usually is highly desirable. $\mathscr{K}_2$ contains the Bartlett, Parzen, and QS kernels, but not the truncated or Tukey-Hanning kernels.

3. FIXED BANDWIDTH HAC ESTIMATORS

In this section, consistency, rate of convergence, and asymptotic truncated MSE properties of fixed bandwidth kernel HAC estimators are determined. Results due to Parzen (1957) for spectral density estimators are utilized. The results of this section and those of Sections 4–7 apply to unconditionally fourth or eighth order stationary random variables (rv's), as specified below. This allows for conditional heteroskedasticity. Many of the results are extended in Section 8 to cover unconditionally nonstationary rv's.

We begin by introducing the basic assumption that controls the temporal dependence of $\{V_t\}$. Let $\kappa_{abcd}(t, t+j, t+m, t+n)$ denote the fourth order cumulant of $(V_{at}, V_{bt+j}, V_{ct+m}, V_{dt+n})$, where V_{at} denotes the ath element of V_t. That is,

$$
\begin{aligned}
(3.1)\quad & \kappa_{abcd}(t, t+j, t+m, t+n) \\
& = E(V_{at} - EV_{at})(V_{bt+j} - EV_{bt+j})(V_{ct+m} - EV_{ct+m})(V_{dt+n} - EV_{dt+n}) \\
& \quad - E(\tilde{V}_{at} - E\tilde{V}_{at})(\tilde{V}_{bt+j} - E\tilde{V}_{bt+j})(\tilde{V}_{ct+m} - E\tilde{V}_{ct+m})(\tilde{V}_{dt+n} - E\tilde{V}_{dt+n}),
\end{aligned}
$$

where $\{\tilde{V}_t\}$ is a Gaussian sequence with the same mean and covariance structure as $\{V_t\}$. Let $\|\cdot\|$ denote the Euclidean norm of a vector or matrix.

ASSUMPTION A: *$\{V_t\}$ is a mean zero, fourth order stationary sequence of rv's with $\sum_{j=-\infty}^{\infty} \|\Gamma(j)\| < \infty$ and $\sum_{j=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \kappa_{abcd}(0, j, m, n) < \infty\ \forall a, b, c, d \leqslant p$.*

Assumption A allows for conditional heteroskedasticity, as well as autocorrelation, but prohibits unconditional heteroskedasticity.

The cumulant condition of Assumption A is standard in the time series literature; e.g., see Anderson (1971, pp. 465, 520, 531) and Hannan (1970, p. 280). In fact, Brillinger (1981) assumes that the cumulant condition of Assumption A holds not only for the fourth order but for all higher orders as well throughout his book (see his Assumption 2.6.1, p. 26). In the Gaussian case, the fourth order cumulants are zero, so the cumulant condition is satisfied trivially. In addition, it is well known that fourth order stationary linear processes (with absolutely summable coefficients and innovations whose fourth moments exist) satisfy the cumulant condition of Assumption A (e.g., see Hannan (1970, p. 211)). The following lemma shows that the cumulant condition

of Assumption A also is implied by an α-mixing (i.e., strong mixing) condition plus a moment condition:

LEMMA 1: *Suppose $\{V_t\}$ is a mean zero (not necessarily fourth order stationary) α-mixing sequence of rv's. If $\sup_{t\geq 1} E\|V_t\|^{4\nu} < \infty$ and $\sum_{j=1}^{\infty} j^2\alpha(j)^{(\nu-1)/\nu} < \infty$ for some $\nu > 1$, then $\sum_{j=0}^{\infty} \sup_{t\geq 1}\|EV_tV_{t+j}'\| < \infty$ and*

$$\sum_{j=1}^{\infty}\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\sup_{t\geq 1}|\kappa_{abcd}(t,t+j,t+m,t+n)| < \infty \quad \forall a,b,c,d \leq p.$$

In particular, if $\{V_t\}$ also is fourth order stationary, then Assumption A holds.

COMMENT: The condition on the mixing numbers in Lemma 1 is satisfied if they are of size $-3\nu/(\nu-1)$ (i.e., $\alpha(j) = O(j^{-\varepsilon-3\nu/(\nu-1)})$ for some $\varepsilon > 0$). The latter condition is slightly stronger than that used by White (1984, Theorem 6.20, p. 155), Newey and West (1987), and Kool (1988). (These authors use the same condition but with 3 replaced by 2.)

Let $\tilde{J}_T$ denote the pseudo-estimator that is identical to $\hat{J}_T$ but is based on the unobserved sequence $\{V_t\} = \{V_t(\theta_0)\}$ rather than $\{\hat{V}_t\} = \{V_t(\hat{\theta})\}$ and is defined without the degrees of freedom correction $T/(T-r)$:

$$\text{(3.2)} \qquad \tilde{J}_T = \sum_{j=-T+1}^{T-1} k(j/S_T)\tilde{\Gamma}(j) \qquad \text{and}$$

$$\tilde{\Gamma}(j) = \begin{cases} \dfrac{1}{T}\displaystyle\sum_{t=j+1}^{T} V_tV_{t-j}' & \text{for } j \geq 0, \\ \dfrac{1}{T}\displaystyle\sum_{t=-j+1}^{T} V_{t+j}V_j' & \text{for } j < 0. \end{cases}$$

First, we summarize well known results for the pseudo-estimator $\tilde{J}_T$. Then, we show that analogous results hold for the estimator $\hat{J}_T$.

The asymptotic bias of kernel estimators depends on the smoothness of the kernel at zero and on the smoothness of the spectral density matrix $f(\lambda)$ of $\{V_t\}$ at zero. Following Parzen (1957), define

$$\text{(3.3)} \qquad k_q = \lim_{x\to 0}\frac{1-k(x)}{|x|^q} \qquad \text{for} \qquad q \in [0,\infty).$$

The smoother is the kernel at zero, the larger is the value of q for which k_q is finite. If q is an even integer, then

$$k_q = -\frac{1}{q!}\frac{d^q k(x)}{dx^q}\bigg|_{x=0}$$

and $k_q < \infty$ if and only if $k(x)$ is q times differentiable at zero. For the truncated kernel, $k_q = 0$ for all $q < \infty$. For the Bartlett kernel, $k_1 = 1$, $k_q = 0$ for $q < 1$, and $k_q = \infty$ for $q > 1$. For the Parzen, Tukey-Hanning, and QS kernels, $k_2 = 6$, $\pi^2/4$, and 1.421223, respectively, $k_q = 0$ for $q < 2$, and $k_q = \infty$ for $q > 2$.

The smoothness of $f(\lambda)$ at $\lambda = 0$ is indexed by

$$f^{(q)} = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} |j|^q \Gamma(j) \qquad \text{for} \qquad q \in [0, \infty). \tag{3.4}$$

If q is even, then

$$f^{(q)} = (-1)^{q/2} \frac{d^q f(\lambda)}{d\lambda^q}\bigg|_{\lambda=0}$$

and $\|f^{(q)}\| < \infty$ if and only if $f(\lambda)$ is q times differentiable at $\lambda = 0$.

Let f denote the spectrum of $\{V_t\}$ at zero, i.e., $f = f(0)$. Define

$$\text{MSE}\left(T/S_T, \tilde{J}_T, W\right) = \frac{T}{S_T} E \operatorname{vec}\left(\tilde{J}_T - J_T\right)' W \operatorname{vec}\left(\tilde{J}_T - J_T\right), \tag{3.5}$$

where W is some $p^2 \times p^2$ weight matrix and $\operatorname{vec}(\cdot)$ is the column by column vectorization function. Let tr denote the trace function and $\otimes$ the tensor (or Kronecker) product operator. Let K_{pp} denote the $p^2 \times p^2$ commutation matrix that transforms $\operatorname{vec}(A)$ into $\operatorname{vec}(A')$, i.e., $K_{pp} = \sum_{i=1}^{p} \sum_{j=1}^{p} e_i e_j' \otimes e_j e_i'$, where e_i is the ith elementary p-vector; see Magnus and Neudecker (1979). Unless indicated otherwise, all limits in the paper are taken as $T \to \infty$.

The following results for $\tilde{J}_T$ are due to Parzen (1957) for the scalar V_t case. Hannan (1970, pp. 280, 283) gives the corresponding vector V_t results.

PROPOSITION 1: *Suppose* $k(\cdot) \in \mathscr{K}_1$, *Assumption A holds*, $S_T \to \infty$ *and* $S_T/T \to 0$. *Then, we have*:

(a) $\lim_{T\to\infty}(T/S_T)\operatorname{Var}(\operatorname{vec} \tilde{J}_T) = 4\pi^2 \int k^2(x)\, dx (I + K_{pp}) f \otimes f$.

(b) *If* $S_T^q/T \to 0$ *for some* $q \in [0, \infty)$ *for which* k_q, $\|f^{(q)}\| \in [0, \infty)$, *then* $\lim_{T\to\infty} S_T^q(E\tilde{J}_T - J_T) = -2\pi k_q f^{(q)}$.

(c) *If* $S_T^{2q+1}/T \to \gamma \in (0, \infty)$ *for some* $q \in (0, \infty)$ *for which* k_q, $\|f^{(q)}\| < \infty$, *then*

$$\lim_{T\to\infty} \text{MSE}\left(T/S_T, \tilde{J}_T, W\right) = 4\pi^2\left(k_q^2 (\operatorname{vec} f^{(q)})' W \operatorname{vec} f^{(q)}/\gamma + \int k^2(x)\, dx \operatorname{tr} W(I + K_{pp}) f \otimes f\right).$$

COMMENT: By Proposition 1(a), the covariance between the (a, b) and (c, d) elements of $\tilde{J}_T$ is $4\pi^2 \int k^2(x)\, dx (f_{ac} f_{bd} + f_{ad} f_{bc})$, where f_{ab} denotes the (a, b) element of f.

Next we state additional assumptions used to obtain results for the estimator of interest $\hat{J}_T$. The first assumption below, together with Assumption A, is sufficient for consistency of $\hat{J}_T$ when $S_T = o(T^{1/2})$. Let Θ denote some convex neighborhood of θ_0.

ASSUMPTION B: (i) $\sqrt{T}(\hat{\theta} - \theta_0) = O_p(1)$. (ii) $\sup_{t\geq 1} E\|V_t\|^2 < \infty$. (iii) $\sup_{t\geq 1} E \sup_{\theta\in\Theta} \|(\partial/\partial\theta') V_t(\theta)\|^2 < \infty$. (iv) $\int_{-\infty}^{\infty} |k(x)|\, dx < \infty$.

Assumption B is not overly restrictive and usually is easy to verify. Its first part follows from asymptotic normality of $\sqrt{T}(\hat{\theta} - \theta_0)$. Its second and third parts

are common conditions used to obtain asymptotic normality of $\sqrt{T}(\hat{\theta} - \theta_0)$. Its fourth part is satisfied by each of the kernels of (2.7) and by almost all other kernels that have been used in practice. The first three parts of Assumption B are identical to assumptions of Newey and West (1987).

The next assumption needs to be imposed in place of Assumption A in order to obtain sharp rate of convergence results and to obtain consistency of $\hat{J}_T$ when S_T is only required to satisfy $S_T = o(T)$.

ASSUMPTION C: (i) *Assumption A holds with V_t replaced by*

$$\left(V_t', \operatorname{vec}\left(\frac{\partial}{\partial\theta'}V_t(\theta_0) - E\frac{\partial}{\partial\theta'}V_t(\theta_0)\right)'\right)'.$$

(ii) $\sup_{t\geqslant 1} E \sup_{\theta\in\Theta}\|(\partial^2/\partial\theta\,\partial\theta')V_{at}(\theta)\|^2 < \infty \quad \forall a = 1,\ldots,p$, *where* $V_t(\theta) = (V_{1t}(\theta),\ldots,V_{pt}(\theta))'$.

Suppose $V_t(\theta)$ is of the form $V(Z_t,\theta)$ for some rv Z_t and some (measurable) function $V(\cdot,\cdot)$. In this case, Assumption C(i) holds if $EV_t = 0 \ \forall t \geqslant 1$, $(V_t', \operatorname{vec}((\partial/\partial\theta')V_t(\theta_0))')'$ is fourth order stationary, $\{Z_t\colon t \geqslant 1\}$ is strong mixing with $\sum_{j=1}^{\infty} j^2\alpha(j)^{(\nu-1)/\nu} < \infty$, and

$$\sup_{t\geqslant 1}\left(E\|V_t\|^{4\nu} + E\left\|\frac{\partial}{\partial\theta'}V_t(\theta_0)\right\|^{4\nu}\right) < \infty \qquad \text{for some } \nu > 1.$$

Under the assumptions above, the effect of using $\hat{\theta}$ rather than θ_0 when constructing $\hat{J}_T$ is at most $o_p(1)$. Nevertheless, if $\hat{\theta}$ has infinite second moment (as occurs, e.g., with the two stage LS estimator in some scenarios) its use can dominate the MSE criterion of (3.5). To circumvent undue influence of $\hat{\theta}$ on the criterion of performance, we replace the MSE criterion with a truncated MSE criterion. Define

$$\text{(3.6)} \qquad \operatorname{MSE}_h(T/S_T, \hat{J}_T, W_T) = E\min\left\{\left|\frac{T}{S_T}\operatorname{vec}(\hat{J}_T - J_T)'W_T\operatorname{vec}(\hat{J}_T - J_T)\right|, h\right\},$$

where W_T is a $p^2 \times p^2$ weight matrix that may be random. The criterion that we use for the optimality results is asymptotic truncated MSE with arbitrarily large truncation point, viz., $\lim_{h\to\infty}\lim_{T\to\infty}\operatorname{MSE}_h(T/S_T, \hat{J}_T, W_T)$. This criterion yields the same value as the asymptotic MSE criterion of Proposition 1 when $\hat{\theta}$ has well defined moments, but does not blow up when $\hat{\theta}$ has infinite second moments.

To obtain the desired asymptotic truncated MSE results, we impose an additional assumption. Let V_{at} denote the ath element of V_t. Let $\kappa_{a_1\cdots a_8}(0, j_1,\ldots,j_7)$ denote the cumulant of $(V_{a_10}, V_{a_2j_1},\ldots,V_{a_8j_7})$ (e.g., see Brillinger (1981, p. 19)), where $a_1,\ldots,a_8$ are positive integers less than $p+1$ and $j_1,\ldots,j_7$ are integers.

ASSUMPTION D: (i) $\{V_t\}$ *is eighth order stationary with*

$$\sum_{j_1=-\infty}^{\infty} \cdots \sum_{j_7=-\infty}^{\infty} \kappa_{a_1 \cdots a_8}(0, j_1, \ldots, j_7) < \infty.$$

(ii) $W_T \xrightarrow{p} W$.

As noted above, Assumption D(i) is part of the assumption utilized by Brillinger (1981, p. 26). It seems likely that an analogue of Lemma 1 could be established which would show that α-mixing plus a moment condition implies Assumption D(i). Without Assumption D(i), the right-hand side of (3.5) with the expectation removed is L^1 bounded. Assumption D(i) is used only to ensure that it is also $L^{1+\delta}$ bounded for some $\delta > 0$. Any other assumption that suffices for this result could be used in place of Assumption D(i).

Utilizing the assumptions above, we have the following theorem.

THEOREM 1: *Suppose* $k(\cdot) \in \mathscr{K}_1$ *and* $S_T \to \infty$.

(a) *If Assumptions A and B hold and* $S_T^2/T \to 0$, *then* $\hat{J}_T - J_T \xrightarrow{p} 0$ *and* $\hat{J}_T - \tilde{J}_T \xrightarrow{p} 0$.

(b) *If Assumptions B and C hold and* $S_T^{2q+1}/T \to \gamma \in (0, \infty)$ *for some* $q \in (0, \infty)$ *for which* k_q, $\|f^{(q)}\| < \infty$, *then* $\sqrt{T/S_T}(\hat{J}_T - J_T) = O_p(1)$ *and* $\sqrt{T/S_T}(\hat{J}_T - \tilde{J}_T) \xrightarrow{p} 0$.

(c) *Under the conditions of part* (*b*) *plus Assumption D*,

$$\begin{aligned}
&\lim_{h\to\infty} \lim_{T\to\infty} \mathrm{MSE}_h\left(T/S_T, \hat{J}_T, W_T\right) \\
&\quad = \lim_{h\to\infty} \lim_{T\to\infty} \mathrm{MSE}_h\left(T/S_T, \tilde{J}_T, W_T\right) \\
&\quad = \lim_{T\to\infty} \mathrm{MSE}\left(T/S_T, \tilde{J}_T, W\right) \\
&\quad = 4\pi^2 \Bigg(k_q^2 (\mathrm{vec}\, f^{(q)})' W \,\mathrm{vec}\, f^{(q)}/\gamma \\
&\qquad\qquad + \int k^2(x)\, dx \,\mathrm{tr}\, W(I + K_{pp}) f \otimes f \Bigg).
\end{aligned}$$

COMMENTS: 1. In contrast to the results of White (1984, pp. 147–161) and Newey and West (1987), the consistency results of Theorem 1(a) apply to kernels with unbounded support and to bandwidth parameter sequences that grow at rate $o(T^{1/2})$ rather than $o(T^{1/4})$. These extensions are useful, because the optimal kernel discussed below has unbounded support and the optimal growth rate of the bandwidth parameters for the Bartlett kernel considered by Newey and West (1987) exceeds $o(T^{1/4})$. (It equals $O(T^{1/3})$.)

