Review of Economic Studies (1994) 61, 631–653 0034-6527/94/00310631$02.00

Automatic Lag Selection in Covariance Matrix Estimation

WHITNEY K. NEWEY
MIT

and

KENNETH D. WEST
University of Wisconsin

First version received September 1992; *final version accepted April* 1994 (*Eds.*)

We propose a nonparametric method for automatically selecting the number of autocovariances to use in computing a heteroskedasticity and autocorrelation consistent covariance matrix. For a given kernel for weighting the autocovariances, we prove that our procedure is asymptotically equivalent to one that is optimal under a mean-squared error loss function. Monte Carlo simulations suggest that our procedure performs tolerably well, although it does result in size distortions.

I. INTRODUCTION

Variance–covariance matrices of estimators of time-series models often must be robust to the presence of heteroskedasticity and autocorrelation of possibly unknown form. In an earlier paper (Newey and West (1987)), we suggested a class of consistent estimators that yielded positive semidefinite matrices by construction. This technique involved calculating weighted sums of estimated autocovariances of cross-products of instruments and residuals. We showed that for a given kernel (a given rule for weighting the autocovariances) it was necessary for consistency to let the bandwidth (the number of autocovariances included) increase with the sample size at an appropriate rate, but otherwise left open the question of how many autocovariances to include, for a given sample.

This is an important theoretical and practical question. An empirical researcher must make a decision on a bandwidth for his chosen kernel not with a sample that is increasing in size, as assumed in asymptotic theory, but with a sample of a specific fixed size. Many rules that asymptotically lead to consistent estimates imply different bandwidths for a given sized sample. While some ambiguity about appropriate choice of bandwidth is inevitable, practitioners would likely find it useful to have a specific, complete rule that could at least be used as a starting point for experimentation with alternative bandwidths. We suggest such a rule, which is data dependent, and is based on both theoretical asymptotic and empirical Monte Carlo results.

As is well known, the matrix we are interested in estimating is proportional to the spectral density of cross-products of instruments and disturbances at frequency zero, and we draw on earlier research on nonparametric density estimation. Such research includes Robinson (1991), who used an approach known as cross-validation to automatically select parameters to smooth spectral density estimates in a non-parametric fashion. Research that is more closely related to ours includes Andrews (1991) and Andrews and Monahan

(1992), who considered the question that we consider (among others), under a similar set of technical conditions, and used, as do we, what is known as a "plug-in" approach. Among other results, they showed how to select a data-dependent bandwidth for a given kernel and sample so as to satisfy an asymptotic mean squared error criterion similar to the one that we use. We build on this work in three ways.

First, their procedures for selecting the bandwidth optimally require the researcher to know the order of the ARMA model governing residual autocorrelation (although misspecification of the order affects only optimality but not consistency). We show how to select the bandwidth optimally when the form of autocorrelation is unknown. Second, we perform Monte Carlo studies that are complementary to theirs. They tried a range of values in some simple models, while we match the point estimates for our artificial data to those from some data used in some actual applications that require a heteroskedasticity and autocorrelation consistent estimator. Third, our procedure is in our opinion somewhat more convenient computationally, since it does not require fitting of an ARMA model, and, for most kernels, allows integer as well as real bandwidths.[1]

Our experiments indicate that selection of bandwidth according to an asymptotically optimal procedure tends to lead to more accurately sized test statistics than do traditional procedures: in one of our two sets of experiments, use of our procedure results in a marked improvement in size of test statistics relative to those of a procedure suggested by Schwert (1987), a (very modest) improvement relative to those of a simple, and in our setup sub-optimal, version of that suggested by Andrews (1991); in another set of experiments, however, all procedures performed roughly comparably. As do Andrews and Monahan (1992), we find that prewhitening with a first-order vector autoregression prior to application of our procedure improves the size of test statistics. But in contrast to Andrews (1991) and Andrews and Monahan (1992), who recommended a kernel called the quadratic spectral, our experiments suggest no firm grounds for preferring this or any other specific kernel. According to these simulations then, if the bandwidth is selected according to our optimal procedure, choice of kernel is of secondary importance, and it may be reasonable to base choice of kernel on grounds such as computational convenience. In our discussion of our Monte Carlo experiments, we suggest a theoretical rationale for why choice of kernel being of second importance is compatible with Andrew's (1991) and Priestley's (1981) proofs of the asymptotic optimality of the quadratic spectral kernel.

We also find, however, that tests often have size distortions even when our procedure is used, as did the tests using the procedures in Andrews (1991) and Andrews and Monahan (1992) for data as serially correlated as are ours. Extensions or refinements to ours or others' procedures therefore remain a priority for future work.

Section II illustrates our procedure in the context of an informal discussion of the relation of this procedure to the literature on estimation of covariance matrices, and is intended to be accessible to the general reader. Section III lays out the theory, and may be skipped without loss of continuity by readers whose main interest is in applying our procedure. Section IV presents the Monte Carlo work. Section V has conclusions. All proofs are in the Appendix. An additional appendix, available on request, contains some proofs and some simulation results omitted from the published paper to save space.

II. INFORMAL OVERVIEW

Suppose that one wishes to estimate the model $y_t = X_t'\theta_0 + u_t$, where y_t and u_t are scalars and X_t and θ_0 are vectors. One has available a $(r \times 1)$ vector of instruments Z_t, with

1. What is computationally convenient to us may not be to others, and some no doubt will find it preferable to fit an ARMA model.

$EZ_t u_t = 0$ and $Z_t u_t$ having serial correlation and heteroskedasticity of unknown form. As is well known (e.g., Hansen (1982)) for inference (and for optimal estimation as well, if the number of instruments r is greater than the number of right-hand side variables) one needs to estimate $S = \sum_{j=-\infty}^{\infty} EZ_t u_t Z_{t-j}' u_{t-j} \equiv \sum_{j=-\infty}^{\infty} \Omega_j = \Omega_0 + \sum_{j=0}^{\infty} (\Omega_j + \Omega_j')$.

Let T be the sample size, let $\hat{\theta}$ be an estimate such that $T^{1/2}(\hat{\theta} - \theta_0)$ is asymptotically normal,[2] let $\hat{u}_t = y_t - X_t \hat{\theta}$, and define the j'th sample autocovariance of $Z_t \hat{u}_t$ as $\hat{\Omega}_j = T^{-1} \sum_{t=j+1}^{T} Z_t \hat{u}_t Z_{t-j} \hat{u}_{t-j}$ for $j \geqq 0$, $\hat{\Omega}_j = \hat{\Omega}_{-j}$ for $j < 0$.[3] As in Andrews (1991), the estimators of S that we consider formally may be written as weighted sums of the $\hat{\Omega}_j$'s. In our Monte Carlo work (but not in our formal analysis), we also consider such estimators applied to the sample autocovariances of residuals from a first-order vector autoregression (VAR) of the $(r \times 1)$ vector $Z_t \hat{u}_t$; see below.