2. Theorem 1(b) yields consistency of $\hat{J}_T$ with S_T only required to be $o(T)$. This extension is of theoretical interest, but is of little practical import, because optimal growth rates typically are less than $T^{1/2}$ (see Section 5 below), and hence, are covered by the results of Theorem 1(a) under weaker assumptions. The main contribution of Theorem 1(b) is the rate of convergence results that it

delivers. These rates are identical to those in the case where no parameters θ_0 are estimated. As indicated by Theorem 1(c), the rates are sharp. If the bandwidth parameters are chosen to grow at the optimal rate determined in Section 5 below, then the rate of convergence for the Bartlett, Parzen, Tukey-Hanning, and QS kernels are $T^{1/3}$, $T^{2/5}$, $T^{2/5}$, and $T^{2/5}$ respectively. In contrast, the rate for parametric estimators typically is $T^{1/2}$.

3. The expression given in Theorem 1(c) for the asymptotic truncated MSE of $\hat{J}_T$ is identical to that of Proposition 1(c) for the asymptotic untruncated MSE of $\tilde{J}_T$. Theorem 1(c) is used below in determining an asymptotically optimal kernel and sequence of fixed bandwidth parameters (see Sections 4 and 5), as well as in determining automatic bandwidth parameters (see Sections 6 and 7).

4. AN OPTIMAL KERNEL

In this section, we show that the QS kernel is best with respect to asymptotic truncated MSE in the class $\mathscr{K}_2$ of kernels that necessarily generate psd estimates. This optimality property holds for any psd (limiting) weight matrix W and any distribution of $\{V_t\}$ such that Assumptions B–D hold.

The asymptotic truncated MSE criterion utilized here is justifiable if $\hat{J}_T$ is used to construct a standard error or variance estimator for $\hat{\theta}$ and one views this as an estimation problem in its own right. If one wants to use $\hat{J}_T$ in forming a test statistic involving $\hat{\theta}$, however, the suitability of the truncated MSE criterion is less clear. A weak argument in its favor is that the asymptotics typically used with such test statistics treat the estimated covariance matrix as though it equals its probability limit. In consequence, in many cases the closer is the covariance matrix estimator to its probability limit, as measured, for example, by truncated MSE, the better is the asymptotic approximation. This is true in the context of the Monte Carlo experiments reported in Section 9 below. On the other hand, there are cases where the deviation of one part of a test statistic from its limiting behavior is offset by the deviation of another part of the statistic from its limiting behavior. In such cases, the argument above breaks down.

The focus on the asymptotic truncated MSE of $\hat{J}_T$ for J_T rather than of $\hat{B}_T\hat{J}_T\hat{B}_T'$ for $B_TJ_TB_T'$ can be justified when the rate of convergence of $\hat{B}_T$ is faster than that of $\hat{J}_T$, as is usually the case. In particular, one can choose the weight matrix W_T in such a way as to obtain asymptotic truncated MSE results for $\hat{B}_T\hat{J}_T\hat{B}_T'$ from the corresponding results for $\hat{J}_T$; see (6.9) and the discussion following it below.

In addition to the results of this section, the QS kernel has been shown to possess optimality properties in the context of spectral density estimation (see Priestley (1962; 1981, pp. 567–571)) and probability density estimation (see Epanechnikov (1969) and Sacks and Yvisacker (1981)). The results of Priestley and Epanechnikov are for an asymptotic maximum relative MSE criterion (where the maximum is over different frequencies or points of support) rather than for a criterion of asymptotic truncated MSE at a given point as is used here. In addition, the present results establish optimality for any given band-

width sequence $\{S_T\}$, whereas each of the other results referred to above establishes optimality only for a particular bandwidth sequence that is optimal in some sense.

Since the kernels in $\mathscr{K}_2$ are not subject to any normalization, it is meaningless to compare two kernels using the same sequence of bandwidth parameters $\{S_T\}$. For example, two kernels that are the same but scaled differently would yield nonidentical results in such a comparison. To make comparisons meaningful, one has to use *comparable bandwidths*. The latter are defined as follows: Given $k(\cdot) \in \mathscr{K}_2$, the QS kernel $k_{QS}(\cdot)$, and a sequence $\{S_T\}$ of bandwidth parameters to be used with the QS kernel, define a comparable sequence $\{S_{Tk}\}$ of bandwidth parameters for use with $k(\cdot)$ such that both kernel/bandwidth combinations have the same asymptotic truncated variance when scaled by the same factor T/S_T. (That is, S_{Tk} is such that

$$\lim_{h\to\infty}\lim_{T\to\infty} \mathrm{MSE}_h\left(T/S_T, \hat{J}_{QST}(S_T) - E\tilde{J}_{QST}(S_T) + J_T, W_T\right)$$
$$= \lim_{h\to\infty}\lim_{T\to\infty} \mathrm{MSE}_h\left(T/S_T, \hat{J}_T(S_{Tk}) - E\tilde{J}_T(S_{Tk}) + J_T, W_T\right),$$

where the subscript QS denotes the estimator is based on the QS kernel.) This definition yields

$$S_{Tk} = S_T / \int k^2(x)\, dx. \tag{4.1}$$

(See footnote 4 for the value of $\int k^2(x)\,dx$ for the kernels of (2.7).)

Note that for the QS kernel $k_{QS}(\cdot)$, $S_{Tk_{QS}} = S_T$, since $\int k_{QS}^2(x)\,dx = 1$. Also note that the use of the QS kernel as the standard for comparability is made for convenience only and does not affect the optimality results.

Let $\hat{J}_{QST}(S_T)$ denote $\hat{J}_T(S_T)$ when the latter is based on the QS kernel.

THEOREM 2: *Suppose Assumptions B–D hold,* $\|f^{(2)}\| < \infty$, *and W is psd. For any sequence of bandwidth parameters* $\{S_T\}$ *such that* $S_T \to \infty$ *and* $S_T^5/T \to \gamma$ *for some* $\gamma \in (0,\infty)$ *and for any kernel* $k(\cdot) \in \mathscr{K}_2$ *that is used to construct* $\hat{J}_T$, *the QS kernel is preferred to* $k(\cdot)$ *in the sense that*

$$\lim_{h\to\infty}\lim_{T\to\infty}\Big(\mathrm{MSE}_h\left(T/S_T, \hat{J}_T(S_{Tk}), W_T\right)$$
$$- \mathrm{MSE}_h\left(T/S_T, \hat{J}_{QST}(S_T), W_T\right)\Big)$$
$$= 4\pi^2(\mathrm{vec}\, f^{(2)})' W\, \mathrm{vec}\, f^{(2)}\left[k_2^2\left(\int k^2(x)\,dx\right)^4 - k_{2QS}^2\right]\Big/\gamma$$
$$\geqslant 0$$

provided $(\mathrm{vec}\, f^{(2)})' W\, \mathrm{vec}\, f^{(2)} > 0$. *The inequality is strict if* $k(x) \neq k_{QS}(x)$ *with positive Lebesgue measure.*

COMMENT: The requirement of the Theorem that $\|f^{(2)}\| < \infty$ is not stringent. Nevertheless, if $\|f^{(q)}\| < \infty$ only for some $1 \leqslant q < 2$, then Theorem 2 does not apply, but the results of Theorem 1 can be used to show that any kernel with

$k_q = 0$ has smaller asymptotic truncated MSE than a kernel with $k_q > 0$. In particular, the QS, Parzen, and Tukey-Hanning kernels have $k_q = 0$ for $1 \leq q < 2$, whereas the Bartlett kernel has $k_q > 0$ for $1 \leq q < 2$. Thus, the asymptotic superiority of the former kernels over the Bartlett kernel holds even if $\|f^{(q)}\| < \infty$ only for $1 \leq q < 2$.

5. OPTIMAL FIXED BANDWIDTH PARAMETERS

In this section, sequences of fixed bandwidth parameters are determined that are optimal in the sense of minimizing asymptotic truncated MSE for a given psd (limiting) weight matrix W. The results apply to each kernel $k(\cdot)$ in $\mathscr{K}_1$ for which $k_q \in (0, \infty)$ for some $q \in (0, \infty)$. This excludes the truncated kernel, but includes all of the other kernels of (2.7). The results are obtained as a simple corollary to Theorem 1(c) above.

Define the optimal bandwidth parameters $\{S_T^*\}$ as follows: Let

$$\alpha(q) = \frac{2(\operatorname{vec} f^{(q)})' W \operatorname{vec} f^{(q)}}{\operatorname{tr} W(I + K_{pp}) f \otimes f} \quad \text{and} \tag{5.1}$$

$$S_T^* = \left(q k_q^2 \alpha(q) T \Big/ \int k^2(x)\, dx \right)^{1/(2q+1)}. \tag{5.2}$$

$\alpha(q)$ is a function of the unknown spectral density matrix $f(\lambda)$. Hence, the optimal bandwidth parameter S_T^* also is unknown in practice. For this reason, estimates of $\alpha(q)$ are considered in Sections 6 and 7 below in order to obtain a feasible analogue of S_T^*.

COROLLARY 1: *Suppose Assumptions B–D hold. Consider a kernel $k(\cdot) \in \mathscr{K}_1$ for which $k_q \in (0, \infty)$ for some $q \in (0, \infty)$. Suppose $\|f^{(q)}\| < \infty$, $\alpha(q) \in (0, \infty)$, and W is psd. For any sequence of bandwidth parameters $\{S_T\}$ such that $S_T^{2q+1}/T \to \gamma$ for some $\gamma \in (0, \infty)$, the sequence $\{S_T^*\}$ is preferred to $\{S_T\}$ in the sense that*

$$\lim_{h \to \infty} \lim_{T \to \infty} \left(\mathrm{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(S_T), W_T\right) - \mathrm{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(S_T^*), W_T\right) \right) \geq 0.$$

The inequality is strict unless $S_T = S_T^* + o(T^{1/(2q+1)})$.

COMMENTS: 1. The values of q in Corollary 1 for the Bartlett, Parzen, Tukey-Hanning, and QS kernels are 1, 2, 2, and 2, respectively. Thus, we have

(5.3)

Bartlett kernel: $S_T^* = 1.1447(\alpha(1)T)^{1/3}$,

Parzen kernel: $S_T^* = 2.6614(\alpha(2)T)^{1/5}$,

Tukey-Hanning kernel: $S_T^* = 1.7462(\alpha(2)T)^{1/5}$,

Quadratic Spectral kernel: $S_T^* = 1.3221(\alpha(2)T)^{1/5}$.

TABLE I

ASYMPTOTICALLY OPTIMAL LAG TRUNCATION / BANDWIDTH VALUES S_T^* FOR THE BARTLETT, PARZEN, TUKEY-HANNING, AND QS ESTIMATORS FOR AR(1) $\{V_t\}$ PROCESSES WITH PARAMETER η[a,b]

	Bartlett Estimator (Newey-West (1987))						Parzen Estimator (Gallant (1987))					
T	ρ .2 η .04	.3 .09	.5 .25	.7 .49	.9 .81	.95 .90	.2 .04	.3 .09	.5 .25	.7 .49	.9 .81	.95 .90
32	.7	1.2	2.4	4.3	10.2	16.6	2.0	2.9	5.1	9.0	24.4	43.4
64	.9	1.5	3.0	5.4	12.9	20.9	2.3	3.3	5.8	10.4	28.0	49.9
128	1.1	1.8	3.8	6.8	16.2	26.3	2.6	3.8	6.7	11.9	32.2	57.3
256	1.4	2.3	4.8	8.6	20.4	33.1	3.0	4.4	7.7	13.7	36.9	65.8
512	1.7	2.9	6.0	10.9	25.7	41.7	3.5	5.0	8.8	15.8	42.4	75.6
1,024	2.1	3.7	7.6	13.7	32.4	52.6	4.0	5.8	10.2	18.1	48.7	86.8
	Tukey-Hanning Estimator						Quadratic Spectral Estimator					
T	ρ .2 η .04	.3 .09	.5 .25	.7 .49	.9 .81	.95 .90	.2 .04	.3 .09	.5 .25	.7 .49	.9 .81	.95 .90
32	1.3	1.9	3.3	5.9	16.0	28.5	1.0	1.4	2.5	4.5	12.1	21.6
64	1.5	2.2	3.8	6.8	18.4	32.7	1.1	1.6	2.9	5.2	13.9	24.8
128	1.7	2.5	4.4	7.8	21.1	37.6	1.3	1.9	3.3	5.9	16.0	28.5
256	2.0	2.9	5.0	9.0	24.2	43.2	1.5	2.2	3.8	6.8	18.4	32.7
512	2.3	3.3	5.8	10.3	27.8	49.6	1.7	2.5	4.4	7.8	21.1	37.5
1,024	2.6	3.8	6.7	11.9	32.0	57.0	2.0	2.9	5.0	9.0	24.2	43.1

[a] The given values of S_T^* are optimal for an iid linear regression model with AR(1) regressors and errors each with AR parameter ρ. This corresponds to $\{V_t\}$ $(=\{U_t X_t\})$ being AR(1) with parameter $\eta = \rho^2$.

[b] The truncation parameters m and $l(n)$ of Newey and West (1987) (Bartlett estimator) and Gallant (1987, pp. 533, 551, 573) (Parzen estimator), respectively, correspond to $S_T^* - 1$; see footnote 3.

2. For illustrative purposes, Table I tabulates S_T^* for the Bartlett, Parzen, Tukey-Hanning, and QS kernels for a linear regression model in which the regressors and errors are mutually independent, homoskedastic, first order autoregressive (AR(1)) random variables each with autoregressive parameter ρ. For this model each element of V_t (except that corresponding to the intercept) has correlation structure identical to that of an AR(1) process with parameter $\eta = \rho^2$. The weight matrix W_T is taken to be a diagonal matrix that gives weight one to the diagonal elements of $\hat{J}_T - J_T$ that correspond to nonconstant regressors and weight zero to all other elements.

3. When the optimal bandwidth parameters $\{S_T^*\}$ are used, the asymptotic truncated MSE is such that the squared bias equals $1/(2q+1)$ of the total MSE (for any limiting psd weight matrix W). Thus, the bias of the Bartlett kernel accounts for a greater fraction of its MSE asymptotically than do the biases of the Parzen, Tukey-Hanning, and QS kernels.

4. When $\{S_T^*\}$ is used, the Parzen and Tukey-Hanning kernels are 8.6% less and .9% more efficient asymptotically than the QS estimator, respectively, for any distribution of $\{V_t\}$ and any limiting psd weight matrix W. (Since the Tukey-Hanning kernel does not necessarily generate psd estimates, i.e., $k_{TH}(x) \notin \mathcal{K}_2$, the latter result does not violate Theorem 2.) Also, the Bartlett kernel is 100% less efficient asymptotically than the Parzen, Tukey-Hanning, and QS

kernels, since its MSE converges to zero at a slower rate. In particular finite sample situations, however, the Bartlett kernel may not perform nearly so poorly in relative terms, depending on the magnitudes of T, $f^{(2)}$, $f^{(1)}$, and f.

5. The only kernels for which $k_q < \infty$ for $q > 2$ are kernels that do not necessarily generate psd estimates. (To see this, note that $k_2 = \int_{-\infty}^{\infty} \lambda^2 K(\lambda)\, d\lambda/2$. Since $k_q < \infty$ for $q > 2$ implies $k_2 = 0$, this implies that $K(\lambda)$ must be negative for some $\lambda \in R$. The discussion of the last paragraph of Section 2 now establishes the assertion.) Thus, the maximal rate of convergence to zero of the truncated MSE for kernels in $\mathscr{K}_2$ is $T^{4/5}$. In contrast, the rate is T for parametric estimators.

6. For asymptotically optimal higher order adjustments to the bandwidth parameters $\{S_T^*\}$, see Andrews (1988, Theorem 4).

6. AUTOMATIC BANDWIDTH ESTIMATORS

This section introduces automatic bandwidth HAC estimators of J_T. These estimators are the same as the kernel estimators of Sections 2–5 except that the bandwidth parameter is a function of the data.

In the density estimation literature, several automatic bandwidth methods have been developed. The two main types are cross-validation (e.g., Beltrao and Bloomfield (1987) and Robinson (1988)) and the "plug-in" method (see Deheuvels (1977) and Sheather (1986)). In the context of spectral density estimation, two additional methods have been suggested by Wahba (1980) and Cameron (1986). Cross-validation and the methods of Wahba and Cameron are suitable if one is interested in estimating a density over an interval, such as the real line, rather than estimating a density at a single point. Hence, they are not well suited to the problem at hand.

Plug-in methods are characterized by the use of an asymptotic formula for an optimal bandwidth parameter (in our case S_T^* of (5.2)) in which estimates are "plugged-in" in place of various unknowns in the formula ($\alpha(q)$ of (5.1)). The estimates that are plugged-in may be parametric or nonparametric. The former yield a less variable bandwidth parameter than the latter, but introduce an asymptotic bias in the estimation of the optimal bandwidth parameter due to the approximate nature of the specified parametric model. (Note that this bias has no effect on the consistency or rate of convergence of the density estimator.)