For most although not all weighting schemes of interest the weights are zero for all $j \geqq m+1$ for some *bandwidth* $m+1 \ll T$, and an estimate $\bar{S}$ is constructed as:

$$\bar{S} = \hat{\Omega}_0 + \sum_{j=1}^{m} \bar{k}_j (\hat{\Omega}_j + \hat{\Omega}_j'),$$

where $\{\bar{k}_j\}$ are the weights. For the Bartlett kernel emphasized in our earlier work, for example, $\bar{k}_j = 1 - j/(m+1)$.

The question is how to choose the bandwidth $m+1$. $\bar{S}$ is consistent if $m \to \infty$ as $T \to \infty$ and $m/T^{1/2} \to 0$. But Anderson (1971) and Andrews (1991) show as well that $\bar{S}$ converges to S at different rates for different rules for choosing m. For most kernels, the most rapid possible rate (which will still be less than the $T^{1/2}$ rate familiar from parametric models, given that the estimator is nonparametric) occurs when m increases at the rate of the fifth root of the sample size; the Bartlett kernel, for which the rate is not $T^{1/5}$ but $T^{1/3}$, is an exception.

For the Bartlett kernel, on which we now focus for concreteness, we thus restrict ourselves to choices of m of the form:

$$m = [\text{parameter} \times T^{1/3}], \tag{2.1}$$

where "$[\cdot]$" denotes "integer part of." Now the question is how to choose the "parameter" in (2.1). For expositional convenience, assume for the moment that $r = 1$ so that $\bar{S}$ and S are scalars. When normalized by an appropriate function of sample size, $\bar{S} - S$ is asymptotically $N(b, v)$ for a certain mean b and variance v, where b and v depend on the data. Given the bias ($b \neq 0$) in the limiting distribution the familiar rule of choosing the "parameter" in (2.1) to minimize asymptotic variance does not seem appealing, and authors such as Priestley (1981, p. 568) suggest choosing it to minimize mean squared error (MSE) $b^2 + v$. In the vector case in which $r > 1$, a natural way to reduce the problem to a scalar one is to specify a $(r \times 1)$ weight vector w and to minimize the asymptotic MSE of $w'(\bar{S} - S)w$. The asymptotic mean and variance depend not only on the data but on w as well, and the optimal "parameter" will in general be different for different w's. For a given weight vector w, let $\sigma_j = w'\Omega_j w$, $s^{(1)} = 2\sum_{j=1}^{\infty} j\sigma_j$, $s^{(0)} = \sigma_0 + 2\sum_{j=1}^{\infty} \sigma_j$. Hannan (1971, p. 286) and Priestley (1981, p. 568) show that the (2.1) "parameter" that is optimal by this MSE criterion is $\gamma = 1.1447(s^{(1)}/s^{(0)})^{2/3}$.

In practice, of course, the Ω_j's are not known, so neither are the σ_j's nor $s^{(1)}$ and $s^{(0)}$. But Andrews (1991) shows how to estimate γ so that in an appropriate sense the resulting estimate of S is optimal by a MSE criterion even when θ_0 is unknown. Before illustrating

2. In our formal work, we maintain the milder assumption that $T^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$. Here and throughout this section we are sloppy about such details, to facilitate presentation of a relatively non-technical discussion.

3. Division by T rather than $T-j$ makes our estimators positive semidefinite.

our own procedure, we comment on the generality of procedures allowed both by Andrews and ourselves. These procedures weight the j-th autocovariance by a smooth function of the ratio of j to the bandwidth $m+1$, and are called "scale parameter" kernels (Priestley (1981, p. 446)). This excludes at least two other classes of estimators that are sometimes used in practice.

The first class estimates S by averaging periodogram ordinates. Priestley (1981, pp. 580–582), however, indicates that certain scale parameter kernels, including in particular one called the Daniell, may be interpreted as approximately averaging periodogram ordinates. The Daniell in turn is dominated by the QS kernel in terms of the MSE criterion, and is nearly as complex computationally. So, while we recognize that theoretical or simulation evidence on the behaviour of such estimators may turn up important results, we do not view such analysis as a pressing priority.

The second class fits a vector AR or ARMA to cross-products of instruments and disturbances, and uses standard filtering formulae (e.g., Priestley (1981, pp. 600–604)) to construct an estimate of S. It is not obvious that this second class is well approximated by one or another scale parameter kernel, and detailed consideration of such estimators is in our view a priority for future research. Andrews and Monahan (1992), however, show that if the number of lags in the autoregressive model is very small relative to sample size, so that conventional parametric theory may be applied to the estimates of that model, one can extend the theory of Andrews (1991) to cover estimates of S that combine (a) prewhitening by a low-order VAR with (b) estimation of the spectral density of the VAR residual using a scale-parameter kernel. When step (b) is done with a procedure such as Andrews' or ours, that step will be asymptotically optimal in the class that prewhiten with a given and fixed number of lags in the vector autoregression, but may or may not asymptotically dominate the same procedure applied to the original, non-prewhitened data.

The formal theory in the next section of this paper does not consider prewhitening; see the working paper version of Andrews and Monahan (1992) for some illustrative calculations of when prewhitening is asymptotically preferable. We do however, experiment with prewhitening in the simulations reported in Section IV and find, as did Andrews and Monahan (1992), that prewhitening tends to improve the accuracy of test statistics. Our recommended procedure, then, includes prewhitening, and proceeds as follows.

For clarity, assume that the Bartlett kernel is used. As above, let $\hat{u}_t$ be the scalar regression residual, Z_t be the $(r \times 1)$ vector of instruments. If the first element of Z_t is the constant term, let the weight vector be

$$w=(0\ 1\ 1\ \ldots\ 1)';$$

other choices of w are possible and are discussed briefly in the next section. Also let

$$\begin{aligned}
&\hat{h}_t=Z_t\hat{u}_t,\ \hat{A}=\textstyle\sum_{t=2}^{T}\hat{h}_t\hat{h}_{t-1}'(\sum_{t=2}^{T}\hat{h}_{t-1}\hat{h}_{t-1}')^{-1},\ \hat{h}_t^{\dagger}\equiv\hat{h}_t-\hat{A}\hat{h}_{t-1},\\
&n=[4(T/100)^{2/9}],\\
&\hat{\sigma}_j=(T-1)^{-1}\textstyle\sum_{t=j+2}^{T}\{(w'\hat{h}_t^{\dagger})(w'\hat{h}_{t-j}^{\dagger})\},j=0,\ldots,n,\\
&\hat{s}^{(1)}=2\textstyle\sum_{j=1}^{n}j\hat{\sigma}_j,\ \hat{s}^{(0)}=\hat{\sigma}_0+2\sum_{j=1}^{n}\hat{\sigma}_j,\ \hat{\gamma}=1.1447(\{\hat{s}^{(1)}/\hat{s}^{(0)}\}^2)^{1/3}.
\end{aligned} \tag{2.2}$$

Thus, $\hat{A}$ is the $(r \times r)$ matrix of VAR(1) regression coefficients obtained by regressing cross-products of instruments and residuals on their first lag, $\hat{h}_t^{\dagger}$ is the resulting $(r \times 1)$ vector of period-t residuals; $\hat{\sigma}_j$ is defined using $(T-1)^{-1}$ rather than T^{-1} to account for the observation lost in fitting the VAR(1). We estimate the population quantities $s^{(1)} \equiv$