The automatic bandwidth parameters considered here are of the plug-in type and use parametric estimates. They deviate from the finite sample optimal S_T values due to error introduced by estimation, the use of approximating parametric models, and the approximation inherent in the asymptotic formula employed. Good performance of a HAC estimator, however, only requires the automatic bandwidth parameter to be near the optimal bandwidth value and not precisely equal to it. The reason is that the MSE's of kernel HAC estimators tend to be somewhat U-shape functions of the bandwidth parameter S_T. This is illustrated in Figure 2, which shows the MSE of the QS estimator as a function of S_T for the AR(1)-HOMO model with $\rho = 0.0$, .3, .5, .7, .9, and .95. More

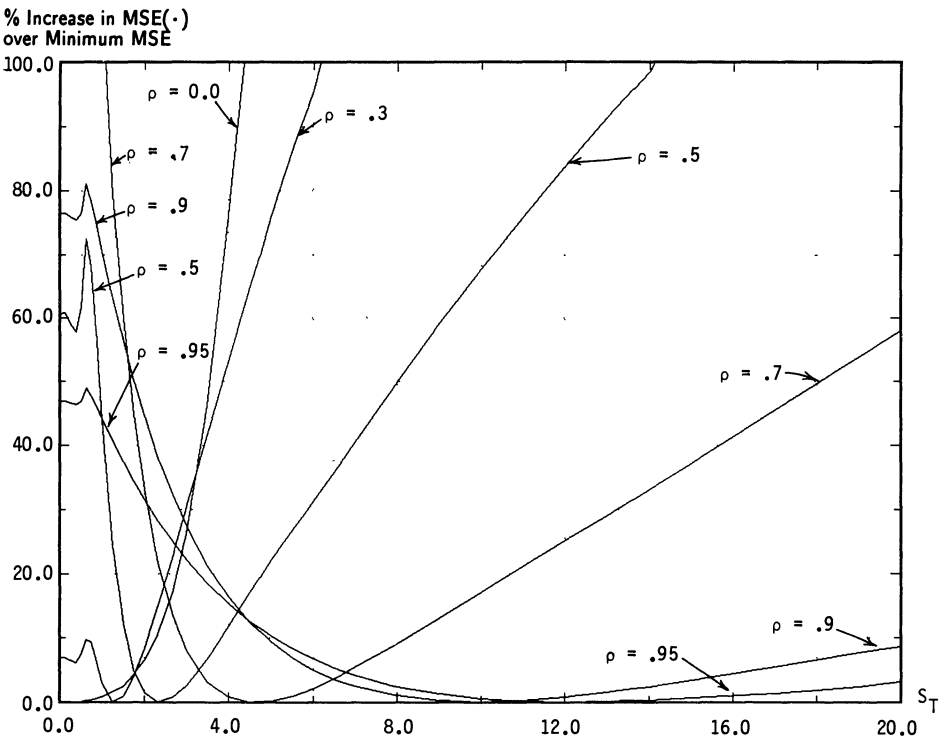

FIGURE 2.—Mean squared error as a function of S_T for the QS estimator in the AR(1)-HOMO model with $\rho = 0.0 - .95$.

precisely, for each value of ρ, Figure 2 graphs the percentage increase in MSE for different S_T values over the minimum MSE over all possible bandwidth values. (As described in Section 9 below, the AR(1)-HOMO model is a linear regression model with regressors and errors that are homoskedastic AR(1) rv's both with AR(1) coefficient ρ.) The automatic bandwidth parameters considered here are designed to produce parameters that are on the flat part of the MSE function even if they are not at the point of minimum MSE.

The automatic bandwidth parameters are defined as follows: First, one specifies p univariate approximating parametric models for $\{V_{at}\}$ for $a = 1, \ldots, p$ (where $V_t = (V_{1t}, \ldots, V_{pt})'$) or one specifies a single multivariate approximating parametric model for $\{V_t\}$. Second, one estimates the parameters of the approximating parametric model(s) by standard methods. Third, one substitutes these estimates into a formula (see below) that expresses $\alpha(q)$ as a function of the parameters of the parametric model(s). This yields an estimate $\hat{\alpha}(q)$ of $\alpha(q)$. $\hat{\alpha}(q)$ is then substituted into the formula (5.2) for the optimal bandwidth parameter S_T^* to yield the automatic bandwidth parameter $\hat{S}_T$:

$$\hat{S}_T = \left(q k_q^2 \hat{\alpha}(q) T \Big/ \int k^2(x)\, dx \right)^{1/(2q+1)}. \tag{6.1}$$

For the kernels of (2.7), we have

(6.2)
$$\begin{aligned}
&\text{Bartlett kernel:} && \hat{S}_T = 1.1447(\hat{\alpha}(1)T)^{1/3},\\
&\text{Parzen kernel:} && \hat{S}_T = 2.6614(\hat{\alpha}(2)T)^{1/5},\\
&\text{Tukey-Hanning kernel:} && \hat{S}_T = 1.7462(\hat{\alpha}(2)T)^{1/5},\\
&\text{Quadratic Spectral kernel:} && \hat{S}_T = 1.3221(\hat{\alpha}(2)T)^{1/5}.^{5,6}
\end{aligned}$$

For general purposes, the suggested approximating parametric models are first order autoregressive (AR(1)) models for $\{V_{at}\}$, $a = 1, \ldots, p$ (with different parameters for each a) or a first order vector autoregressive (VAR(1)) model for $\{V_t\}$. These models are parsimonious. If some other model(s) seem more appropriate for a particular problem, however, they should be used instead. For example, it may be necessary to use models that allow for seasonal patterns or it may be preferable to use first order autoregressive moving average (ARMA(1, 1)) or mth order moving average (MA(m)) models.

The use of p univariate approximating parametric models has advantages of simplicity and parsimony over the use of a single multivariate model, but requires a simple form for the weight matrix W that appears in the formula (5.1) for $\alpha(q)$. In particular, it requires that W give weight only to the diagonal elements of $\hat{J}_T$. Let $\{w_a\colon a = 1, \ldots, p\}$ denote these weights. In this case, (5.1) reduces to

$$\alpha(q) = \sum_{a=1}^{p} w_a \left(f_{aa}^{(q)}\right)^2 \Big/ \sum_{a=1}^{p} w_a f_{aa}^2, \tag{6.3}$$

where $f_{aa}^{(q)}$ and f_{aa} denote the ath diagonal elements of $f^{(q)}$ and f respectively. The usual choice for w_a is one for $a = 1, \ldots, p$ or one for all a except that which corresponds to an intercept parameter and zero for the latter. In linear regression models, the latter choice of weights has the advantage that it yields a scale invariant HAC estimator of the covariance matrix of the LS estimator, provided the estimator $\hat{\alpha}(q)$ (defined below) is scale invariant.

We now provide formulae for $\hat{\alpha}(q)$ for several different approximating parametric models for $\{V_{at}\}$. First, consider AR(1) models for $\{V_{at}\}$. Let (ρ_a, σ_a^2) denote the autoregressive and innovation variance parameters, respectively, for $a = 1, \ldots, p$. Let $\{(\hat{\rho}_a, \hat{\sigma}_a^2)\colon a = 1, \ldots, p\}$ denote the corresponding estimates.

[5] An automatic bandwidth parameter $\hat{S}_T$ is not given in (6.2) for the truncated kernel, because the formula (5.2) for S_T^* does not apply with this kernel. Monte Carlo results, however, show that the formula $\hat{S}_T = .6611(\hat{\alpha}(2)T)^{1/5}$ works quite well for the truncated kernel. This formula is obtained by treating the truncated kernel as though its value of k_2 is finite and equal to the corresponding value for the QS kernel (i.e., $k_2 = k_{2QS}/(\int k^2(x)\,dx)^2 = .3553$).

[6] See Footnote 3 for the relation between the bandwidth parameters S_T and $\hat{S}_T$ used here and the lag truncation parameters as defined by White (1984), Newey and West (1987), Gallant (1987), and Gallant and White (1988).

Then,

$$\hat{\alpha}(2) = \sum_{a=1}^{p} w_a \frac{4\hat{\rho}_a^2\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^8} \Bigg/ \sum_{a=1}^{p} w_a \frac{\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^4} \quad \text{and} \tag{6.4}$$

$$\hat{\alpha}(1) = \sum_{a=1}^{p} w_a \frac{4\hat{\rho}_a^2\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^6(1+\hat{\rho}_a)^2} \Bigg/ \sum_{a=1}^{p} w_a \frac{\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^4}.$$

For ARMA(1,1) models with ARMA parameters (ρ_a, ψ_a) and innovation variance σ_a^2 for $a = 1, \ldots, p$,[7] we have

$$\hat{\alpha}(2) = \sum_{a=1}^{p} w_a \frac{4\left(1+\hat{\rho}_a\hat{\psi}_a\right)^2\left(\hat{\rho}_a+\hat{\psi}_a\right)^2\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^8} \Bigg/ \sum_{a=1}^{p} w_a \frac{\left(1+\hat{\psi}_a\right)^4\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^4} \quad \text{and} \tag{6.5}$$

$$\hat{\alpha}(1) = \sum_{a=1}^{p} w_a \frac{4\left(1+\hat{\rho}_a\hat{\psi}_a\right)^2\left(\hat{\rho}_a+\hat{\psi}_a\right)^2\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^6(1+\hat{\rho}_a)^2} \Bigg/ \sum_{a=1}^{p} w_a \frac{\left(1+\hat{\psi}_a\right)^4\hat{\sigma}_a^4}{(1-\hat{\rho}_a)^4}. \tag{6.6}$$

For MA(m) models with MA parameters $\{\psi_{au}: u = 1, \ldots, m\}$ and innovation variances σ_{au}^2 for $a = 1, \ldots, p$, we have

$$\hat{\alpha}(q) = \frac{\sum_{a=1}^{p} w_a \left[2\sum_{j=1}^{m} j^q \left(\hat{\psi}_{aj} + \sum_{u=1}^{m-j} \hat{\psi}_{au}\hat{\psi}_{au+j}\right)\right]^2 \hat{\sigma}_a^4}{\sum_{a=1}^{p} w_a \left[\sum_{j=-m}^{m} \left(\hat{\psi}_{a|j|} + \sum_{u=1}^{m-|j|} \hat{\psi}_{au}\hat{\psi}_{au+|j|}\right)\right]^2 \hat{\sigma}_a^4}. \tag{6.7}$$

Next, consider a VAR(1) model with $p \times p$ AR parameter matrix A and $p \times p$ innovation covariance matrix Σ. With this multivariate approximating parametric model one can use any psd $p^2 \times p^2$ weight matrix W_T. We have

$$\hat{\alpha}(q) = \frac{2\left(\operatorname{vec} \hat{f}^{(q)}\right)' W_T \operatorname{vec} \hat{f}^{(q)}}{\operatorname{tr} W_T (I + K_{pp})\hat{f} \otimes \hat{f}}, \quad \text{where} \tag{6.8}$$

$$\hat{f} = \frac{1}{2\pi}(I-\hat{A})^{-1}\hat{\Sigma}(I-\hat{A}')^{-1},$$

$$\hat{f}^{(2)} = \frac{1}{2\pi}(I-\hat{A})^{-3}\left(\hat{A}\hat{\Sigma} + \hat{A}^2\hat{\Sigma}\hat{A}' + \hat{A}^2\hat{\Sigma} - 6\hat{A}\hat{\Sigma}\hat{A}' + \hat{\Sigma}(\hat{A}')^2 + \hat{A}\hat{\Sigma}(\hat{A}')^2 + \hat{\Sigma}\hat{A}'\right)(I-\hat{A}')^{-3},$$

$$\hat{f}^{(1)} = \frac{1}{2\pi}(\hat{H} + \hat{H}'), \quad \text{and} \quad \hat{H} = (I-\hat{A})^{-2}\hat{A}\sum_{j=0}^{\infty} \hat{A}^j\hat{\Sigma}(\hat{A}')^j.$$

[7] The ARMA(1,1) model is parametrized as $V_{at} = \rho_a V_{at-1} + \varepsilon_{at} + \psi_a\varepsilon_{at-1}$ with $\operatorname{Var}(\varepsilon_{at}) = \sigma_a^2$. The MA($m$) model considered below is parameterized as $V_{at} = \sum_{u=0}^{m} \psi_{au}\varepsilon_{at-u}$ with $\psi_{a0} = 1$ and $\operatorname{Var}(\varepsilon_{at}) = \sigma_a^2$.

As above, K_{pp} is the $p^2 \times p^2$ commutation matrix; see Magnus and Neudecker (1979).

A natural choice for the weight matrix W_T in this case is

$$W_T = (\hat{B}_T \otimes \hat{B}_T)'\tilde{W}(\hat{B}_T \otimes \hat{B}_T), \tag{6.9}$$

where $\hat{B}_T$ is the estimator defined just below (2.2) and $\tilde{W}$ is an $r^2 \times r^2$ diagonal weight matrix. This choice of W_T corresponds to the loss function

$$\operatorname{vec}(\hat{B}_T\hat{J}_T\hat{B}_T' - \hat{B}_T J_T \hat{B}_T')'\tilde{W} \operatorname{vec}(\hat{B}_T\hat{J}_T\hat{B}_T' - \hat{B}_T J_T \hat{B}_T'),$$

where $\hat{B}_T\hat{J}_T\hat{B}_T'$ is the covariance matrix estimator of $\hat{\theta}$. If $\hat{B}_T - B_T = o_p(S_T/T)$ (as is usually the case, since $\hat{B}_T$ is usually a $\sqrt{T}$-consistent parametric estimator of B_T), then the asymptotic truncated MSE with this loss function is the same as when $\hat{B}_T J_T \hat{B}_T'$ is replaced by $B_T J_T B_T'$ in the loss function. Thus, the choice of loss function (6.9) for $\hat{J}_T$ gives the asymptotic truncated MSE of $\hat{B}_T\hat{J}_T\hat{B}_T'$ for estimating $B_T J_T B_T'$ with weight matrix $\tilde{W}$. The diagonal matrix $\tilde{W}$ should be chosen to suitably weight the elements of $\hat{B}_T\hat{J}_T\hat{B}_T'$. For example, to give equal weight to each nonredundant element of $\hat{B}_T\hat{J}_T\hat{B}_T'$, one takes $\tilde{W}$ to have ones for diagonal elements that correspond to nondiagonal elements $\hat{B}_T\hat{J}_T\hat{B}_T'$ and twos for diagonal elements that correspond to diagonal elements of $\hat{B}_T\hat{J}_T\hat{B}_T'$.

Last, consider a VAR(m) model $V_t = \sum_{j=1}^m A_j V_{t-j} + \Sigma_t$, where $\{A_j\colon\ j = 1, \ldots, m\}$ are $p \times p$ parameter matrices and Σ_t is a $p \times 1$ innovation vector with covariance matrix Σ. In this case, $\hat{\alpha}(2)$ is as in (6.8) with $q = 2$,

$$\hat{f} = (1/2\pi)\Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1} \hat{\Sigma} \Big(I - \sum_{j=1}^m \hat{A}_j'\Big)^{-1}, \quad \text{and}$$

$$\hat{f}^{(2)} = \frac{1}{2\pi}(\hat{M}_1 + \hat{M}_2 + \hat{M}_2'), \quad \text{where}$$

$$\hat{M}_1 = -2\Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1}\Big(\sum_{j=1}^m j\hat{A}_j\Big)\Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1} \hat{\Sigma}\Big(I - \sum_{j=1}^m \hat{A}_j'\Big)^{-1} \times \Big(\sum_{j=1}^m j\hat{A}_j'\Big)\Big(I - \sum_{j=1}^m \hat{A}_j'\Big)^{-1},$$

$$\hat{M}_2 = \Big[2\Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1}\Big(\sum_{j=1}^m j\hat{A}_j\Big)\Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1} \times \Big(\sum_{j=1}^m j\hat{A}_j\Big)\Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1} + \Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1} \times \Big(\sum_{j=1}^m j^2\hat{A}_j\Big)\Big(I - \sum_{j=1}^m \hat{A}_j\Big)^{-1}\Big] \hat{\Sigma}\Big(I - \sum_{j=1}^m \hat{A}_j'\Big)^{-1}.$$

(The latter formulae were provided by Ken Vetzal.)

The choice between using p univariate approximating parametric models (such as AR(1) models) or a single multivariate model (such as a VAR(1)

model) depends upon a tradeoff between simplicity and parsimony on one hand and flexibility in the choice of weight matrix on the other. For the Monte Carlo results of Section 9 below, AR(1) univariate approximating models are used.

In practice, the value of a HAC estimator can be sensitive to the choice of the bandwidth parameter. Hence, it often is wise to calculate several bandwidth values centered about the automatic bandwidth value given by (6.1) in order to assess the degree of sensitivity of the estimator. These additional bandwidth values can be chosen by replacing the estimated parameters of the approximating parametric models used in (6.1) by the estimated parameters plus or minus one or two standard deviations of their values. For example, with AR(1) approximating models, one would replace $\hat{\rho}_a$ by $\hat{\rho}_a \pm 1/\sqrt{T}$ or $\hat{\rho}_a \pm 2/\sqrt{T}$.

7. PROPERTIES OF THE AUTOMATIC BANDWIDTH ESTIMATORS

In this section, we establish consistency, rate of convergence, and asymptotic truncated MSE results for kernel HAC estimators that are constructed using the automatic bandwidth parameters $\{\hat{S}_T\}$ introduced in Section 6.