$2\sum_{j=1}^{\infty} j\sigma_j$ and $s^{(0)} \equiv \sigma_0 + 2\sum_{j=1}^{\infty} \sigma_j$ by truncating these infinite sums at a point indexed by the *lag selection parameter n*. For a given value of m, the estimate of S is constructed as

$$(I-\hat{A})^{-1}\hat{S}^{\dagger}(I-\hat{A})^{-1'},$$

$$\hat{S}^{\dagger} = \{\hat{\Omega}_0 + \sum_{j=1}^{m} \{1-j/(m+1)\}(\hat{\Omega}_j + \hat{\Omega}_j')\}, \hat{\Omega}_j = (T-1)^{-1}\sum_{t=j+2}^{T} \hat{h}_t^{\dagger}\hat{h}_{t-j}^{\dagger\prime}, j=0,\ldots,m.$$

(Note that we have redefined $\hat{\Omega}_j$.) We recommend initially setting

$$m = [\hat{\gamma}T^{1/3}]$$

and then exercising some judgment about sensitivity of results to exact choice of n and m—say, by increasing and decreasing n.

This procedure still involves choice of a parameter, the lag selection parameter n. But there is some evidence from the literature on density estimation that the final result (here, an estimate of the variance–covariance matrix) is less sensitive to n than to m (Silverman (1986, p. 58)).

III. THEORY

We assume that estimation has exploited an orthogonality condition $Eh_t(\theta_0) = 0$, where the $(r \times 1)$ vector h_t is mean zero and covariance stationary (see below) and θ_0 is the unknown parameter. In an ordinary linear least-squares regression, for example, h_t is the vector of cross-products of right-hand side variables and regression disturbance. As is well known (e.g., Hansen (1982)), for inference on $\hat{\theta}$, it is necessary to estimate

$$S = \sum_{j=-\infty}^{\infty} \Omega_j = \Omega_0 + \sum_{j=1}^{\infty} (\Omega_j + \Omega_j'), \Omega_j \equiv Eh_t h_{t-j}'. \tag{3.1}$$

Apart from a factor of 2π, S is the spectral density of h_t at frequency zero.

As in Anderson (1971) and Andrews (1991), the estimators of S that we analyze formally can be written as

$$\hat{S} = \sum_{j=-T+1}^{T-1} k(j/\hat{m}(T))\hat{\Omega}_j = \hat{\Omega}_0 + \sum_{j=1}^{T-1} k(j/\hat{m}(T))(\hat{\Omega}_j + \hat{\Omega}_j'), \tag{3.2}$$

where T is the sample size, $\hat{m}(T)$ is a data dependent *bandwidth*, k is a *kernel*, and $\hat{\Omega}_j$ is an estimate of Ω_j defined below.[4] Examples of kernels include the Bartlett, Parzen and quadratic spectral (hereafter, QS). We take the kernel as given. Our aim in this part of the paper is to develop an automatic procedure for choosing $\hat{m}(T)$ that will be optimal in a sense defined below. On choice of kernel, see Priestley (1981) and Andrews (1991), who, using an asymptotic mean squared error criterion, recommend the QS.

We make the following assumptions on the kernel:

Assumption 1. (a) $k(0) = 1$, $k(x) = k(-x)$, $|k(x)|$ bounded, $\int_{-\infty}^{\infty} |k(x)|\,dx < \infty$, $k(x)$ continuous at zero and all but a finite number of other points; there is a finite, non-zero $q > 0$ that is the largest real number such that

$$\lim_{|x|\to 0} \{1-k(x)\}/|x|^q = c_k,$$

for some $0 < c_k < \infty$.

(b) $|k(x)-k(y)| < c|x-y|$ for some $c > 0$.

4. The formal theory presented in this section does not allow for prewhitening, although it is clear from Andrews and Monahan (1992) our procedure remains valid given the $T^{1/2}$ consistency of the VAR regression coefficients.

(c) $k(x)$ has $[q]+1$ continuous, bounded derivatives on $[0, x]$, for some $x>0$, with the derivatives at $x=0$ evaluated as $x\to 0^+$.

Assumptions 1(a) and 1(b) are standard, appearing in Anderson (1971) or Andrews (1991). Assumption 1(c) is new here, but will not be binding in practice since it holds for all of the kernels in Anderson (1971) and Andrews (1991).

For a matrix $A=(a_{ij})$, let $|\cdot|$ denote the max norm $\max_{i,j}|a_{ij}|$. We make the following assumptions about the data and estimator of θ_0:

Assumption 2. (a) $h_t(\theta)=h(z_t, \theta)$, where $h(z, \theta)$ is measurable in z for all θ, and twice continuously differentiable in θ in a neighbourhood N of θ_0, with probability one.

(b) Let $h_{t\theta}=\partial h_t/\partial\theta$, $h_{it\theta\theta}=\partial^2 h_{it}/\partial\theta\partial\theta'$, where h_{it} is the i'th component of h_t. There is a measurable function $f(z)$ such that $\sup_N |h_t(\theta)|<f(z)$, $\sup_N |h_{t\theta}(\theta)|<f(z)$, $\sup_N |h_{it\theta\theta}(\theta)|<f(z)$, $i=1, \ldots, r$, where for some finite constant D, $E\{f(z_t)^2\}<D$.

(c) $(h_t(\theta_0)'$, vec $(h_{t\theta_0}-Eh_{t\theta_0})')'$ is zero mean and stationary to fourth order; S (defined in 3.1) is positive definite; $(h_t'$, vec $(h_{t\theta_0}-Eh_{t\theta_0})')'$ has absolutely summable fourth cumulants and p-summable autocovariances (so, e.g., $\sum_{j=-\infty}^{\infty}|j|^p|\Omega_j|<\infty$, for Ω_j defined in 3.1).

(d) $T^{1/2}(\hat{\theta}-\theta_0)=O_p(1)$.

Assumption 2 is also made in Andrews (1991).

We make the following assumptions on the relationship between the data and the kernel:

Assumption 3. (a) $p>q+0{\cdot}5+0{\cdot}25q^{-1}$ (where p is defined in Assumption 2, q in Assumption 1).

(b) Let $d=0{\cdot}5\{(2q+1)/(2p+1)\}$. Then $|k(x)|<c|x|^{-b}$ for some $c>0$ and some b satisfying $b>1+\{(1-2d)q-d\}^{-1}$.

Andrews (1991) and Andrews and Monahan (1992) maintain conditions similar to those given in Assumption 3. Note that Assumption 3 implies that $0<\{(1-2d)q-d\}^{-1}$. For any kernel for which $k(x)=0$ for $|x|>1$ (a group that, to our knowledge, includes all kernels used in practice except the QS), Assumption 3(b) holds trivially for arbitrarily large b. For the QS kernel, Assumption 3(b) requires $p>23/4$, a constraint that we discuss below.