The results of this section apply to kernels in the following class:

$$\begin{aligned}\mathscr{K}_3 = \{k(\cdot) \in \mathscr{K}_1 &: \text{(i) } |k(x)| \leqslant C_1|x|^{-b} \text{ for some } b > 1 + 1/q \\ &\text{and some } C_1 < \infty, \text{ where } q \in (0,\infty) \text{ is such that} \\ &k_q \in (0,\infty), \text{ and (ii) } |k(x) - k(y)| \leqslant C_2|x-y| \ \ \forall x, \\ &y \in R \text{ for some constant } C_2 < \infty\}.\end{aligned} \tag{7.1}$$

This class contains the Bartlett, Parzen, Tukey-Hanning, and QS kernels, but not the truncated kernel, because the latter does not satisfy the Lipschitz condition.

For consistency of $\hat{J}_T(\hat{S}_T)$, $\hat{\alpha}(q)$ only needs to satisfy the following assumption.

ASSUMPTION E: $\hat{\alpha}(q) = O_p(1)$ *and* $1/\hat{\alpha}(q) = O_p(1)$.

For rate of convergence and asymptotic truncated MSE results, stronger conditions on $\hat{\alpha}(q)$ are needed. Let $\hat{\xi}$ denote the estimator of the parameter of the approximating parametric model(s) introduced in Section 6. (For example, with univariate AR(1) approximating parametric models, $\hat{\xi} = (\hat{\rho}_1, \hat{\sigma}_1^2, \ldots, \hat{\rho}_p, \hat{\sigma}_p^2)'$.) Let ξ denote the probability limit of $\hat{\xi}$. $\hat{\alpha}(q)$ is the value of $\alpha(q)$ that corresponds to $\hat{\xi}$. The probability limit of $\hat{\alpha}(q)$ depends on ξ and is denoted α_ξ. For the results referred to above, we make the following assumption.

ASSUMPTION F: $\sqrt{T}(\hat{\alpha}(q) - \alpha_\xi) = O_p(1)$ *for some* $\alpha_\xi \in (0,\infty)$.

Note that α_ξ equals the optimal value $\alpha(q)$ if the approximating parametric model indexed by ξ actually is correct. In general, however, α_ξ deviates from $\alpha(q)$.

The fixed bandwidth sequence that is closest to $\{\hat{S}_T\}$ is defined by replacing $\hat{\alpha}(q)$ by α_ξ in the definition of $\hat{S}_T$. Let

$$S_{\xi T} = \left(qk_q^2 \alpha_\xi T \Big/ \int k^2(x)\,dx\right)^{1/(2q+1)}. \tag{7.2}$$

The asymptotic properties of $\hat{J}_T(\hat{S}_T)$ are shown to be equivalent to those of $\hat{J}_T(S_{\xi T})$.

For the rate of convergence and asymptotic truncated MSE results, we also require the following assumption. Let $\lambda_{\max}(A)$ denote the largest eigenvalue of the matrix A.

ASSUMPTION G: $\lambda_{max}(\Gamma(j)) \leqslant C_3 j^{-m}$ $\forall j \geqslant 0$, *for some* $C_3 < \infty$ *and some* $m > \max\{2, 1 + 2q/(q+2)\}$, *where* q *is as in* $\mathscr{K}_3$.

If $\{V_t\}$ is strong mixing with mixing numbers of size $-\max\{2, 1 + 2q/(q+2)\} \times \nu/(\nu - 1/2)$ for some $\nu > 1$ such that $\sup_{t \geqslant 1} E\|V_t\|^{4\nu} < \infty$, then Assumption G holds. In particular, in the cases of interest $q \leqslant 2$, so the size condition is $-3\nu/(\nu - 1/2)$. This is less stringent than the size condition $-3\nu/(\nu - 1)$ which is sufficient for Assumption A.

The main result of this section is the following:

THEOREM 3: *Suppose* $k(\cdot) \in \mathscr{K}_3$, q *is as in* $\mathscr{K}_3$, *and* $\|f^{(q)}\| < \infty$.

(a) *If Assumptions A, B, and E hold and* $q > 1/2$, *then* $\hat{J}_T(\hat{S}_T) - J_T \to^p 0$.

(b) *If Assumptions B, C, F, and G hold, then* $\sqrt{T/S_{\xi T}}(\hat{J}_T(\hat{S}_T) - J_T) = O_p(1)$ *and* $\sqrt{T/S_{\xi T}}(\hat{J}_T(\hat{S}_T) - \hat{J}_T(S_{\xi T})) \to^p 0$.

(c) *If Assumptions B–D, F, and G hold, then*

$$\begin{aligned}
&\lim_{h\to\infty}\lim_{T\to\infty} \mathrm{MSE}_h\left(T/S_{\xi T}, \hat{J}_T(\hat{S}_T), W_T\right) \\
&\quad = \lim_{h\to\infty}\lim_{T\to\infty} \mathrm{MSE}_h\left(T/S_{\xi T}, \hat{J}_T(S_{\xi T}), W_T\right) \\
&\quad = 4\pi^2\Big(k_q^2(\mathrm{vec}\, f^{(q)})' W \,\mathrm{vec}\, f^{(q)}/\gamma_\xi \\
&\qquad\qquad + \int k^2(x)\,dx \,\mathrm{tr}\, W(I + K_{pp}) f \otimes f\Big),
\end{aligned}$$

where $\gamma_\xi = qk_q^2\alpha_\xi / \int k^2(x)\,dx$.

If $\hat{\alpha}(q) \to^p \alpha(q)$ (i.e., $\alpha_\xi = \alpha(q)$), as occurs if the approximate parametric model indexed by ξ is correct, then $\{\hat{S}_T\}$ exhibits some optimality properties as a result of Theorem 1(c) and Corollary 1. In particular, given a kernel $k(\cdot) \in \mathscr{K}_3$, let $\{\dot{S}_T\}$ be any sequence of automatic bandwidth parameters such that for some fixed sequence $\{S_T\}$, which satisfies $S_T^{2q+1}/T \to \gamma$ for some $\gamma \in (0, \infty)$, we have

$$\begin{aligned}
\lim_{h\to\infty}\lim_{T\to\infty} \Big(&\mathrm{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(\dot{S}_T), W_T\right) \\
&- \mathrm{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(S_T), W_T\right)\Big) = 0.
\end{aligned} \tag{7.3}$$

Then, $\{\hat{S}_T\}$ is preferred to $\{\dot{S}_T\}$:

COROLLARY 2: *Suppose Assumptions B–D, F, and G hold. Consider a kernel* $k(\cdot) \in \mathscr{K}_3$. *Let* q *be as in* $\mathscr{K}_3$. *Suppose* $\|f^{(q)}\| < \infty$, $\alpha(q) \in (0, \infty)$, *and* W *is psd. Let* $\{\dot{S}_T\}$ *be any sequence of automatic bandwidth parameters that satisfies* (7.3). *If* $\alpha_\xi = \alpha(q)$ (*i.e., if* $\hat{\alpha}(q)$ *converges in probability to the optimal value* $\alpha(q)$), *then* $\{\hat{S}_T\}$ *is preferred to* $\{\dot{S}_T\}$ *in the sense that*

$$\lim_{h \to \infty} \lim_{T \to \infty} \left(\mathrm{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(\dot{S}_T), W_T\right) - \mathrm{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(\hat{S}_T), W_T\right)\right) \geqslant 0.$$

The inequality is strict unless $S_T = S_T^* + o(T^{1/(2q+1)})$.

8. EXTENSION TO NONSTATIONARY RANDOM VARIABLES

Thus far this paper has considered unconditionally weakly stationary rv's (Assumption A). We now make note of sufficient conditions for the consistency and rate of convergence results of the paper to hold for unconditionally heteroskedastic rv's. Here we do not discuss asymptotic truncated MSE results or optimality results for kernels and bandwidth sequences for unconditionally nonstationary rv's. Such results can be found in Andrews (1988). They use lower and upper bounds on the MSE and a minimax MSE criterion for optimality.

Consider the following generalizations of Assumptions A, C, and G:

ASSUMPTION A*: $\{V_t\}$ *is a mean zero sequence of rv's with*

$$\sum_{j=0}^{\infty} \sup_{t \geqslant 1} \|EV_t V_{t+j}'\| < \infty \quad \textit{and}$$

$$\sum_{j=1}^{\infty} \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} \sup_{t \geqslant 1} |\kappa_{abcd}(t, t+j, t+m, t+n)| < \infty \quad \forall a, b, c, d \leqslant p.$$

ASSUMPTION C*: *Assumption C holds with reference to Assumption A* rather than Assumption A in part* (*i*).

ASSUMPTION G*: *Assumption G holds with* $\lambda_{\max}(\Gamma(j))$ *replaced by* $\sup_{t \geqslant 1} \lambda_{\max}(EV_t V_{t+j}')$.

By Lemma 1, Assumption A* holds if $\{V_t\}$ is a mean zero α-mixing sequence of rv's with $\sup_{t \geqslant 1} E\|V_t\|^{4\nu} < \infty$ and $\sum_{j=1}^{\infty} j^2 \alpha(j)^{(\nu-1)/\nu} < \infty$ for some $\nu > 1$.

If $V_t(\theta)$ is of the form $V(Z_t, \theta)$ for some rv Z_t and some (measurable) function $V(\cdot, \cdot)$, then Assumptions A* and C* hold if (i) $EV_t = 0$ $\forall t \geqslant 1$, (ii) $\{Z_t: t \geqslant 1\}$ is α-mixing with

$$\sum_{j=1}^{\infty} j^2 \alpha(j)^{(\nu-1)/\nu} < \infty \text{ and } \sup_{t \geqslant 1} \left(E\|V_t\|^{4\nu} + E\|(\partial/\partial\theta')\, V_t(\theta_0)\|^{4\nu}\right) < \infty$$

for some $\nu > 1$, and (iii) $\sup_{t \geqslant 1} E \sup_{\theta \in \Theta} \|\partial^2 V_{at}(\theta)/\partial\theta\, \partial\theta'\|^2 < \infty$ $\forall a = 1, \ldots, p$. If, in addition, $q \leqslant 2$ in Assumption G* (as is the case for the QS, Parzen, Tukey-Hanning, and Bartlett kernels), then Assumption G* also holds.

Now, under Assumption A* rather than A, Proposition 1 continues to hold by Lemmas 1 and 2 and Theorem 1 of Andrews (1988) with the following changes: In part (a), $\lim_{T\to\infty}, \operatorname{vec} \tilde{J}_T$, $=$, and $(I+K_{pp})f\otimes f$ are replaced by $\overline{\lim}_{T\to\infty}$, $b'\tilde{J}_T b$, $\leqslant$, and $2(b'fb)^2$, respectively, for arbitrary $b\in R^p$. In part (b), $\|f^{(q)}\|$, $\lim_{T\to\infty}$, $(E\tilde{J}_T - J_T)$, $=$, and $f^{(q)}$ are replaced by $(1/2\pi)\sum_{j=-\infty}^{\infty}|j|^q \sup_{t\geqslant 1}|Eb'V_tV'_{t+|j|}b|$, $\overline{\lim}_{T\to\infty}$, $|Eb'\tilde{J}_T b - b'J_T b|$, $\leqslant$, and $(1/2\pi)\sum_{j=-\infty}^{\infty}|j|^q \sup_{t\geqslant 1}|Eb'V_tV'_{t+|j|}b|$, respectively, for arbitrary $b\in R^p$. In part (c), $\|f^{(q)}\|$ is changed as above and the result is changed to: $\forall b\in R^p$,

$$\overline{\lim_{T\to\infty}}\ \frac{T}{S_T}E\left(b'\tilde{J}_T b - b'J_T b\right)^2$$

$$\leqslant 4\pi^2\left(k_q^2\left[\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}|j|^q \sup_{t\geqslant 1}\left|Eb'V_tV'_{t+|j|}b\right|\right]^2\Big/\gamma + \int k^2(x)\,dx\, 2(b'fb)^2\right).$$

Theorem 1(a) continues to hold with Assumption A replaced by Assumption A*. Theorem 1(b) continues to hold with Assumption C replaced by Assumption C* and $\|f^{(q)}\|$ replaced by $(1/2\pi)\sum_{j=-\infty}^{\infty}|j|^q \sup_{t\geqslant 1}\|EV_tV'_{t+j}\|$. The proof of these results is a trivial extension of the proof of Theorem 1(a) and (b) in the Appendix using the results of the previous paragraph.

Theorem 3(a) and (b) continues to hold if $f^{(q)}$ and Assumptions A, C, and G are replaced by $(1/2\pi)\sum_{j=-\infty}^{\infty}|j|^q \sup_{t\geqslant 1}\lambda_{\max}(EV_tV'_{t+|j|})$ and Assumptions A*, C*, and G* respectively. Using the results of the previous paragraph and Lemmas 1 and 2 of Andrews (1988), the proof is a straightforward extension of that given in the Appendix for Theorem 3(a) and (b). Thus, automatic bandwidth kernel estimators are consistent with nonstationary as well as fourth order stationary rv's.

9. MONTE CARLO RESULTS

In this section, simulation methods are used to evaluate the asymptotic results obtained in Sections 3–8. In particular, we are interested in evaluating the results of Theorem 2 regarding the optimal kernel and of Theorem 3 and Corollary 2 regarding automatic bandwidth parameters.

The models we consider are linear regression models, each with an intercept and four regressors; see (2.1). The estimand of interest is the variance of the LS estimator of the first nonconstant regressor. (That is, the estimand is the second diagonal element of $\operatorname{Var}(\sqrt{T}(\hat{\theta}-\theta_0))$ in (2.1).) Four basic regression models are considered: AR(1)-HOMO, in which the errors and regressors are homoskedastic AR(1) processes; AR(1)-HET1 and AR(1)-HET2, in which the errors and regressors are AR(1) processes with multiplicative heteroskedasticity overlaid on the errors; and MA(1)-HOMO, in which the errors and regressors are homoskedastic MA(1) processes. (Details are given below.) A range of six to

eight parameter values are considered for each model. Each parameter value corresponds to a different degree of autocorrelation.

Estimators based on the five kernels of (2.7) are evaluated. They are: truncated (TRUNC), Bartlett (BART), Parzen (PARZ), Tukey-Hanning (TUK), and quadratic spectral (QS). The performance of each kernel estimator is determined for a variety of different bandwidths. These bandwidths include the asymptotically optimal bandwidth of (5.2), the automatic bandwidth of (6.1) based on univariate AR(1) approximating models with (ρ_a, σ_a^2) estimated by LS for each a, and a grid of fixed bandwidths that are used to obtain the finite sample optimal bandwidth. For the former two bandwidths, the weights $\{w_a\}$ are taken to be zero for the intercept and one for the others.

For comparative purposes, three estimators are considered in addition to the kernel estimators described above: the heteroskedasticity consistent estimator of Eicker (1967) and White (1980), denoted INID; the standard LS variance estimator for iid errors, denoted IID; and a parametric estimator that assumes that the errors are homoskedastic AR(1) random variables, denoted PARA. More specifically,

$$\text{(9.1)}\qquad \text{INID} = \left[\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\left(\frac{1}{T-5}\sum_{t=1}^{T}\hat{U}_t^2 X_t X_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\right]_{22},$$

$$\text{IID} = \left(\frac{1}{T-5}\sum_{t=1}^{T}\hat{U}_t^2\right)\left[\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\right]_{22}, \qquad \text{and}$$

$$\text{PARA} = \left[\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\left(\frac{1}{T-5}\sum_{t=1}^{T}\hat{U}_t^2\right)\cdot\left(\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\hat{\rho}^{|s-t|} X_s X_t'\right)\left(\frac{1}{T}\sum_{t=1}^{T} X_t X_t'\right)^{-1}\right]_{22},$$

where $\hat{U}_t = Y_t - X_t'\hat{\theta}$, $\hat{\rho}_{LS}$ is the LS estimator of ρ from the regression of $\hat{U}_t$ on $\hat{U}_{t-1}$ for $t = 2, \ldots, T$, $\hat{\rho} = \min(.97, \hat{\rho}_{LS})$, and $[\cdot]_{22}$ denotes the (2,2) element of $\cdot$.[8]

For each variance estimator and each scenario, the following performance criteria are estimated by Monte Carlo simulation: (i) the exact bias, variance, mean squared error (MSE), and mean absolute error (MAE) of the variance estimator and (ii) the true confidence levels of the nominal 99%, 95%, and 90% regression coefficient confidence intervals (CI's) based on the t statistic constructed using the LS coefficient estimator and the variance estimator. (The nominal $100(1-\alpha)\%$ CI's are based on an asymptotic normal approximation. For the INID, IID, and PARA estimators, this normal approximation is not valid asymptotically in some of the scenarios under consideration.) The control

[8] The truncated estimator $\hat{\rho}$, rather than $\hat{\rho}_{LS}$, is used to construct PARA because we do not want the performance of PARA to be dominated by a few observations for which $\hat{\rho}_{LS}$ is near or greater than one. Since $\hat{\rho}_{LS}$ has a large downward bias when ρ is large (say .9 or .95), the truncation at .97 occurs seldomly even when ρ is large.

variate method of Davidson and MacKinnon (1981) is used to estimate the true confidence levels in (ii). Sample sizes of 64, 128, and 256 are investigated. One thousand repetitions are used for each scenario.

The distributions of all of the variance estimators considered here are invariant with respect to the regression coefficient vector θ_0 in the model. Hence, we set $\theta_0 = \mathbf{0}$ in each model and do so without loss of generality.