We assume that one is interested in estimating $w'Sw$, for some $(r\times 1)$ weight vector w. One has available a sequence of estimates $\{w_T\}$ that converge in probability to w at a suitable rate:

Assumption 4. $T^{q/(2q+1)}(w_T-w)\xrightarrow{p}0$.

It is possible that w_T is non-stochastic, say $w_T=w=(0\ 1\ 1\ \ldots\ 1)'$, as in our Monte Carlo work below. Alternatively, since, in general, one is ultimately interested in estimating not S but a variance–covariance matrix, say, V, which is, say, $(a\times a)$, one might have $w'Sw=(\alpha'H)S(H'\alpha)=\alpha'V\alpha$, for some $(a\times 1)$ weight vector α and some $(a\times r)$ matrix H, with $w_T=\alpha'H_T$. For ordinary least squares, for example, with a vector of right-hand side variables X_t, $H=(EX_tX_t')^{-1}$, $H_T=(T^{-1}\sum_{t=1}^{T}X_tX_t')^{-1}$. In this example and more generally $T^{1/2}(w_T-w)$ is bounded in probability, so Assumption 4 is easily satisfied.

Consider first optimal choice of bandwidth $m(T)$ in the hypothetical case where $h_t \equiv h(z_t, \theta_0)$ rather than $\hat{h}_t \equiv h(z_t, \hat{\theta})$ is observed. Let

$$\sigma_j = w'\Omega_j w, \tag{3.3}$$

$$s^{(q)} = \sum_{j=-\infty}^{\infty} |j|^q w'\Omega_j w \equiv \sum_{j=-\infty}^{\infty} |j|^q \sigma_j \quad \text{for } q \geqq 0;\ 0^0 \equiv 1 \Rightarrow s^{(0)} \equiv \sum_{j=-\infty}^{\infty} \sigma_j, \tag{3.4}$$

$$\tilde{\Omega}_j = T^{-1} \sum_{t=j+1}^{T} h_t h_{t-j}' \quad \text{for } j \geqq 0,\ \tilde{\Omega}_j = \tilde{\Omega}_{-j}' \quad \text{for } j < 0. \tag{3.5}$$

Lemma 1. *Let*

$$\tilde{S} = \sum_{j=-T+1}^{T-1} k_j \tilde{\Omega}_j = \tilde{\Omega}_0 + \sum_{j=1}^{T-1} k_j (\tilde{\Omega}_j + \tilde{\Omega}_j') \tag{3.6}$$

where $k_j \equiv k(x_j)$, $x_j \equiv j/m(T)$, and $\{m(T)\}$ is a non-stochastic sequence of bandwidths such that $m(T) \to \infty$ as $T \to \infty$, with $m(T)^q/T \to 0$, $m(T)/T \to 0$. Suppose that $s^{(q)} \neq 0$. Then no sequence of bandwidths yields a smaller asymptotic mean squared error (smaller $\lim_{T\to\infty} E\{\text{normalized } w'(\tilde{S} - S)w\}^2$) than if $m(T) = \gamma T^{1/(2q+1)}$, where

$$\gamma = c_\gamma \{s^{(q)}/s^{(0)}\}^{2/(2q+1)},\ c_\gamma \equiv \left(q c_k^2 / \int_{-\infty}^{\infty} k^2(x)dx \right)^{1/(2q+1)}. \tag{3.7}$$

The bias-squared component of the MSE is $\{\gamma^{-q} c_k s^{(q)}\}^2$, the variance component is $2\gamma (s^{(0)})^2 \int_{-\infty}^{\infty} k^2(x)dx$; the normalization factor is $T^{q/(2q+1)}$.

Lemma 1 follows by finding the γ that sets to zero the derivative of the standard expression for the mean squared error (Hannan (1971, p. 286), Priestley (1981, p. 568)). See Andrews (1991) for a proof under the conditions assumed here.

In practice, of course, γ is not known, nor can $\tilde{\Omega}_j$ be computed since h_t is not observed. Suitable sample counterparts to the objects in Lemma 1 must therefore be used. To this end define:

$$\hat{\Omega}_j = T^{-1} \sum_{t=j+1}^{T} \hat{h}_t \hat{h}_{t-j}' \quad \text{for } j \geqq 0,\ \hat{\Omega}_j = \hat{\Omega}_{-j}' \quad \text{for } j < 0, \tag{3.8}$$

$$\hat{\sigma}_j = w_T' \hat{\Omega}_j w_T = T^{-1} \sum_{t=j+1}^{T} w_T' \hat{h}_t \hat{h}_{t-j}' w_T, \tag{3.9}$$

$$\hat{s}^{(q)} = \sum_{j=-n}^{n} |j|^q \hat{\sigma}_j \quad \text{for } n, q \geqq 0;\ 0^0 \equiv 1 \Rightarrow \hat{s}^{(0)} = \sum_{j=-n}^{n} \hat{\sigma}_j, \tag{3.10}$$

$$\hat{\gamma} = c_\gamma \{\hat{s}^{(q)}/\hat{s}^{(0)}\}^{2/(2q+1)}, \tag{3.11}$$

$$\hat{k}_j = k(\hat{x}_j),\ \hat{x}_j \equiv j/\hat{\gamma} T^{1/(2q+1)}, \tag{3.12}$$

$$\bar{k}_j = k(\bar{x}_j),\ \bar{x}_j \equiv j/([\hat{\gamma} T^{1/(2q+1)}] + 1), \tag{3.13}$$

$$\hat{S} = \sum_{j=-T+1}^{T-1} \hat{k}_j \hat{\Omega}_j = \hat{\Omega}_0 + \sum_{j=1}^{T-1} \hat{k}_j (\hat{\Omega}_j + \hat{\Omega}_j'), \tag{3.14}$$

$$\bar{S} = \sum_{j=-T+1}^{T-1} \bar{k}_j \hat{\Omega}_j = \hat{\Omega}_0 + \sum_{j=1}^{T-1} \bar{k}_j (\hat{\Omega}_j + \hat{\Omega}_j'). \tag{3.15}$$

In (3.10), the dependence of $\hat{s}^{(q)}$ and $\hat{s}^{(0)}$ on the *lag selection parameter* n is suppressed for simplicity.

As is suggested by (3.10) and (3.11), we will propose obtaining $\hat{\gamma}$ through use of a truncated autocovariance estimator of $\hat{s}^{(0)}$ and $\hat{s}^{(q)}$. In principle, one could use a kernel other than the truncated. We recommend the truncated because it is the easiest to compute and is efficient in a mean squared error sense (Anderson (1971)). Equation (3.11) is to be applied by first squaring the quantity in braces, and then computing the $1/(2q+1)$ root

of the result. It follows that $\hat{\gamma}$ will be positive even if $\hat{s}^{(0)}$ or $\hat{s}^{(q)}$ are negative, so our earlier (1987) reasons for recommending against using a truncated estimator for S itself are not relevant.