Next we describe the four models used in the Monte Carlo study. The AR(1)-HOMO model consists of mutually independent errors and regressors. The errors are mean zero, homoskedastic, stationary, AR(1), normal random variables with variance 1 and AR parameter ρ. The four regressors are generated by four independent draws from the same distribution as that of the errors, but then are transformed to achieve a diagonal $(1/T)\Sigma_{t=1}^{T} X_t X_t'$ matrix.[9] The values considered for the AR(1) parameter ρ are 0, .3, .5, .7, .9, .95, $-.3$, and $-.5$.

The AR(1)-HET1 and AR(1)-HET2 models are constructed by introducing multiplicative heteroskedasticity to the errors of the AR(1)-HOMO model. Suppose $\{x_t, \tilde{U}_t\colon t = 1, \ldots, T\}$ are the nonconstant regressors and errors generated by the AR(1)-HOMO model (where $X_t = (1, x_t')'$). Let $U_t = |x_t'\zeta| \times \tilde{U}_t$. Then, $\{x_t, U_t\colon t = 1, \ldots, T\}$ are the nonconstant regressors and errors for the AR(1)-HET1 and AR(1)-HET2 models when $\zeta = (1, 0, 0, 0)'$ and $\zeta = (1/2, 1/2, 1/2, 1/2)'$ respectively. In the AR(1)-HET1 model, the heteroskedasticity is related only to the regressor whose coefficient estimator's variance is being estimated, whereas in the AR(1)-HET2 model, the heteroskedasticity is related to all of the regressors.[10] The same values of ρ are considered as in the AR(1)-HOMO model.

The MA(1)-HOMO model is exactly the same as the AR(1)-HOMO model except that the errors and the (pretransformed) regressors are homoskedastic, stationary, MA(1) random variables with variance 1 and MA parameter ψ. The values of ψ that are considered are .3, .5, .7, .99, $-.3$, and $-.7$.

The first table of simulation results, Table II, provides a comparison of the five kernels of (2.7). The table presents ratios of the finite sample MSE's of the TRUNC, BART, PARZ, and TUK estimators to those of the QS estimator for each model scenario and $T = 128$. Each estimator has its bandwidth parameter

[9] The transformation used is described as follows. Let $\tilde{x}$ denote the $T \times 4$ matrix of pretransformed, randomly generated, AR(1) regressor variables. Let $\bar{x}$ denote $\tilde{x}$ with its column means subtracted off. Let $x = \bar{x}(\bar{x}'\bar{x}/T)^{-1/2}$. Define the $T \times 5$ matrix of transformed regressors to be $X = [1_T\colon x]$. By construction, $X'X = TI_5$.

Since $E\tilde{x} = \mathbf{0}$ and $E\tilde{x}'\tilde{x} = I_4$, this transformation is close to the identity map with high probability. With this transformation, the estimand and the estimators simplify and the computational burden is reduced considerably. The estimand becomes just the product of the second diagonal elements of the three 5×5 matrices multiplied together in (2.1). Two of these diagonal elements are known—only one has to be estimated, viz., the second diagonal element of the J_T matrix. Without the transformation, one has to compute all twenty-five elements of the estimated J_T matrix, rather than a single element, in order to compute the performance criteria described above.

[10] When the regressor transformation map is the identity map, the errors in the AR(1)-HET1 and AR(1)-HET2 models are mean zero, variance one, AR(1) sequences with AR parameter ρ^2 and innovations that are uncorrelated (unconditionally and conditionally on $\{X_t\}$) but not independent. Hence, the errors have an AR(1) correlation structure even after the introduction of heteroskedasticity.

TABLE II

RATIO OF MSE OF TRUNCATED, BARTLETT, PARZEN, AND TUKEY-HANNING ESTIMATORS TO MSE OF QS ESTIMATOR USING FINITE SAMPLE OPTIMAL S_T VALUES – $T = 128$

		ρ							
Model	Estimator	0	.3	.5	.7	.9	.95	−.3	−.5
AR(1)-HOMO	TRUNC	1.00	1.09	.93	.93	.95	.97	1.09	.94
	BART	1.00	1.00	1.05	1.09	1.06	1.04	1.01	1.05
	PARZ	1.00	1.01	1.01	1.02	1.01	1.01	1.01	1.01
	TUK	1.00	1.00	1.00	1.01	1.00	1.00	1.01	1.00
AR(1)-HET1	TRUNC	1.00	1.03	.98	.97	.97	.98	1.02	1.13
	BART	1.00	1.00	1.02	1.04	1.03	1.02	1.02	1.13
	PARZ	1.00	1.00	1.01	1.01	1.01	1.01	1.02	1.13
	TUK	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.13
AR(1)-HET2	TRUNC	1.00	1.00	1.07	.98	.96	.98	1.00	1.09
	BART	1.00	1.00	1.00	1.03	1.04	1.03	1.00	1.00
	PARZ	1.00	1.00	1.00	1.00	1.01	1.01	1.00	1.00
	TUK	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		ψ							
		.3	.5	.7	.99	−.3	−.7		
MA(1)-HOMO	TRUNC	1.04	1.02	.99	.99	1.02	.98		
	BART	.99	.99	1.04	1.05	.99	1.02		
	PARZ	.99	.99	1.01	1.02	.99	1.00		
	TUK	.99	.97	1.00	1.00	.99	.99		

set equal to its nonrandom finite sample optimal value (determined by grid search) to ensure comparability of the kernels.

The table shows that the QS estimator is slightly more efficient than the PARZ estimator and very slightly more efficient than the TUK estimator in the scenarios considered. These results are basically consistent with the asymptotic results for kernel comparisons given in Theorem 2 and Corollary 1 Comment 4. The finite sample advantage of the QS kernel over the PARZ kernel, however, is clearly less than its asymptotic advantage. For these kernels, results corresponding to those of Table II, but for sample sizes $T = 64$ and $T = 256$, are quite similar to those of Table II.

In Table II, the three estimators QS, PARZ, and TUK consistently exhibit a distinct, but not large, advantage over the BART estimator. This advantage is predicted by the asymptotic results of Theorem 1 (also see Corollary 1 Comment 4). It is interesting to note that for sample size $T = 256$ (not reported here), the MSE advantage of the QS, PARZ, and TUK estimators over the BART estimator is more pronounced than in Table II where $T = 128$. This is expected given the asymptotic results.

For all of the estimators, the results of Table II are not changed much when the MSE criterion is replaced by the MAE criterion. The only change is that the differences between the estimators are somewhat less pronounced.

The TRUNC estimator exhibits wide fluctuations in its MSE relative to that of the QS estimator and the other three estimators. In the AR(1)-HOMO model, it ranges from being 9% less efficient to 7% more efficient than the QS

estimator. For most scenarios, however, it is more efficient than the QS estimator. This is what is suggested by the asymptotic results (see Proposition 1(b) and Theorem 1(c)), since the bias of the TRUNC estimator declines at a faster rate than it does for the other estimators. Results corresponding to Table II but with sample sizes $T = 64$ and $T = 256$ show that the relative efficiency of the TRUNC estimator is increasing with T (i.e., the ratios of MSE's are declining) in most scenarios, but at a fairly slow rate.

Comparisons of the true confidence levels of the CI's constructed using the five different variance estimators are not given in the tables, because they are quite similar to the comparisons based on MSE's given in Table II. In all cases, the true confidence levels of the CI's fall short of their nominal confidence levels. Thus, the best CI's are the ones whose confidence levels are the largest. Of the BART, PARZ, TUK, and QS-based CI's, the QS-based CI's are fairly consistently the best, but only by a slight margin over the PARZ and TUK-based CI's. The margin is larger with respect to the BART-based CI's. There are two reasons why the BART-based CI's do worse than the other CI's. First, the BART variance estimator has greater MSE's than do the other estimators, and second, its squared bias-variance ratio is significantly larger than that of the other estimators in most cases. The latter property is to be expected given the asymptotics (see Corollary 1 Comment 3).

The true confidence level results for the TRUNC-based CI's are similar to the TRUNC estimator's MSE results. In some scenarios they are the best and in some scenarios they are the worst. The scenarios in which they are best and worst are the same scenarios where the TRUNC estimator has lowest and highest MSE's, respectively, in Table II.

One drawback of the TRUNC estimator (as well as the TUK estimator) is that it does not necessarily generate nonnegative variance estimates. In the Monte Carlo experiments, however, a significant number of negative estimates arise only when there is very heavy autocorrelation. For example, in the AR(1)-HOMO model with $\rho = .95$, the percentages of negative TRUNC estimates are 7.6, 1.2, and 0 for $T = 64$, 128, and 256, respectively (using the finite sample optimal bandwidth parameter). For smaller values of ρ and for the TUK estimator, the percentages are zero for all sample sizes considered.

For brevity, we only discuss results for the QS estimator in the remainder of this section. For the most part, in the tables that follow, the relative performances of the other kernel estimators in comparison with the QS estimator follow patterns similar to those observed in Table II. Tables analogous to those given here, but including the other kernel estimators, are available from the author upon request.

Table III assesses the performance of the automatic bandwidth procedure $\hat{S}_T$ of (6.1). In all scenarios, the approximating parametric models used by the automatic bandwidth procedure are univariate AR(1) models.

Table III shows that in general the $\hat{S}_T$ bandwidth values work very well. This is true in both the homoskedastic and heteroskedastic cases. The $\hat{S}_T$ values work much better with positive serial dependence than with negative serial

TABLE III

RATIO OF MSE OF QS ESTIMATOR USING AUTOMATIC S_T VALUE, $\hat{S}_T$, TO MSE OF QS ESTIMATOR USING FINITE SAMPLE OPTIMAL S_T VALUE

Model	T	ρ 0	.3	.5	.7	.9	.95	−.3	−.5
AR(1)-HOMO	64	1.09	1.16	1.07	1.02	1.01	1.01	1.17	1.09
	128	1.05	1.14	1.14	1.05	1.01	1.01	1.23	1.12
	256	1.06	1.10	1.05	1.06	1.03	1.01	1.14	1.07
AR(1)-HET1	64	1.12	1.02	1.00	1.01	1.01	1.01	1.45	3.05
	128	1.10	1.02	1.02	1.02	1.01	1.01	1.68	4.18
	256	1.01	1.03	1.03	1.04	1.03	1.01	1.93	5.17
AR(1)-HET2	64	1.06	1.13	1.07	1.02	1.01	1.01	1.20	1.34
	128	1.05	1.16	1.17	1.04	1.03	1.01	1.18	1.10
	256	1.07	1.23	1.07	1.12	1.01	1.02	1.22	1.22
		ψ .3	.5	.7	.99	−.3	−.7		
MA(1)-HOMO	64	1.15	1.05	1.12	1.14	1.15	1.23		
	128	1.02	1.16	1.17	1.32	1.11	1.21		
	256	1.05	1.21	1.28	1.47	1.06	1.29		

dependence. No clear improvement or deterioration of the MSE ratios occurs as T increases from 64 to 128 to 256.

The analogue of Table III (not reported here) that uses true confidence levels rather than MSE's as the performance criterion puts the automatic bandwidth parameter $\hat{S}_T$ in an even better light than does Table III. In virtually every case, the use of $\hat{S}_T$ incurs only a small reduction in the true confidence level from the true level obtained using the best fixed S_T value. (The latter confidence level, in turn, is always less than or equal to the nominal level.) For example, in most scenarios, the reduction in the confidence level for the nominal 95% CI's is in the range of 0 to 1%.

In conclusion, the automatic bandwidth procedure $\hat{S}_T$ performs quite well in terms of MSE and true confidence levels in comparison with the optimal finite sample bandwidth (in the models considered).

Tables IV–VI aim to show how well kernel HAC estimators perform in comparison with other types of variance estimators, viz., INID, IID, and PARA. The kernel estimator used for all three tables is the QS estimator with the automatic bandwidth parameter $\hat{S}_T$ discussed above. The results for other kernels and other bandwidth choices (such as S_T^* and the finite sample optimal S_T value) can be deduced reasonably well from the comparative results given above.

Table IV presents detailed results for the AR(1)-HOMO model with sample size $T = 128$. Table V presents analogous, but less detailed, results for a subset of parameter values in the AR(1)-HET1, AR(1)-HET2, and MA(1)-HOMO models with $T = 128$. Table VI presents a selected set of results for all four models with $T = 256$.

TABLE IV

BIAS, VARIANCE, AND MSE OF QS ESTIMATOR WITH AUTOMATIC S_T VALUE, $\hat{S}_T$, AND TRUE CONFIDENCE LEVELS OF NOMINAL 99%, 95%, AND 90% CONFIDENCE INTERVALS CONSTRUCTED USING THE QS ESTIMATOR WITH AUTOMATIC S_T VALUE FOR THE AR(1)-HOMO MODEL – $T = 128$

ρ	Value of Estimand	Estimator	Bias	Variance	MSE	99%	95%	90%
0	1.00	QS	−.050	.045	.047	98.2	93.9	88.0
		INID	−.048	.043	.045	98.1	93.8	88.3
		IID	.0040	.016	.016	98.5	94.5	89.4
		PARA	.0045	.017	.017	98.5	94.5	89.5
.3	1.18	QS	−.15	.088	.11	97.7	91.5	85.5
		INID	−.24	.044	.10	97.4	90.8	83.9
		IID	−.19	.018	.56	97.9	92.0	86.4
		PARA	−.032	.037	.038	98.9	94.0	88.9
.5	1.59	QS	−.31	.25	.34	97.2	89.7	83.3
		INID	−.68	.050	.51	94.5	84.6	76.8
		IID	−.62	.026	.41	95.3	86.2	78.9
		PARA	−.095	.14	.15	98.8	94.1	87.5
.7	2.65	QS	−.88	.66	1.44	94.6	86.5	79.4
		INID	−1.81	.057	3.32	84.6	73.5	64.4
		IID	−1.74	.037	3.07	85.7	75.9	66.4
		PARA	−.43	.56	.75	97.1	91.2	84.8
.9	6.41	QS	−3.94	3.43	19.0	85.9	74.9	65.9
		INID	−5.79	.071	33.6	56.2	44.2	38.0
		IID	−5.72	.059	32.8	60.1	47.2	39.7
		PARA	−2.96	4.23	13.0	92.8	82.7	75.7
.95	8.62	QS	−6.52	3.31	45.8	74.3	63.1	55.8
		INID	−8.18	.059	67.0	43.8	33.5	27.9
		IID	−8.13	.065	66.1	47.1	35.6	29.8
		PARA	−5.58	5.01	36.1	83.1	72.2	65.0

The first feature of note in Tables IV–VI is that the QS estimator basically dominates INID, and PARA basically dominates IID, over all model scenarios. When ρ or ψ equals zero, INID and IID are at most slightly better than QS and PARA, respectively. When ρ or ψ is nonzero, QS and PARA usually are distinctly superior to INID and IID, respectively. Thus, when no autocorrelation is present, one pays a small price for using a HAC estimator with an automatic bandwidth parameter rather than a heteroskedasticity consistent estimator of the Eicker-White form. On the other hand, when autocorrelation is present, one stands to gain significantly from the use of a HAC estimator rather than an Eicker-White type estimator.

The next feature of note in Tables IV–VI is the very poor performance of all of the estimators in the AR(1) models when $\rho = .9$ or .95. This is expected for INID and IID, but it also is true for QS and PARA. For the QS estimator, this poor performance is not due to poor choices of S_T or to the choice of kernel—the results are improved little or none if $\hat{S}_T$ is replaced by the finite

TABLE V

BIAS AND MSE OF QS ESTIMATOR WITH AUTOMATIC S_T VALUE, $\hat{S}_T$, AND TRUE CONFIDENCE LEVEL OF NOMINAL 95% CONFIDENCE INTERVAL CONSTRUCTED USING THE QS ESTIMATOR WITH AUTOMATIC S_T VALUE FOR THE AR(1)-HET1, AR(1)-HET2, AND MA(1)-HOMO MODELS – $T = 128$

Model/Estimator		Bias	MSE	95%	Model/Estimator		Bias	MSE	95%
AR(1)-HET1	QS	−.32	1.35	92.8	AR(1)-HET1	QS	−1.1	2.9	89.0
$\rho = 0$	INID	−.33	1.23	92.9	$\rho = .3$	INID	−1.2	2.9	88.0
(2.94)[a]	IID	−1.95	3.86	75.4	(3.89)	IID	−2.9	8.5	69.6
	PARA	−1.95	3.86	75.0		PARA	−2.8	8.0	71.1
AR(1)-HET1	QS	−2.0	7.5	87.4	AR(1)-HET1	QS	−18.	352.	60.5
$\rho = .5$	INID	−2.7	9.0	82.6	$\rho = .9$	INID	−22.	478.	38.8
(5.31)	IID	−4.4	19.0	58.7	(23.4)	IID	−23.	515.	27.7
	PARA	−4.0	16.6	64.7		PARA	−21.	442.	46.2
AR(1)-HET2	QS	−.15	.34	91.5	AR(1)-HET2	QS	−.23	.59	91.0
$\rho = 0$	INID	−.15	.32	91.6	$\rho = .3$	INID	−.32	.50	90.4
(1.47)	IID	−.49	.28	88.6	(1.67)	IID	−.70	.54	86.7
	PARA	−.49	.29	88.5		PARA	−.61	.44	87.3
AR(1)-HET2	QS	−.52	1.17	89.5	AR(1)-HET2	QS	−4.5	26.5	71.3
$\rho = .5$	INID	−.85	1.09	85.8	$\rho = .9$	INID	−6.3	40.5	48.7
(2.15)	IID	−1.19	1.47	81.3	(7.18)	IID	−6.5	42.1	45.6
	PARA	−.88	.91	87.6		PARA	−4.6	25.3	72.2
MA(1)-HOMO	QS	−.24	.16	91.3	MA(1)-HOMO	QS	−.22	.27	91.0
$\psi = .5$	INID	−.37	.18	89.2	$\psi = .99$	INID	−.55	.35	85.5
(1.31)	IID	−.32	.13	91.2	(1.48)	IID	−.49	.27	88.4
	PARA	−.049	.058	93.7		PARA	−.064	.089	94.4

[a] The numbers in parentheses in columns 1 and 6 are the values of the estimand.

sample optimal S_T value or if the QS kernel is replaced by any of the other four kernels.

A comparison of the QS and PARA estimators for sample size $T = 128$ (Tables IV and V) shows that PARA is better than QS in the AR(1)-HOMO and MA(1)-HOMO models in terms of MSE and true confidence levels. The differences in MSE are quite large for $\rho \leqslant .7$; the differences in true confidence levels are much smaller. In the AR(1)-HET1 model, the reverse is true. The QS estimator is much better than PARA in terms of both MSE and true confidence levels over the entire range of ρ values. In the AR(1)-HET2 model, neither QS nor PARA is dominant. PARA enjoys an edge in MSE, but QS is better in terms of true confidence levels.

In sum, for $T = 128$, the PARA is the best all-round estimator if one ignores the AR(1)-HET1 model. Even PARA performs very poorly in each of the AR(1) models, however, when $\rho = .9$ or .95. If one includes the AR(1)-HET1 model, then the QS estimator is the best all-round estimator, since PARA does very poorly in this model. Nevertheless, the QS estimator pays a significant price for attaining its versatility, as the comparison with PARA in the AR(1)-HOMO model attests.

TABLE VI

BIAS AND MSE OF QS ESTIMATOR WITH AUTOMATIC S_T VALUE, $\hat{S}_T$, AND TRUE CONFIDENCE LEVEL OF NOMINAL 95% CONFIDENCE INTERVAL CONSTRUCTED USING THE QS ESTIMATOR WITH AUTOMATIC S_T VALUE FOR THE AR(1)-HOMO, AR(1)-HET1, AR(1)-HET2, AND MA(1)-HOMO MODELS – $T = 256$

Model/Estimator		Bias	MSE	95%	Model/Estimator		Bias	MSE	95%
AR(1)-HOMO	QS	−.03	.025	93.7	AR(1)-HOMO	QS	−.119	.062	93.0
$\rho = 0$	INID	−.03	.024	93.7	$\rho = .3$	INID	−.23	.076	91.2
(1.00)[a]	IID	−.00	.0098	94.5	(1.19)	IID	−.20	.049	92.2
	PARA	−.00	.0098	94.5		PARA	−.01	.022	95.0
AR(1)-HOMO	QS	−3.54	19.0	81.0	AR(1)-HOMO	QS	−8.2	79.	70.9
$\rho = .9$	INID	−6.96	48.5	45.4	$\rho = .95$	INID	−12.3	152.	32.7
(7.72)	IID	−6.90	47.6	46.6	(12.9)	IID	−12.3	150.	34.3
	PARA	−2.23	11.6	88.8		PARA	−6.3	54.	79.4
AR(1)-HET1	QS	−.81	1.73	92.2	AR(1)-HET1	QS	−1.7	5.3	89.7
$\rho = .3$	INID	−1.11	1.89	90.4	$\rho = .5$	INID	−2.7	8.1	83.4
(3.92)	IID	−2.92	8.54	71.9	(5.44)	IID	−4.5	20.0	59.3
	PARA	−2.80	7.91	74.3		PARA	−4.1	16.8	66.9
AR(1)-HET2	QS	−.14	.38	93.3	AR(1)-HET2	QS	−.36	.61	91.9
$\rho = .3$	INID	−.26	.31	92.2	$\rho = .5$	INID	−.82	.86	87.6
(1.70)	IID	−.71	.53	84.8	(2.22)	IID	−1.24	1.57	80.3
	PARA	−.60	.40	87.0		PARA	−.87	.83	87.4
MA(1)-HOMO	QS	−.17	.089	92.7	MA(1)-HOMO	QS	−.086	.17	93.3
$\psi = .5$	INID	−.34	.140	90.4	$\psi = .99$	INID	−.53	.31	87.9
(1.31)	IID	−.32	.110	90.1	(1.49)	IID	−.50	.26	88.8
	PARA	−.026	.026	95.2		PARA	−.039	.043	94.4

[a] The numbers in parentheses in columns 1 and 6 are the values of the estimand.

Next we discuss the changes that occur in the results when the sample size is increased from 128 to 256 (see Table VI). For the INID and IID estimators, there is not much change. When $\rho = 0$ or $\psi = 0$ there are improvements in their MSE's and some improvements in their true confidence levels. But, when $\rho > 0$ or $\psi > 0$, there is not much improvement in either. In consequence, the dominance of QS over INID and PARA over IID is enhanced when the sample size is increased.

For the QS and PARA estimators, the increase in sample size from 128 to 256 causes a substantial improvement in their MSE's and true confidence levels in the AR(1)-HOMO model, especially for large values of ρ. The gap between the true confidence levels of the QS and PARA estimators is narrowed. In the AR(1)-HET1 and AR(1)-HET2 models, the QS estimator exhibits similar improvements when the sample size is increased. The PARA estimator, however, shows no improvement in the AR(1)-HET1 model and only small improvements in the AR(1)-HET2 model. In consequence, the dominance of QS over PARA in the AR(1)-HET1 model is accentuated when $T = 256$, and the lack of dominance of either QS or PARA in the AR(1)-HET2 model when $T = 128$ is replaced by dominance of QS when $T = 256$. In the MA(1)-HOMO model, QS

and PARA both improve in MSE with the sample size increase; QS also improves in true confidence levels, but PARA does not.

In sum, the increase in sample size from 128 to 256 improves the overall performance of the QS estimator absolutely and relatively to the PARA, INID, and IID estimators. As when $T = 128$, QS has the best overall performance of the four estimators when $T = 256$ if one includes the AR(1)-HET1 model. PARA is the best estimator overall if this model is excluded. In the latter case, the preference for PARA over QS in terms of true confidence levels is much less when $T = 256$ than when $T = 128$.

10. CONCLUSION

The results of this paper are summarized as follows:

(i) The paper establishes the consistency of kernel HAC estimators under conditions that are more general in most respects than other results in the literature. In particular, they are more general with respect to the class of kernels considered and the allowable rate of increase of the bandwidth parameters. In addition, the paper establishes rate of convergence and asymptotic truncated MSE results for kernel HAC estimators.

(ii) The paper compares different kernel HAC estimators in the literature via asymptotic and simulation methods. The paper establishes an asymptotically optimal kernel, viz., the QS kernel, from the class of kernels that generate psd estimates. The latter includes the Bartlett and Parzen kernels. The Monte Carlo results (including those reported here and those available from the author upon request) substantiate the optimality of the QS kernel within this class in terms of both MSE and true confidence level performance. The Monte Carlo results indicate, however, that the differences between the kernels are not large. They indicate that the Bartlett kernel, used by Newey and West (1987), is somewhat inferior to the other kernels considered.

(iii) The paper determines suitable fixed and automatic bandwidth parameters for use with HAC estimators. The latter are based on the plug-in method. They are found to perform surprisingly well in most cases in the simulations.

(iv) The paper compares the performance of kernel HAC estimators to that of other types of covariance matrix estimators via Monte Carlo simulation. The other estimators considered are the Eicker-White heteroskedasticity consistent estimator, the standard LS covariance matrix estimator (IID), and a parametric estimator (PARA) that assumes that the errors are homoskedastic and AR(1). The QS HAC estimator more or less dominates the Eicker-White estimator and is the most versatile estimator of those considered. But, it pays a significant price for its versatility, as is illustrated by its performance relative to that of PARA in those scenarios for which PARA is designed.

All of the estimators considered perform very poorly in an absolute sense when the amount of autocorrelation is large. For the HAC estimators, this is found to be true even if the finite sample optimal bandwidth parameters are used.

Cowles Foundation for Research in Economics, Yale University, P. O. Box 2125, Yale Station, New Haven, CT 06520-2125, U.S.A.

Manuscript received August, 1988; final revision received June, 1990.

APPENDIX

PROOF OF LEMMA 1: First, consider the case where $\{V_t\}$ is fourth order stationary. For notational simplicity, suppose $p = 1$. Using a standard α-mixing inequality (see Hall and Hyde (1980, Corollary A.2, p. 278)), we obtain the first condition of Assumption A:

$$\text{(A.1)} \qquad \sum_{j=-\infty}^{\infty} |EV_tV_{t-j}| \leqslant \sum_{j=-\infty}^{\infty} 8\big(E|V_0|^{2\nu}\big)^{1/\nu}\alpha(j)^{(\nu-1)/\nu} < \infty.$$

To establish the cumulant condition of Assumption A, it suffices to show that

$$\text{(A.2)} \qquad H = \sum_{j=1}^{\infty}\sum_{m=1}^{\infty}\sum_{n=1}^{\infty} |EV_0V_jV_mV_n - E\tilde{V}_0\tilde{V}_j\tilde{V}_m\tilde{V}_n| < \infty$$

and the analogous result with $\sum_{n=1}^{\infty}$ replaced by $\sum_{n=-\infty}^{-1}$. The latter follows by a similar argument to that used to prove (A.2). Hence, we only prove (A.2) here.

There are 3! orderings of (j, m, n). Hence,

$$\begin{aligned}
\text{(A.3)} \qquad H &\leqslant 3!\sum_{j=1}^{\infty}\sum_{m=j}^{\infty}\sum_{n=m}^{\infty} |EV_0V_jV_mV_n - E\tilde{V}_0\tilde{V}_j\tilde{V}_m\tilde{V}_n| \\
&= 6\sum_{j=1}^{\infty}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} |EV_0V_jV_{j+m}V_{j+m+n} - E\tilde{V}_0\tilde{V}_j\tilde{V}_{j+m}\tilde{V}_{j+m+n}| \\
&\leqslant 6\sum_{0\leqslant m,n\leqslant j}\Big(\big|EV_0\big(V_jV_{j+m}V_{j+m+n}\big)\big| + \big|E\tilde{V}_0\big(\tilde{V}_j\tilde{V}_{j+m}\tilde{V}_{j+m+n}\big)\big|\Big) \\
&\quad + 6\sum_{0\leqslant j,n\leqslant m}\Big(\big|EV_0V_j\big(V_{j+m}V_{j+m+n}\big) - EV_0V_jEV_{j+m}V_{j+m+n}\big| \\
&\qquad\qquad + \big|E\tilde{V}_0\tilde{V}_j\big(\tilde{V}_{j+m}\tilde{V}_{j+m+n}\big) - E\tilde{V}_0\tilde{V}_jE\tilde{V}_{j+m}\tilde{V}_{j+m+n}\big|\Big) \\
&\quad + 6\sum_{0\leqslant j,m\leqslant n}\Big(\big|E\big(V_0V_jV_{j+m}\big)V_{j+m+n}\big| + \big|E\big(\tilde{V}_0\tilde{V}_j\tilde{V}_{j+m}\big)\tilde{V}_{j+m+n}\big|\Big).
\end{aligned}$$

The last inequality uses the fact that $\{V_t\}$ and $\{\tilde{V}_t\}$ have the same autocovariances.

Using the mixing inequality referred to above, we get

$$\begin{aligned}
\text{(A.4)} \qquad &\big|EV_0\big(V_jV_{j+m}V_{j+m+n}\big)\big| \leqslant 8\big(E|V_0|^{4\nu}\big)^{1/\nu}\alpha(j)^{(\nu-1)/\nu}, \\
&\big|EV_0V_j\big(V_{j+m}V_{j+m+n}\big) - EV_0V_jEV_{j+m}V_{j+m+n}\big| \leqslant 8\big(E|V_0|^{4\nu}\big)^{1/\nu}\alpha(m)^{(\nu-1)/\nu}, \\
&\big|E\big(V_0V_jV_{j+m}\big)V_{j+m+n}\big| \leqslant 8\big(E|V_0|^{4\nu}\big)^{1/\nu}\alpha(n)^{(\nu-1)/\nu}.
\end{aligned}$$

By expressing $E\tilde{V}_0\tilde{V}_j\tilde{V}_{j+m}\tilde{V}_{j+m+n}$ in terms of the covariances of $(\tilde{V}_0, \tilde{V}_j, \tilde{V}_{j+m}, \tilde{V}_{j+m+n})$, which equal the covariances of $(V_0, V_j, V_{j+m}, V_{j+m+n})$, and by bounding the latter covariances using the mixing inequality, we get

$$\begin{aligned}
\text{(A.5)} \qquad |E\tilde{V}_0\tilde{V}_j\tilde{V}_{j+m}\tilde{V}_{j+m+n}| &\leqslant C(\alpha^{\tau}(j)\alpha^{\tau}(n) + \alpha^{\tau}(j+m)\alpha^{\tau}(m+n) \\
&\qquad + \alpha^{\tau}(j+m+n)\alpha^{\tau}(m)), \\
|E\tilde{V}_0\tilde{V}_j\tilde{V}_{j+m}\tilde{V}_{j+m+n} - E\tilde{V}_0\tilde{V}_jE\tilde{V}_{j+m}\tilde{V}_{j+m+n}| &\leqslant C(\alpha^{\tau}(j+m)\alpha^{\tau}(m+n) \\
&\qquad + \alpha^{\tau}(j+m+n)\alpha^{\tau}(m))
\end{aligned}$$

for some $C < \infty$, where $\tau = (\nu - 1)/\nu$.

Next, we have

$$\text{(A.6)}\qquad \sum_{0\leqslant m,n\leqslant j}\alpha(j)^{(\nu-1)/\nu}\leqslant \sum_{j=0}^{\infty}\sum_{m=0}^{j}\sum_{n=0}^{j}\alpha(j)^{(\nu-1)/\nu}\leqslant \sum_{j=0}^{\infty}(j+1)^2\alpha(j)^{(\nu-1)/\nu}<\infty.$$

Equations (A.3)–(A.6) combine to yield (A.2).

It is straightforward to adjust the above proof for the case where $\{V_t\}$ is not fourth order stationary. For example, in (A.2), $E|V_0V_jV_mV_n-E\tilde{V}_0\tilde{V}_j\tilde{V}_m\tilde{V}_n|$ is replaced by

$$\sup_{t\geqslant 1}E|V_tV_{t+j}V_{t+m}V_{t+n}-E\tilde{V}_t\tilde{V}_{t+j}\tilde{V}_{t+m}\tilde{V}_{t+n}|. \qquad \textit{Q.E.D.}$$

PROOF OF PROPOSITION 1: For the scalar V_t case, part (a) is given by Theorem 5A of Parzen (1957). For the vector case, Theorem 9 of Hannan (1970, p. 280) gives the asymptotic covariance between any two elements of $\tilde{J}_T$. The commutation-tensor product formula of part (a) is obtained by observing that the asymptotic covariances between $[\tilde{J}_T]_{ij}$ and $[\tilde{J}_T]_{mn}$ for $i,j,m,n\leqslant p$ are of the same form as the covariances between X_iX_j and X_mX_n, where $\boldsymbol{X}=(X_1,\ldots,X_p)'\sim N(\boldsymbol{0},\Sigma)$ (e.g., see Muirhead (1982, p. 20)). By Magnus and Neudecker (1979, Theorem 4.3(iv)), $\operatorname{Var}(\operatorname{vec}\boldsymbol{XX}')=\operatorname{Var}(\boldsymbol{X}\otimes\boldsymbol{X})=(I+K_{pp})\Sigma\otimes\Sigma$. The given formula follows.

For the scalar V_t case, part (b) of the Theorem is given by Theorem 5B of Parzen (1957). For the vector case, it is given by Theorem 10 of Hannan (1970, p. 283). (Note that the proofs of Hannan's Theorems 9 and 10 go through even if $k(\cdot)$ is not continuous everywhere, as he assumes, but only at zero and all but a finite number of points.)