Andrews (1991) suggests obtaining $\hat{s}^{(0)}$ and $\hat{s}^{(q)}$ by fitting ARMA models of fixed order and using the estimated coefficients to compute the implied infinite sums; Andrews and Monahan (1992) suggest prewhitening with a fixed-order AR before following the Andrews (1991) procedure. Our own procedure obviously is useful when h_t does not follow an ARMA process. But an advantage likely to be of more relevance in practice will be when it is reasonable to suppose that h_t is well approximated by an ARMA process but the order of the process is unknown. Andrews (1991) and Andrews and Monahan (1992) require a fixed order for the ARMA process and for the prewhitening AR process. But we allow $n \to \infty$ as sample size $\to \infty$, so our procedure still yields the optimal estimator in such a case. Our procedure is also simpler computationally, in our opinion, especially when accurate estimation of $s^{(0)}$ and $s^{(q)}$ requires estimation of a high-order ARMA process for h_t. That high orders might be required in practice is suggested by Cochrane (1988), who, in a related context, has argued that for economic data low-order ARMA processes tend to yield poor estimates of infinite sums of autocovariances such as $s^{(0)}$ and $s^{(q)}$.

In (3.14) and (3.15), $\hat{S}$ and $\bar{S}$ differ only in that $\hat{S}$ uses a real bandwidth (as suggested by Andrews (1991)), $\bar{S}$ an integer bandwidth. We show that the two are asymptotically equivalent for most kernels used in practice (an exception is the QS). Choice between the two thus depends on convenience. To prevent confusion, we note that even in a model with a single instrument ($r=1$), there is a difference between $\hat{s}^{(0)}$ and $\hat{S}$. $\hat{s}^{(0)}$ is a truncated autocovariance estimate, while $\hat{S}$ uses a non-trivial kernel, which according to Assumption 1, must not be the truncated kernel.

The key question is how to choose the lag selection parameter n in (3.10) as a function of sample size and data. For real bandwidths, this is considered in Theorem 1, whose proof is in the Appendix.

Theorem 1. *Assume* $s^{(q)} \neq 0$. *Let* $\hat{\gamma}$ *be estimated so that*

$$T^{\{1-4\varepsilon/(2q+1)\}/2(p-q)} n^{-1} \to 0, \tag{3.16a}$$

$$T^{-4\varepsilon/(2q+1)^2} n \to 0, \tag{3.16b}$$

for some ε *such that*

$$(q+0{\cdot}5)\boldsymbol{d} < \varepsilon < 0{\cdot}5q - (2b-2)^{-1}. \tag{3.17}$$

Then

$$T^{q/(2q+1)}(w_T' \hat{S} w_T - w' \tilde{S} w) \xrightarrow{p} 0. \tag{3.18}$$

The smoothness of the spectral density of $(h_t', \text{vec}\,(h_{t\theta_0})')'$ at frequency 0, which is indexed by p, sets a lower bound on how fast one can increase the lag selection parameter n; the characteristics of the kernel, which are indexed by q and b, set an upper bound. Assumption 3(b) guarantees that in (3.17) $(q+0{\cdot}5)\boldsymbol{d} < 0{\cdot}5q - (2b-2)^{-1}$. In our discussion of Assumption 3 above, we noted that it will hold for the QS kernel only if $p > 23/4$. It is doubtful that a constraint like $p > 23/4$, or any other manifestation of the potentially tight bounds on the rate of increase in n implied by Theorem 1, will be binding in practice, since it is highly unlikely that an investigator would suspect that the autocorrelations of his data die out at a specific slow rate such as that suggested by $p \leqq 23/4$. Insofar as an investigator has a prior on p, it is often that implied by the assumption that h_t and $h_{t\theta_0}$ $\sim$ ARMA of finite order, in which case $p = \infty$ and $\boldsymbol{d} = 0$. Then Assumption 3 will not bind

for any kernels used in practice, (3.16a) is satisfied as long as $n \to \infty$, and the left-most inequality in (3.17) is satisfied as long as $0 < \varepsilon$. Together with (3.16b), the right-most inequality in (3.17) yields the following implied rate of increase of n for some common kernels:

Bartlett: $n \to \infty, n/T^{2/9} \to 0$,
Parzen: $n \to \infty, n/T^{4/25} \to 0$,
Quadratic Spectral: $n \to \infty, n/T^{2/25} \to 0$.

The rate for the Parzen (Bartlett) also applies to the many other kernels for which $q=2$ ($q=1$) and $k(x)=0$ for $|x|>1$; see Anderson (1971) and Priestley (1981) for examples.

We now consider integer bandwidths, which may make for more convenient computation.

Theorem 2. *Let the assumptions of Theorem* 1 *hold, with Assumption* 3(*b*) *strengthened so that*

$$b > \max\,(1 + \{(1-2d)q - d\}^{-1}, 3). \tag{3.19}$$

Then

$$T^{q/(2q+1)}(w_T' \bar{S} w_T - w' \tilde{S} w) \xrightarrow{p} 0. \tag{3.20}$$

The condition (3.19) applies to any kernel for which $k(x)=0$ for $|x|>1$, but does exclude one kernel sometimes used in practice, the QS.

What happens if h_t happens to be serially uncorrelated, but the researcher does not know this and applies our procedure? Theorem 3 establishes that a consistent estimate will still result.

Theorem 3. *Assume that* $Eh_t h_{t-j}' = 0$ *for* $j \neq 0$, *so that* $s^{(q)} = \gamma = 0$. *Then under the conditions of Theorem* 1,

$$\hat{S} \xrightarrow{p} S. \tag{3.22}$$

Under the conditions of Theorem 2,

$$\bar{S} \xrightarrow{p} S. \tag{3.22}$$

It should be noted that it is possible to have $s^{(q)}=0$ even if h_t is serially correlated. A scalar example is $h_t \sim \text{MA}(2)$, $Eh_t^2 = 1$, $Eh_t h_{t-1} = 0$, $Eh_t h_{t-2} = -0{\cdot}25 \Rightarrow s^{(2)} = Eh_t^2 + 4Eh_t h_{t-2} = 0$. It is possible that our procedure will then lead to an inconsistent estimate of S, but such cases clearly are singular.

IV. MONTE CARLO RESULTS

A. Description of Estimators

Using OLS estimation of various regression models, we experiment with Bartlett, Parzen, quadratic spectral (QS) and truncated kernels, in some cases with the VAR(1) prewhitening described in Section II and below. We used integer bandwidths for the Bartlett and Parzen kernels, which satisfy the conditions of Theorem 2, real bandwidths for the QS kernel, which satisfies Theorem 1 but not Theorem 2. The truncated kernel satisfies the conditions of neither theorem (assumptions 1(a) and 1(b) both fail). For this kernel, as well as in some of the computations with the Bartlett and QS kernels, we used bandwidths not chosen in the data-dependent fashion suggested by our theory.

TABLE I

Summary of kernels and estimators

A. Formulas for estimators of S

(1) Bartlett	$\hat{\Omega}_0+\sum_{j=1}^{m}\{1-j/(m+1)\}(\hat{\Omega}_j-\hat{\Omega}_j')$
(2) Parzen	$\hat{\Omega}_0+\sum_{j=1}^{[(m+1)/2]}\{1-6(j/\{m+1\})^2+6(j/\{m+1\})^3\}(\hat{\Omega}_j+\hat{\Omega}_j')$ $+\sum_{j=[(m+1)/2]+1}^{m}2\{1-(j/\{m+1\})\}^3(\hat{\Omega}_j+\hat{\Omega}_j')$
(3) Quadratic Spectral	$\hat{\Omega}_0+\sum_{j=1}^{T-1}k_j(\hat{\Omega}_j+\hat{\Omega}_j')$ $k_j=k_j(x_j)=\frac{25}{12\pi^2x_j^2}\left(\frac{\sin(6\pi x_j/5)}{6\pi x_j/5}-\cos(6\pi x_j/5)\right)$
(4) Truncated	$\hat{\Omega}_0+\sum_{j=1}^{m}(\hat{\Omega}_j+\hat{\Omega}_j')$
(5) Population	$\Omega_0+\sum_{j=1}^{\infty}(\Omega_j+\Omega_j')$

B. Key parameters

	(1) q	(2) c_γ	(3) Max rate of increase of lag selection parameter n	(4) Asymptotic MSE relative to QS
(1) Bartlett	1	1·1447	$o(T^{2/9})$	∞
(2) Parzen	2	2·6614	$o(T^{4/25})$	1·086
(3) QS	2	1·3221	$o(T^{2/25})$	1·0
(4) Truncated	∞	n.a.	n.a.	0·0
(5) Population	n.a.	n.a.	n.a.	0·0

Notes:
1. $\hat{\Omega}_j$ is an estimate of the j-th autocovariance of either (a) the cross-products of instruments and regression disturbances ($\hat{h}_t$), or (b) the residual from a VAR(1) estimated for such cross-products ($\hat{h}_t^\dagger$).
2. Let $\hat{\sigma}_j=w'\hat{\Omega}_j w$, where a given experiment's weight vector w is defined below; let $\hat{\gamma}=c_\gamma\{\hat{s}^{(q)}/\hat{s}^{(0)}\}^{2/(2q+1)}$, where $\hat{s}^{(q)}\equiv 2\sum_{j=1}^{n}j^q\hat{\sigma}_j$, $\hat{s}^{(0)}\equiv\hat{\sigma}_0+2\sum_{j=1}^{n}\hat{\sigma}_j$, q, c_γ, and n are as in Table IB, and the exact values of n used in the experiments are given below; T is the sample size. Then in our recommended procedure, $m=[\hat{\gamma}T^{1/(2q+1)}]$ for the Bartlett and Parzen kernels, where "$[\cdot]$" denotes "integer part of" and m is as in Table IA, $x_j=j/(\hat{\gamma}T^{1/(2q+1)})$ for the QS kernel, where x_j is as in Table IA.

Panel A in Table I lists the formulas for the four kernels, as well as one for the population value that is to be estimated. For notational simplicity, we suppress any data-dependence of the bandwidth m. The formula for the Bartlett kernel, for example, maps into the previous discussion by defining $\bar{k}_j=1-j/(\hat{m}+1)$ for $j\leqq\hat{m}\equiv[\hat{\gamma}T^{1/3}]$, $\bar{k}_j=0$ for $j>\hat{m}$. Panel B lists some key parameters for each kernel. For prewhitened kernels, the estimate of S was adjusted by the estimate of the VAR(1) regression coefficients as described below.

Table II presents a complete list of the kernels used in the Monte Carlo experiments. Panel A describes some estimators that do not use our procedure to obtain the bandwidth or lag truncation parameter. We include these for comparison. In line (1), the rule for selecting the bandwidth of the Bartlett estimator is one used by Schwert (1987) in a study that considered in part finite-sample properties of this estimator in a unit-root context. In line (2), the bandwidth is selected by the sort of procedure suggested by Andrews (1991) and Andrews and Monahan (1992). Let $\hat{\phi}$ be the estimated first-order autocorrelation coefficient of $w'\hat{h}_t$ or $w'\hat{h}_t^\dagger$ ($\hat{h}_t^\dagger$ is the residual after VAR(1) prewhitening, see below). The bandwidth is set to $1{\cdot}3221\ \{4\hat{\phi}^2/(1-\hat{\phi})^4\}^{1/5}T^{1/5}\equiv\tilde{\gamma}T^{1/5}$, which is optimal if $w'h_t$ (or the residual after prewhitening) follows an AR(1) in population (Andrews (1991)). Since that is not the case with our data-generating processes, this procedure is consistent but not as efficient as the one proposed here. In line (3), the truncated estimator was used by, e.g., Hansen and Hodrick (1980). The formula "$[4(T/100)^{1/5}]$" was chosen by analogy to that

TABLE II

Estimators used in Monte Carlo experiments

A. Feasible estimators, bandwidth not chosen by our procedure

Kernel	Formula for bandwidth m	Prewhitened?	Relevant lines in: Table IV	Table V
(1) Bartlett	$[4(T/100)^{1/4}]$	No	1	1
(2) QS[2]	$\tilde{\gamma}T^{1/5}$	No	2	2
		Yes	3	3
(3) Truncated	$[4(T/100)^{1/5}]$	No	4	None
	12		None	4

B. Estimators that are not feasible

(4) Population S	n.a.	No	5	5
(5) Bartlett	$[\gamma T^{1/3}]$	No	6	6

C. Feasible Estimators, Bandwidth Chosen By Our Procedure

Kernel	Formula for lag selection parameter n	Prewhitened?	Relevant lines in: Table IV	Table V
(6) Bartlett	$[4(T/100)^{2/9}]$	No	7	7
	$[12(T/100)^{2/9}]$	No	8	None
	12	No	None	8
	$[3(T/100)^{2/9}]$	Yes	9	9
(7) Parzen	$[4(T/100)^{4/25}]$	No	10	10
	$[12(T/100)^{4/25}]$	No	11	None
	12	No	None	11
	$[3(T/100)^{2/25}]$	Yes	12	12
(8) QS	$[4(T/100)^{2/25}]$	No	13	None
	12	No	None	13

Notes:
1. See notes to Table I.
2. In line (2) $\tilde{\gamma}$ was estimated by computing $s^{(2)}$ and $s^{(0)}$ from an AR(1) fitted to $w'\hat{h}_t$ or $w'\hat{h}_t^{\dagger}$, where a given experiment's w is defined in Table III and $\hat{h}_t$ and $\hat{h}_t^{\dagger}$ are defined in note 1 of Table I.

for the Bartlett estimator in line (1), with the "1/5" exponent somewhat arbitrary; the smaller this exponent, the more efficient asymptotically is this estimator. The formula "$n=$ 12" for the other truncated estimator was chosen because it was known *a priori* in the experiments reported in Table V that all autocovariances after the twelfth were zero.

Panel B describes two estimators that are not feasible in actual application, which we use to gauge the effects of sampling error in estimation of γ and the autocovariances of h_t. The estimator in line (4) uses the population spectral density, the one in line (5) a Bartlett estimator using the population value of γ. The latter sets the number of lagged autocovariance used (m, in the notation of line (1) of Table I) to the non-stochastic vlaue $[\gamma T^{1/3}]$.