In part (c), $T/S_T=S_T^{2q}/(S_T^{2q+1}/T)=S_T^{2q}/(\gamma+o(1))$. Thus,

$$\text{(A.7)}\qquad \lim_{T\to\infty}\operatorname{MSE}\left(T/S_T,\tilde{J}_T,W\right)$$

$$=\lim_{T\to\infty}S_T^{2q}\left(E\tilde{J}_T-J_T\right)'W\left(E\tilde{J}_T-J_T\right)/(\gamma+o(1))+\lim_{T\to\infty}\frac{T}{S_T}\operatorname{tr}W\operatorname{Var}\left(\operatorname{vec}\tilde{J}_T\right).$$

Part (c) now follows from parts (a) and (b). *Q.E.D.*

The following two simple lemmas are used in the proof of Theorem 1:

LEMMA A1: *If $\{\xi_T\}$ is a bounded sequence of rv's such that $\xi_T\to^p 0$, then $E\xi_T\to 0$.*

PROOF OF LEMMA A1: Convergence in probability to zero implies weak convergence to zero. For bounded rv's, the latter implies convergence of expectations to zero by the definition of weak convergence. *Q.E.D.*

LEMMA A2: *Let $\{X_T\}$ be a sequence of nonnegative rv's for which $\sup_{T\geqslant 1}EX_T^{1+\delta}<\infty$ for some $\delta>0$. Then, $\lim_{h\to\infty}\lim_{T\to\infty}(E\min\{X_T,h\}-EX_T)=0$.*

PROOF OF LEMMA A2: The following establishes the Lemma:

$$\text{(A.8)}\qquad 0\leqslant\lim_{h\to\infty}\lim_{T\to\infty}E\left(X_T-\min\{X_T,h\}\right)\leqslant\lim_{h\to\infty}\lim_{T\to\infty}EX_T1(X_T\geqslant h)$$

$$\leqslant\lim_{h\to\infty}\lim_{T\to\infty}\left(EX_T^{1+\delta}\right)^{1/(1+\delta)}P(X_T\geqslant h)^{\delta/(1+\delta)}$$

$$\leqslant\left(\sup_{T\geqslant 1}EX_T^{1+\delta}\right)^{1/(1+\delta)}\lim_{h\to\infty}\lim_{T\to\infty}(EX_T/h)^{\delta/(1+\delta)}=0. \qquad \textit{Q.E.D.}$$

PROOF OF THEOREM 1: By definition of $\mathscr{K}_1$ and Assumption A, $k_0=0$ and $|f^{(0)}|<\infty$. Hence, under the assumptions of Theorem 1(a), Proposition 1(a) and (b) (with $q=0$ in the latter) gives $\tilde{J}_T-J_T=o_p(1)$. Similarly, under the assumptions of Theorem 1(b), Proposition 1(c) yields $\sqrt{T/S_T}(\tilde{J}_T-J_T)=O_p(1)$. Thus, Theorem 1(a) and (b) holds if the second result stated in each of these parts

holds. The latter hold if and only if they hold with $\hat{J}_T - \tilde{J}_T$ replaced by $b'\hat{J}_T b - b'\tilde{J}_T b$ for arbitrary $b \in R^p$. In consequence, we suppose $\hat{J}_T$ and $\tilde{J}_T$ are scalars without loss of generality. In addition, we suppose $\hat{J}_T$ is defined without any degrees of freedom adjustment since this simplifies the expressions without affecting the results.

We now show that $(\sqrt{T}/S_T)(\hat{J}_T - \tilde{J}_T) = O_p(1)$ provided $S_T \to \infty$ and Assumption B holds. This yields the second result of Theorem 1(a). Let $\tilde{J}_T(\theta)$ denote the "estimator" calculated using $k(\cdot)$, S_T, and $\{V_t(\theta)\}$. A mean value expansion of $\tilde{J}_T(\hat{\theta})$ $(= \hat{J}_T)$ about θ_0 yields

$$\text{(A.9)} \qquad \frac{\sqrt{T}}{S_T}\left(\hat{J}_T - \tilde{J}_T\right) = \frac{1}{S_T}\frac{\partial}{\partial \theta'}\tilde{J}_T(\bar{\theta})\sqrt{T}\left(\hat{\theta} - \theta_0\right)$$

$$= \frac{1}{S_T}\sum_{j=-T+1}^{T-1} k(j/S_T)\frac{\partial}{\partial \theta'}\hat{\Gamma}(j)\bigg|_{\theta=\bar{\theta}} \sqrt{T}\left(\hat{\theta} - \theta_0\right)$$

for some $\bar{\theta}$ on the line segment joining $\hat{\theta}$ and θ_0. In addition, we have

$$\text{(A.10)} \qquad \sup_{j \geqslant 1}\left\|\frac{\partial}{\partial \theta}\hat{\Gamma}(j)\right\|\bigg|_{\theta=\bar{\theta}} = \sup_{j \geqslant 1}\left\|\frac{1}{T}\sum_{t=|j|+1}^{T}\left(V_t(\bar{\theta})\frac{\partial}{\partial \theta}V_{t-|j|}(\bar{\theta}) + V_{t-|j|}(\bar{\theta})\frac{\partial}{\partial \theta}V_t(\bar{\theta})\right)\right\|$$

$$\leqslant 2\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta \in \Theta} V_t^2(\theta)\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\theta \in \Theta}\left\|\frac{\partial}{\partial \theta}V_t(\theta)\right\|^2\right)^{1/2} = O_p(1),$$

where the second equality follows using Assumption B(ii) and (iii) by applying Markov's inequality to each of the terms in parentheses (and noting that $\sup_{t \geqslant 1} E \sup_{\theta \in \Theta}\|V_t(\theta)\|^2 < \infty$ under Assumptions B(ii) and (iii) by a mean value expansion argument). This result, Assumption B(i), and the fact that $(1/S_T)\sum_{j=-T+1}^{T-1}|k(j/S_T)| \to \int_{-\infty}^{\infty}|k(x)|\,dx < \infty$ imply that the right-hand side of (A.9) is $O_p(1)$ and the proof of Theorem 1(a) is complete.

Next we show that $\sqrt{T/S_T}(\hat{J}_T - \tilde{J}_T) = o_p(1)$ under the assumptions of Theorem 1(b). A two-term Taylor expansion gives

$$\text{(A.11)} \qquad \sqrt{T/S_T}\left(\hat{J}_T - \tilde{J}_T\right) = \left[\frac{\partial}{\partial \theta'}\tilde{J}_T(\theta_0)/\sqrt{S_T}\right]\sqrt{T}\left(\hat{\theta} - \theta_0\right)$$

$$+ \frac{1}{2}\sqrt{T}\left(\hat{\theta} - \theta_0\right)'\left[\frac{\partial^2}{\partial \theta\,\partial \theta'}\tilde{J}_T(\bar{\theta})/\sqrt{TS_T}\right]\sqrt{T}\left(\hat{\theta} - \theta_0\right)$$

$$= L_{1T}'\sqrt{T}\left(\hat{\theta} - \theta_0\right) + \frac{1}{2}\sqrt{T}\left(\hat{\theta} - \theta_0\right)'L_{2T}\sqrt{T}\left(\hat{\theta} - \theta_0\right),$$

where $L_{1T}(\in R^p)$ and $L_{2T}(\in R^{p \times p})$ are defined implicitly and $\bar{\theta}$ lies on the line segment joining $\hat{\theta}$ and θ_0. Manipulations similar to those of (A.10) and Assumptions B(ii), B(iii), and C(ii) yield

$$\text{(A.12)} \qquad \|L_{2T}\| \leqslant \left(\frac{1}{TS_T}\right)^{1/2}\sum_{j=-T+1}^{T-1}|k(j/S_T)|\frac{1}{T}\sum_{t=|j|+1}^{T}\sup_{\theta \in \Theta}\left\|\frac{\partial^2}{\partial \theta\,\partial \theta'}V_t(\theta)V_{t-|j|}(\theta)\right\|$$

$$= \left(\frac{S_T}{T}\right)^{1/2}\left(\frac{1}{S_T}\sum_{j=-T+1}^{T-1}|k(j/S_t)|\right)O_p(1) = o_p(1)$$

(using Markov's inequality to show that the sample averages that arise are $O_p(1)$).

To obtain $L_{1T} = o_p(1)$, we use Assumption C(i) and apply Proposition 1(a) and (b) to $\tilde{J}_T$ with the latter constructed using $(V_t', \partial V_t/\partial \theta' - E(\partial V_t/\partial \theta'))'$ rather than just V_t. The first row and column of off-diagonal elements of this $\tilde{J}_T$ matrix (written as column vectors) are

$$\text{(A.13)} \qquad \sum_{j=-T+1}^{T-1} k(j/S_T)\frac{1}{T}\sum_{t=|j|+1}^{T} V_t\left(\frac{\partial}{\partial \theta}V_{t-|j|} - \lambda\right) \qquad \text{and}$$

$$\sum_{j=-T+1}^{T-1} k(j/S_T)\frac{1}{T}\sum_{t=|j|+1}^{T}\left(\frac{\partial}{\partial \theta}V_t - \lambda\right)V_{t-|j|},$$

where $\lambda = E(\partial/\partial\theta)V_t$. By Proposition 1(a) and (b), these vectors are $O_p(1)$. L_{1T} is equal to the sum of the two expressions in (A.13) times $1/\sqrt{S_T}$ plus $D_T\lambda$, where

$$D_T = \sum_{j=-T+1}^{T-1} k(j/S_T)\frac{1}{T}\sum_{t=|j|+1}^{T}(V_t + V_{t-|j|})/\sqrt{S_T}. \tag{A.14}$$

In consequence, $L_{1T} = o_p(1)$ if $D_T = o_p(1)$. We have

$$\begin{aligned} ED_T^2 &\leqslant \frac{1}{S_T}\sum_{i=-T+1}^{T-1}\sum_{j=-T+1}^{T-1}|k(i/S_T)k(j/S_T)|\frac{4}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T}|EV_sV_t| \\ &\leqslant \frac{S_T}{T}\left(\frac{1}{S_T}\sum_{j=-T+1}^{T-1}|k(j/S_T)|\right)^2\sum_{u=-T+1}^{T-1}|\Gamma(u)| = o(1). \end{aligned} \tag{A.15}$$

Since L_{1T} and L_{2t} are $o_p(1)$, so is the right-hand side of (A.11) and the proof of Theorem 1(b) is complete.

To establish the first equality of Theorem 1(c), we apply Lemma A1 with

$$\begin{aligned} \xi_T = &\min\left\{\frac{T}{S_T}\left|\operatorname{vec}\left(\hat{J}_T - J_T\right)'W_T\operatorname{vec}\left(\hat{J}_T - J_T\right)\right|, h\right\} \\ &- \min\left\{\frac{T}{S_T}\left|\operatorname{vec}\left(\tilde{J}_T - J_T\right)'W_T\operatorname{vec}\left(\tilde{J}_T - J_T\right)\right|, h\right\}. \end{aligned} \tag{A.16}$$

Since $\sqrt{T/S_T}(\hat{J}_T - \tilde{J}_T) = o_p(1)$ and $\sqrt{T/S_T}(\hat{J}_T - J_T) = O_p(1)$ by Theorem 1(b), $\xi_T \to^p 0$. Also, $|\xi_T| \leqslant h$. Hence, $E\xi_T \to 0$. Since this holds for all h, the first equality of Theorem 1(c) holds.

The second equality of Theorem 1(c) is obtained by showing that

$$\lim_{h\to\infty}\lim_{T\to\infty}\left(\operatorname{MSE}_h\left(T/S_T, \tilde{J}_T, W_T\right) - \operatorname{MSE}_h\left(T/S_T, \tilde{J}_T, W\right)\right) = 0 \quad \text{and} \tag{A.17}$$

$$\lim_{h\to\infty}\lim_{T\to\infty}\operatorname{MSE}_h\left(T/S_T, \tilde{J}_T, W\right) = \lim_{T\to\infty}\operatorname{MSE}\left(T/S_T, \tilde{J}_T, W\right). \tag{A.18}$$

Under Assumption D(ii), (A.17) holds by applying Lemma A1. Equation (A.18) holds by applying Lemma A2 with $X_T = |(T/S_T)\operatorname{vec}(\tilde{J}_T - J_T)'W\operatorname{vec}(\tilde{J}_T - J_T)|$. We get $\sup_{T\geqslant 1} EX_T^2 < \infty$, as required by Lemma A2, if $E(\sqrt{T/S_T}[\tilde{J}_T - J_T]_{ab})^4 = O(1)\ \forall a, b \leqslant p$, where $[\cdot]_{ab}$ denotes the (a, b) element of the matrix. This fourth moment equals $\kappa_{T4} + 4\kappa_{T3}\kappa_{T1} + 3\kappa_{T2}^2 + 6\kappa_{T2}\kappa_{T1}^2 + \kappa_{T1}^4$, where κ_{Tj} denotes the jth cumulant of $\sqrt{T/S_T}[\tilde{J}_T - J_T]_{ab}$ (e.g., see Stuart and Ord (1987, p. 86)). Under Assumption D(i), κ_{T3} and κ_{T4} are $o(1)$ by the proof of Theorem 7.7.1 of Brillinger (1981, pp. 262, 441–444). (Note that Brillinger's Assumptions 2.6.2(1) and 7.7.1, which are assumed in his Theorem 7.7.1 but are not assumed here, are used in his proof only for the results concerning first and second order cumulants, and hence, are not needed here.) Also, κ_{T1} and κ_{T2} equal the mean and variance of $\sqrt{T/S_T}[\tilde{J}_T - J_T]_{ab}$, and hence, are $O(1)$ by Proposition 1(c). In consequence, $\sup_{T\geqslant 1} EX_T^2 < \infty$ and the second equality of Theorem 1(c) holds.

The third equality of Theorem 1(c) holds by Proposition 1(c). *Q.E.D.*

PROOF OF THEOREM 2: We apply Theorem 1(c) with the kernel $k(\cdot)$, the bandwidth sequence $\{S_{Tk}\}$, and $q = 2$. Since $S_{Tk}^5/T \to \gamma/(\int k^2(x)\,dx)^5$ and $T/S_T = (1/\int k^2(x)\,dx)T/S_{Tk}$, Theorem 1(c) gives

$$\begin{aligned} &\lim_{h\to\infty}\lim_{T\to\infty}\operatorname{MSE}_h\left(T/S_T, \hat{J}_T(S_{Tk}), W_T\right) \\ &\quad = 4\pi^2\left(k_2^2\left(\int k^2(x)\,dx\right)^4\left(\operatorname{vec} f^{(2)}\right)'W\operatorname{vec} f^{(2)}/\gamma + \operatorname{tr} W\left(I + K_{pp}\right)f\otimes f\right) \end{aligned} \tag{A.19}$$

provided $k_2 < \infty$. Since $\int k_{QS}^2(x)\,dx = 1$, this yields the equality in the result of Theorem 2. If $k_2 = \infty$, then the left-hand side of (A.19) equals infinity since the bias term is unbounded. This can be proved along the lines of Andrews (1988, Lemma 2).

Let $K(\cdot)$ and $K_{QS}(\cdot)$ denote the spectral window generators of $k(\cdot)$ and $k_{QS}(\cdot)$ respectively (as defined at the end of Section 2). By standard calculations, we have $k_2 = \int_{-\infty}^{\infty} \lambda^2 K(\lambda)\,d\lambda$, $k(0) = \int_{-\infty}^{\infty} K(\lambda)\,d\lambda$, and $\int_{-\infty}^{\infty} k^2(x)\,dx = \int_{-\infty}^{\infty} K^2(\lambda)\,d\lambda$. Thus,

$$\text{(A.20)} \qquad k_2\left(\int k^2(x)\,dx\right)^2 \geqslant k_{2QS} \qquad \text{for all } k(\cdot) \in \mathscr{K}_2$$

if and only if $K_{QS}(\cdot)$ minimizes

$$\text{(A.21)} \qquad \int_{-\infty}^{\infty} \lambda^2 K(\lambda)\,d\lambda \left(\int_{-\infty}^{\infty} K^2(\lambda)\,d\lambda\right)^2$$

subject to (a) $\int_{-\infty}^{\infty} K(\lambda)\,d\lambda = 1$, (b) $K(\lambda) \geqslant 0$ $\forall \lambda \in R$, and (c) $K(\lambda) = K(-\lambda)$ $\forall \lambda \in R$, where $K_{QS}(\lambda) = (5/8\pi)(1 - \lambda^2/c^2)$ for $|\lambda| \leqslant c$ and $K_{QS}(\lambda) = 0$ otherwise for $c = 6\pi/5$.