Panel C of Table II describes the kernels that choose bandwidths optimally. The "2/9" exponent in line (6), as well as the "4/25" and "2/25" exponents in lines (7) and (8), were chosen to let the lag selecton parameter n increase at the maximum rate allowed by the theory.[5] Results when a "1/9" exponent was used for the Bartlett kernel were similar to the "2/9" results.

5. Strictly speaking, they let n increase at slightly too fast a rate. We use them nonetheless, for notational simplicity, since, for our samples sizes ($\leqq 1000$), the resulting values of n would be the same if we used technically acceptable exponents of 200/901 in line (6), 400/2501 in line (7) and 200/2501 in line (8).)

In panel C, the factors of 4 and 12 were chosen to mimic Schwert (1987). For prewhitened kernels, the factor of $3<4$ was chosen because prewhitening will tend to reduce serial correlation. All prewhitened experiments were also done with a "6" replacing the "3," with very similar results. Once again, some of the kernels set $n=12$ independent of sample size because it was known *a priori* in the experiments reported in Table V that all autocovariances after the twelfth were zero.

All our experiments used OLS estimation. For y_t and X_t defined below, write the OLS regression as $y_t=X_t'\theta_0+u_t$. Let $\hat{\theta}=(\sum_{t=1}^{T} X_tX_t')^{-1}\sum_{t=1}^{T} X_ty_t$ be the OLS estimate, $\hat{u}_t=y_t-X_t'\hat{\theta}$ the OLS residual, $\hat{h}_t=X_t\hat{u}_t$ cross-products of right-hand side variables and residual, $\hat{A}_{LS}\equiv\sum_{t=2}^{T}\hat{h}_t\hat{h}_{t-1}'(\sum_{t=2}^{T}\hat{h}_{t-1}\hat{h}_{t-1}')^{-1}$ the VAR(1) regression estimate, $\hat{A}$ be $\hat{A}_{LS}$ adjusted in a fashion that guarantees that $\hat{A}$ has eigenvalues of modulus less than 0·97 (see below), $\hat{h}_t^{\dagger}\equiv\hat{h}_t-\hat{A}\hat{h}_{t-1}$ the residuals corresponding to the adjusted VAR(1) estimate. Then the asymptotic variance covariance matrix used in computing test statistics was

$$(T^{-1}\textstyle\sum_{t=1}^{T} X_tX_t')^{-1}\text{ (estimate of } S)\,(T^{-1}\textstyle\sum_{t=1}^{T} X_tX_t')^{-1}. \tag{4.1}$$

For non-prewhitened estimators, the estimate of S in (4.1) was computed as indicated in Table I, using the sample autocovariances of $\hat{h}_t$. For prewhitened estimators, the estimate of S in (4.1) was computed as

$$(I-\hat{A})^{-1}\bar{S}^{\dagger}(I-\hat{A})^{-1\prime},\quad\text{or } (I-\hat{A})^{-1}\hat{S}^{\dagger}(I-\hat{A})^{-1\prime},$$

where $\bar{S}^{\dagger}$ and $\hat{S}^{\dagger}$ were computed as indicated in Table I, using the sample autocovariances of $\hat{h}_t^{\dagger}$. In all experiments, the first element of X_t was a constant term and the weight vector w was set to $(0\ 1\ 1\ \ldots\ 1)'$.

The adjustment of $\hat{A}_{LS}$ to insure eigenvalues of modulus less than 0·97 is as in Andrews and Monahan (1992): Let $\hat{B}$ and $\hat{C}$ be $(r\times r)$ matrices whose columns are the eigenvectors of $\hat{A}_{LS}\hat{A}_{LS}'$ and $\hat{A}_{LS}'\hat{A}_{LS}$, $\hat{\Delta}_{LS}=\hat{B}'\hat{A}_{LS}\hat{C}$, $\hat{\Delta}$ the matrix that results when the diagonal elements of $\hat{\Delta}_{LS}$ greater than 0·97 are replaced by 0·97 and those less than $-0{\cdot}97$ are replaced by $-0{\cdot}97$; then $\hat{A}=\hat{B}\hat{\Delta}\hat{C}'$. (As reported in the additional appendix available on request, our results showed little sensitivity to this adjustment, so the procedure that we recommend in part II omits such an adjustment.)

B. Overview of experiments

We performed two sets of experiments, each motivated by a different body of empirical literature. In each experiment, the number of repetitions was 1000, and the same 1000 sets of data were used for all kernels. The first of our two sets, which consisted of two experiments denoted A1 and A2, was stimulated by the literature on testing for Granger causality in the bivariate money-income process. One way to test for Granger causality from money to income is to estimate by OLS a two-sided projection of money onto income and test the null that the coefficients on future money are zero (Sims (1972)). The residual from this projection will in general display serial correlation of an unknown form, and so the procedure we have developed here is relevant.

To calibrate these experiments, we obtained monthly data on M1 and industrial production, seasonally adjusted, 1959:9–1988:2. After taking both monthly and quarterly log differences, we estimated the one-sided projection of $y_t\equiv$ growth in M1 on $x_t\equiv$ growth in industrial production. We also estimated univariate processes for $u_t\equiv$ the residual to this projection and for growth in industrial production. The sample size was 342 for the monthly regression, 110 for the quarterly. Two data-generating processes were then

TABLE III

Description of Artificial Data

A. Experiments A1 and A2

$$y_t = \theta_1 + \theta_2 x_t + \cdots + \theta_6 x_{t-4} + u_t \quad \text{(A1)}$$

$$= \theta_1 + \theta_2 x_t + \cdots + \theta_6 x_{t-4} + \theta_7 x_{t-5} + \theta_8 x_{t-6} + u_t \quad \text{(A2)}$$

$x_t \sim \text{AR}(p)$ with i.i.d. normal innovations,
$u_t \sim \text{AR}(p)$ with i.i.d. normal innovations,
X_t, u_s independent all t, s;

A1: $p=4$; A2: $p=6$;

θ_i's and parameters of x_t and u_t processes given in footnote 6;

A1: estimate $y_t = \theta_1 + \theta_2 x_t + \cdots + \theta_6 x_{t-4} + \theta_7 x_{t+1} + \theta_8 x_{t+2} + u_t$, test H_0: $\theta_7 = \theta_8 = 0$.

A2: estimate $y_t = \theta_1 + \theta_2 x_t + \cdots + \theta_8 x_{t-6} + \theta_9 x_{t+1} + \theta_{10} x_{t+2} + \theta_{11} x_{t+3} + u_t$, test H_0: $\theta_9 = \theta_{10} = \theta_{11} = 0$.

A1: First 10 autocorrelations of $w'h_t = (0\ 1 \ldots 1)X_t u_t = 0{\cdot}226, 0{\cdot}121, 0{\cdot}133, 0{\cdot}124, 0{\cdot}050, 0{\cdot}029, 0{\cdot}024, 0{\cdot}018, 0{\cdot}011, 0{\cdot}008$.

A2: First 10 autocorrelations of $w'h_t = (0\ 1 \ldots 1)X_t u_t = 0{\cdot}307, 0{\cdot}101, 0{\cdot}158, 0{\cdot}021, 0{\cdot}106, 0{\cdot}088, 0{\cdot}021, 0{\cdot}029, 0{\cdot}013, 0{\cdot}007$.