The minimization problem (A.21) is the same as that which arises in a result of Priestley (1981, p. 570) who considers a maximum (over different frequencies) relative MSE criterion. Using a calculus of variations argument, Priestley shows that $K_{QS}(\cdot)$ solves (A.21). Hence, (A.20) holds and combined with (A.19) this establishes the inequality in the result of the theorem. If $k(x) \neq k_{QS}(x)$ with positive Lebesgue measure, then (A.20) holds with a strict inequality and so does the result of the Theorem. *Q.E.D.*

PROOF OF COROLLARY 1: By Theorem 1(c) and the fact that

$$T^{2q/(2q+1)} = \left(S_T^{2q+1}/T\right)^{1/(2q+1)} T/S_T = \left(\gamma^{1/(2q+1)} + o(1)\right) T/S_T,$$

we get

$$\text{(A.22)} \qquad \lim_{h\to\infty} \lim_{T\to\infty} \mathrm{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(S_T), W_T\right)$$

$$= \gamma^{1/(2q+1)} 4\pi^2 \left(k_q^2 (\operatorname{vec} f^{(q)})' W \operatorname{vec} f^{(q)}/\gamma + \int k^2(x)\,dx \operatorname{tr} W(I + K_{pp}) f \otimes f\right).$$

It is straightforward to show that the last line above is uniquely minimized over $\gamma \in (0, \infty)$ by $\gamma^* = q k_q^2 \alpha(q)/\int k^2(x)\,dx$ (provided $0 < \alpha(q) < \infty$ and W is psd) and that a sequence $\{S_T\}$ satisfies $S_T^{2q+1}/T \to \gamma^*$ if and only if $S_T = S_T^* + o(T^{1/(2q+1)})$. *Q.E.D.*

PROOF OF THEOREM 3: First we establish Theorem 3(b). By Theorem 1(b), $\sqrt{T/S_{\xi T}}(\hat{J}_T(S_{\xi T}) - J_T) = O_p(1)$. Hence, it suffices to establish the second result of Theorem 3(b). Without loss of generality, assume V_t is a scalar rv and no degrees of freedom correction is made to $\hat{J}_T$. Let v be a constant in the interval $(\max\{1 + 1/(2b-2), q/(m-1)\}, 1 + q/2)$ and let $r(T) = [(S_{\xi T})^v]$, where $[\cdot]$ denotes the integer part of $\cdot$. We have

$$\text{(A.23)} \qquad T^{q/(2q+1)}\left(\hat{J}_T(\hat{S}_T) - \hat{J}_T(S_{\xi T})\right) = 2T^{q/(2q+1)} \sum_{j=1}^{r(T)} \left(k\left(j/\hat{S}_T\right) - k\left(j/S_{\xi T}\right)\right)\hat{\Gamma}(j)$$

$$+ 2T^{q/(2q+1)} \sum_{j=r(T)+1}^{T-1} k\left(j/\hat{S}_T\right)\hat{\Gamma}(j)$$

$$- 2T^{q/(2q+1)} \sum_{j=r(T)+1}^{T-1} k\left(j/S_{\xi T}\right)\hat{\Gamma}(j)$$

$$= 2M_{1T} + 2M_{2T} - 2M_{3T}.$$

We show $M_{1T} \to^p 0$ as follows: Using the Lipschitz condition on $k(\cdot)$,

$$(A.24) \qquad |M_{1T}| \leqslant T^{q/(2q+1)} \sum_{j=1}^{r(T)} C_2 \left|1/\hat{S}_T - 1/S_{\xi T}\right| j \left|\hat{\Gamma}(j)\right|$$

$$\leqslant C\sqrt{T}\left|\hat{\alpha}(q)^{1/(2q+1)} - \alpha_\xi^{1/(2q+1)}\right| \left(\hat{\alpha}(q)\alpha_\xi\right)^{-1/(2q+1)} T^{-3/(4q+2)} \sum_{j=1}^{r(T)} j\left|\hat{\Gamma}(j)\right|$$

for some constant $C < \infty$. By Assumption F and the delta method, it suffices to show that $G_{1T} + G_{2T} + G_{3T} \to^p 0$, where

$$(A.25) \qquad G_{1T} = T^{-3/(4q+2)} \sum_{j=1}^{r(T)} j\left|\hat{\Gamma}(j) - \tilde{\Gamma}(j)\right|,$$

$$G_{2T} = T^{-3/(4q+2)} \sum_{j=1}^{r(T)} j\left|\tilde{\Gamma}(j) - \Gamma_T(j)\right|, \qquad \text{and}$$

$$G_{3T} = T^{-3/(4q+2)} \sum_{j=1}^{r(T)} j\left|\Gamma_T(j)\right|.$$

By a mean value expansion, we have

$$(A.26) \qquad G_{1T} \leqslant T^{-3/(4q+2)-1/2} r(T) \sum_{j=1}^{r(T)} \left|\left(\frac{\partial}{\partial\theta'}\hat{\Gamma}(j)\Big|_{\theta=\bar{\theta}}\right)\sqrt{T}\left(\hat{\theta} - \theta_0\right)\right|$$

$$\leqslant CT^{(-3-(2q+1)+4\upsilon)/(4q+2)} \sup_{j\geqslant 1}\left\|\frac{\partial}{\partial\theta}\hat{\Gamma}(j)\Big|_{\theta=\bar{\theta}}\right\| \sqrt{T}\|\hat{\theta} - \theta_0\| \to^p 0,$$

since $\upsilon < 1 + q/2$, $\sqrt{T}\|\hat{\theta} - \theta_0\| = O_p(1)$ by Assumption B(i), and $\sup_{j\geqslant 1}\|(\partial/\partial\theta)\hat{\Gamma}(j)|_{\theta=\bar{\theta}}\| = O_p(1)$ by (A.10) and Assumption B(ii) and (iii), where $\bar{\theta}$ is on the line segment joining $\hat{\theta}$ and θ_0.

We have

$$(A.27) \qquad EG_{2T}^2 \leqslant T^{-3/(2q+1)-1} r^4(T) \sup_{j\geqslant 1} T \operatorname{Var}\left(\tilde{\Gamma}(j)\right) \leqslant CT^{-3/(2q+1)-1+4\upsilon/(2q+1)} \to 0$$

for some $C < \infty$, since $\sup_{j\geqslant 1} T \operatorname{Var}(\tilde{\Gamma}(j)) = O(1)$ by Hannan (1970, equation (3.3), p. 209) and $\upsilon < 1 + q/2$. Also, using Assumption G,

$$(A.28) \qquad G_{3T} \leqslant T^{-3/(4q+2)} C_3 \sum_{j=1}^{\infty} j^{1-m} \to 0,$$

since $m > 2$ implies that $\sum_{j=1}^{\infty} j^{1-m} < \infty$. Equations (A.24)–(A.28) imply $M_{1T} \to^p 0$.

We show $M_{2T} \to^p 0$ as follows: $M_{2T} = A_{1T} + A_{2T} + A_{3T}$, where

$$(A.29) \qquad A_{1T} = T^{q/(2q+1)} \sum_{j=r(T)+1}^{T-1} k\left(j/\hat{S}_T\right)\left(\hat{\Gamma}(j) - \tilde{\Gamma}(j)\right),$$

$$A_{2T} = T^{q/(2q+1)} \sum_{j=r(T)+1}^{T-1} k\left(j/\hat{S}_T\right)\left(\tilde{\Gamma}(j) - \Gamma_T(j)\right), \qquad \text{and}$$

$$A_{3T} = T^{q/(2q+1)} \sum_{j=r(T)+1}^{T-1} k\left(j/\hat{S}_T\right)\Gamma_T(j).$$

By a mean value expansion and the definition of $\mathscr{K}_3$,

$$|A_{1T}| \leq T^{q/(2q+1)-1/2} \sum_{j=r(T)}^{T-1} C_1\left(j/\hat{S}_T\right)^{-b} \left| \left(\frac{\partial}{\partial \theta'} \hat{\Gamma}(j) \Big|_{\theta=\bar{\theta}} \right) \sqrt{T}\left(\hat{\theta}-\theta_0\right) \right| \tag{A.30}$$

$$= T^{q/(2q+1)-1/2+b/(2q+1)} \left(\sum_{j=r(T)}^{\infty} j^{-b} \right) O_p(1)$$

$$= T^{(2q-2q-1+2b-2v(b-1))/(4q+2)} O_p(1) \to^p 0,$$

where the first equality uses (A.10) and Assumption B, and the convergence to zero uses $v > (2b-1)/(2b-2)$. Again by the definition of $\mathscr{K}_3$,

$$|A_{2T}| \leq T^{q/(2q+1)} \sum_{j=r(T)}^{T-1} C_1\left(j/\hat{S}_T\right)^{-b} \left|\tilde{\Gamma}(j) - \Gamma_T(j)\right| \tag{A.31}$$

$$= C_1\left(qk_q^2\hat{\alpha}(q)\right)^{b/(2q+1)} T^{(2b-1)/(4q+2)} \sum_{j=r(T)}^{T-1} j^{-b}\sqrt{T}\left|\tilde{\Gamma}(j) - \Gamma_T(j)\right| \quad \text{and}$$

$$\times E\left(T^{(2b-1)/(4q+2)} \sum_{j=r(T)}^{T-1} j^{-b}\sqrt{T}\left|\tilde{\Gamma}(j) - \Gamma_T(j)\right| \right)^2$$

$$\leq T^{(2b-1)/(2q+1)} \left(\sum_{j=r(T)}^{T-1} j^{-b}\sqrt{T}\,\mathrm{Var}^{1/2}\left(\tilde{\Gamma}(j)\right) \right)^2 \tag{A.32}$$

$$\leq T^{(2b-1)/(2q+1)} \left(\sum_{j=r(T)}^{\infty} j^{-b} \right)^2 O(1) = T^{(2b-1-2v(b-1))/(2q+1)} O(1) \to 0,$$

since $v > 1 + 1/(2b-2)$ and $\sup_{j\geq 1} T\,\mathrm{Var}(\tilde{\Gamma}(j)) = O(1)$ as above. Equations (A.31) and (A.32) combine to yield $A_{2T} \to^p 0$, since $\hat{\alpha}(q) = O_p(1)$.

Using Assumption G and $|k(\cdot)| \leq 1$, we obtain

$$|A_{3T}| \leq T^{q/(2q+1)} \sum_{j=r(T)}^{T-1} C_3 j^{-m} \leq C T^{(q-v(m-1))/(2q+1)} \alpha_\xi^{-v(m-1)/(2q+1)} \to 0 \tag{A.33}$$

for some constant $C < \infty$, since $v > q/(m-1)$ and $\alpha_\xi > 0$.

Equations (A.29)–(A.33) combine to give $M_{2T} \to^p 0$. An analogous argument yields $M_{3T} \to^p 0$. Combined with $M_{1T} \to^p 0$ and (A.23), these results complete the proof of Theorem 3(b).

Next we prove Theorem 3(a). For arbitrary $\alpha_\xi \in (0,\infty)$, $\hat{J}_T(S_{\xi T}) - J_T = o_p(1)$ by Theorem 1(a) (since $q > 1/2$ implies $S_T^2/T \to 0$). Hence, it suffices to show $\hat{J}_T(\hat{S}_T) - \hat{J}_T(S_{\xi T}) = o_p(1)$. This result differs from the result of Theorem 3(b) only because the scale factor $T^{q/(2q+1)}$ does not appear, Assumption F is replaced by Assumption E, Assumption G is not imposed, and $q > 1/2$. The proof of Theorem 3(b) goes through with the following changes: $v \in (1-(2q-1)/(2b-2), 1)$; $T^{q/(2q+1)}$ is deleted in (A.23) and (A.24);

$$\left|\hat{\alpha}(q)^{1/(2q+1)} - \alpha_\xi^{1/(2q+1)}\right| \left(\hat{\alpha}(q)\alpha_\xi\right)^{1/(2q+1)} = O_p(1)$$

in (A.24) by Assumption E; $T^{-3/(4q+2)}$ is replaced by $T^{-1/(2q+1)}$ in (A.25), (A.26), and (A.28); $T^{-3/(2q+1)}$ is replaced by $T^{-2/(2q+1)}$ in (A.28); (A.26) and (A.27) hold provided $v < 3/4 + q/2$; (A.28) is replaced by

$$T^{-1/(2q+1)} \sum_{j=1}^{r(T)} j\left|\Gamma_T(j)\right| \leq T^{-1/(2q+1)} r(T) \sum_{j=1}^{\infty} |\Gamma(j)| = C T^{(v-1)/(2q+1)} \to 0 \tag{A.34}$$

since $v < 1$; $T^{q/(2q+1)}$ is deleted in (A.29)–(A.31) and (A.33); (A.30) holds since $v > 1 - (2q - 1)/(2b - 2)$; $T^{(2b-1)/(4q+2)}$ is replaced by $T^{(2b-2q-1)/(4q+2)}$ in (A.32); (A.32) holds since $v > 1 - (2q - 1)/(2b - 2)$; (A.31), (A.32), and the assumption $\hat{\alpha}(q) = O_p(1)$ yield $A_{2T} \to^p O$; (A.33) is replaced by

$$\text{(A.35)} \qquad \left| \sum_{j=r(T)+1}^{T-1} k\left(j/\hat{S}_T\right)\Gamma_T(j) \right| \leqslant \sum_{j=r(T)}^{\infty} |\Gamma(j)| \to 0,$$

which concludes the proof of Theorem 1(a).

The first equality of Theorem 3(c) holds by applying Lemma A1 in the same way as in the proof of the first equality of Theorem 1(c) (with the reference to Theorem 1(b) changed to Theorem 3(b)). The second equality of Theorem 3(c) holds by Theorem 1(c). *Q.E.D.*

PROOF OF COROLLARY 2: By (7.3) and Theorem 3(c), the left-hand side of the result of the Corollary equals

$$\text{(A.36)} \qquad \lim_{h\to\infty} \lim_{T\to\infty} \left(\text{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(S_T), W_T\right) - \text{MSE}_h\left(T^{2q/(2q+1)}, \hat{J}_T(S_{\xi T}), W_T\right)\right).$$

Since, $\alpha_\xi = \alpha(q)$ implies $S_{\xi T} = S_T^*$, Corollary 1 implies that the expression in (A.36) is $\geqslant 0$ with the inequality being strict unless $S_T = S_T^* + o(T^{1/(2q+1)})$. *Q.E.D.*

REFERENCES

ANDERSON, T. W. (1971): *The Statistical Analysis of Time Series*. New York: Wiley.

ANDREWS, D. W. K. (1988): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," Cowles Foundation Discussion Paper No. 877, Yale University.

——— (1989): "Asymptotics for Semiparametric Econometric Models: I. Estimation and Testing," Cowles Foundation Discussion Paper No. 908R, Yale University.

ANDREWS, D. W. K., AND R. C. FAIR (1988): "Inference in Nonlinear Econometric Models with Structural Change," *Review of Economic Studies*, 55, 615–640.

ANDREWS, D. W. K., AND J. C. MONAHAN (1990): "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator," Cowles Foundation Discussion Paper No. 942, Yale University.

BELTRAO, K. I., AND P. BLOOMFIELD (1987): "Determining the Bandwidth of a Kernel Spectrum Estimate," *Journal of Time Series Analysis*, 8, 21–38.

BRILLINGER, D. R. (1981): *Time Series: Data Analysis and Theory*. New York: Holden-Day.

CAMERON, M. A. (1987): "An Automatic Non-parametric Spectrum Estimator," *Journal of Time Series Analysis*, 8, 379–387.

DAVIDSON, R., AND J. G. MACKINNON (1981): "Efficient Estimation of Tail-area Probabilities in Sampling Experiments," *Economic Letters*, 8, 73–77.

DEHEUVELS, P. (1977): "Estimation Non-parametrique de la Densite par Histogrammes Generalises," *Revue de la Statistique Appliquee*, 25, 5–42.

EICKER, F. (1967): "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. Berkeley: University of California Press, 59–82.

EPANECHNIKOV, V. A. (1969): "Non-parametric Estimation of a Multivariate Probability Density," *Theory of Probability and Its Applications*, 14, 153–158.

GALLANT, A. R. (1987): *Nonlinear Statistical Models*. New York: Wiley.

GALLANT, A. R., AND H. WHITE (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. New York: Basil Blackwell.

HALL, P., AND C. C. HYDE (1980): *Martingale Limit Theory and Its Application*. New York: Academic Press.

HANNAN, E. J. (1970): *Multiple Time Series*. New York: Wiley.

HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

KEENER, R. W., J. KMENTA, AND N. C. WEBER (1987): "Estimation of the Covariance Matrix of the Least Squares Regression Coefficients When the Disturbance Covariance Matrix is of Unknown Form," unpublished manuscript, Department of Statistics, University of Michigan.

KOOL, H. (1988): "A Note on Consistent Estimation of Heteroskedastic and Autocorrelated Covariance Matrices," Research Memorandum 21, Department of Econometrics, Free University, Amsterdam.

LEVINE, D. (1983): "A Remark on Serial Correlation in Maximum Likelihood," *Journal of Econometrics*, 23, 337–342.

MAGNUS, J. R., AND H. NEUDECKER (1979): "The Commutation Matrix: Some Properties and Applications," *Annals of Statistics*, 7, 381–394.

MUIRHEAD, R. J. (1982): *Aspects of Multivariate Statistical Theory*. New York: Wiley.

NEWEY, W. K., AND K. D. WEST (1987): "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

PARZEN, E. (1957): "On Consistent Estimates of the Spectrum of a Stationary Time Series," *Annals of Mathematical Statistics*, 28, 329–348.

PHILLIPS, P. C. B. (1987): "Time Series Regression with a Unit Root," *Econometrica*, 55, 277–301.

PHILLIPS, P. C. B., AND S. OULIARIS (1988): "Testing for Cointegration Using Principal Components Methods," *Journal of Economic Dynamics and Control*, 12, 205–230.

PRIESTLEY, M. B. (1962): "Basic Considerations in the Estimation of Spectra," *Technometrics*, 4, 551–564.

——— (1981): *Spectral Analysis and Time Series*, Volumes I and II. New York: Academic Press.

ROBINSON, P. M. (1988): "Automatic Generalized Least Squares," unpublished manuscript, London School of Economics.

SACKS, J., AND D. YLVISACKER (1981): "Asymptotically Optimum Kernels for Density Estimation at a Point," *Annals of Statistics*, 9, 334–346.

SHEATHER, S. J. (1986): "An Improved Data-based Algorithm for Choosing the Window Width when Estimating the Density at a Point," *Computational Statistics and Data Analysis*, 4, 61–65.

STUART, A., AND J. K. ORD (1987): *Kendall's Advanced Theory of Statistics*, Vol. 1, 5th ed. New York: Oxford University Press.

WAHBA, G. (1980): "Automatic Smoothing of the Log Periodogram," *Journal of the American Statistical Association*, 75, 122–132.

WHITE, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test of Heteroskedasticity," *Econometrica*, 48, 817–838.

——— (1984): *Asymptotic Theory for Econometricians*. New York: Academic Press.

WHITE, H., AND I. DOMOWITZ (1984): "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143–161.