B. Experiments B1 and B2

$$x_t = e_t - f_{t-13},\ e_t = e_{t-1} + \eta_t + \varepsilon_{t-1},\ f_t = e_t + \varepsilon_t;$$

η_t, ε_s independent, $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$; $\sigma_\varepsilon^2 = 5$;

$\eta_t \sim \text{GARCH}(1, 1)$, $\eta_t/(\sigma_{\eta t}^2)^{1/2} \sim N(0, 1)$,

$$\sigma_{\eta t}^2 = 1 + 0{\cdot}05\,\eta_{t-1}^2 + 0{\cdot}85 b \sigma_{\eta, t-1}^2 \quad \text{(B1)}$$

$$\sigma_{\eta t}^2 = 1 + 0{\cdot}30\,\eta_{t-1}^2 + 0{\cdot}60 b \sigma_{\eta, t-1}^2 \quad \text{(B2)}$$

B1 and B2: Estimate $x_t = \theta_1 + \theta_2 x_{t-13} + \theta_3 x_{t-14} + u_t$, test H_0: $\theta_1 = \theta_2 = \theta_3 = 0$.

B1: First 10 autocorrelations of $(0\ 1\ 1)X_t u_t = w'h_t = 0{\cdot}883, 0{\cdot}738, 0{\cdot}606, 0{\cdot}487, 0{\cdot}381, 0{\cdot}288, 0{\cdot}208, 0{\cdot}141, 0{\cdot}087, 0{\cdot}046$

B2: First 10 autocorrelations of $(0\ 1\ 1)X_t u_t = w'h_t = 0{\cdot}874, 0{\cdot}722, 0{\cdot}587, 0{\cdot}468, 0{\cdot}365, 0{\cdot}275, 0{\cdot}200, 0{\cdot}137, 0{\cdot}086, 0{\cdot}048$

Notes:

1. Sample sizes:

A1	A2	B1	B2
100	300	300, 1000	300, 1000

2. In all experiments number of repetitions = 1000.

defined, one in which the parameters were matched to those estimated for the quarterly data (experiment A1), with the other to the monthly estimates (experiment A2).

Panel A of Table III describes the regression models, where, in a slight abuse of notation, the scalar elements of the unknown parameter vector θ_0 are denoted $\theta_1, \theta_2, \ldots, \theta_r$, $r=6$ (experiment A1) or $r=8$ (A2). The data were generated by using the indicated AR(p)'s ($p=4$ for Al, $p=6$ for A2) to generate $T=100$ (A1) or $T=300$ (A2) observations on x_t and u_t, then using the parameters listed in a footnote[6] to generate y_t.

6. Experiment A1: $y_t = 0{\cdot}1575792E\text{-}01 + 0{\cdot}1364678x_t - 0{\cdot}1199439x_{t-1} - 0{\cdot}3374262E\text{-}01x_{t-2} - 0{\cdot}3113678E\text{-}01x_{t-3} - 0{\cdot}1205284E\text{-}01x_{t-4} + u_t$, $x_t = 0{\cdot}005911203 + 0{\cdot}4717572x_{t-1} - 0{\cdot}07913229x_{t-2} + 0{\cdot}04288376x_{t-3} - 0.07724863x_{t-4} + \varepsilon_{1t}$; $\varepsilon_{1t} \sim N(0, \sigma_1^2)$, $\sigma_1 = 0{\cdot}01873203$; $u_t = 0{\cdot}18866313u_{t-1} + 0{\cdot}05309064u_{t-2} + 0{\cdot}1041030u_{t-3} + 0{\cdot}1213361u_{t-4} + \varepsilon_{2t}$; $\varepsilon_{2t} \sim N(0, \sigma_2^2)$, $\sigma_2 = 0{\cdot}0096223922$. Experiment A2: $y_t = 0{\cdot}5031414E\text{-}02 + 0{\cdot}2002929E\text{-}01x_t + 0{\cdot}2521308\text{-}02E\text{-}02x_{t-1} + 0{\cdot}2052117E\text{-}01x_{t-2} - 0{\cdot}4798466E\text{-}01x_{t-3} - 0{\cdot}3796021E\text{-}01x_{t-4} + 0{\cdot}1133342E\text{-}03x_{t-5} - 0{\cdot}1039753E\text{-}01x_{t-6} + u_t$; $x_t = 0{\cdot}001327165 + 0{\cdot}3821891x_{t-1} + 0{\cdot}04460943x_{t-2} + 0{\cdot}05424138x_{t-3} + 0{\cdot}08620320x_{t-4} - 0{\cdot}06192332x_{t-5} + 0{\cdot}0186440x_{t-6} + \varepsilon_{1t}$; $\varepsilon_{1t} \sim N(0, \sigma_1^2)$, $\sigma_1 = 0{\cdot}00811155$; $u_t = 0{\cdot}3367596u_{t-1} - 0{\cdot}07424469u_{t-2} + 0{\cdot}1934595u_{t-3} - 0{\cdot}1437814u_{t-4} + 0{\cdot}1710659u_{t-5} + 0{\cdot}01611501u_{t-6} + \varepsilon_{2t}$; $\varepsilon_{2t} \sim N(0, \sigma_2^2)$, $\sigma_2 = 0{\cdot}00454424$.

The actual initial historical values of money growth were used as initial conditions in generating x_t. The initial u_t's were set to zero, and the first 100 observations generated were discarded (i.e., observations 101–200 (A1) or 101–400 (A2) were used to generate y_t). The innovations in x_t and u_t were assumed normal and independent, so there is no conditional heteroskedasticity. The population autocorrelations of $w'h_t$ reported in the Table were computed analytically. So, too, were the autocovariances of h_t, the first 60 of which were used to compute the population value of S (row 4 of Table II, row 5 of Table IV). (The infinite sum was approximated by the sum of the first 60 terms because for both A1 and A2 the last 20 lags (lags 41 through 60) caused $s \equiv w'Sw$ to change by less than 10^{-4} of one percent.)

Panel B of Table III describes the setup of our second set of experiments, which was motivated by the literature that tests whether the log of a forward exchange rate (f_{t-13} in the notation of panel B) is an efficient predictor of the log of the corresponding spot rate (denoted e_t)—i.e., whether $f_{t-13} = E_{t-13}e_t$.[7] The lag of 13 comes from Hansen and Hodrick (1980), who used weekly data and a 13-week-ahead forward rate. The test we use is one of theirs, obtained by regressing the realized difference between the two rates ($x_t \equiv e_t - f_{t-13}$) on a constant and x_{t-13} and x_{t-14}, and testing whether all three coefficients are zero. Under the null, cross-products of regressors and the disturbance will follow a MA(12) process.

As indicated in panel B, we assume that e_t follows a random walk with GARCH(1, 1) disturbances, a process consistent with the results of many recent studies of weekly bilateral dollar exchange rates (e.g., West and Cho (1994)); h_t will therefore be conditionally heteroskedastic in these experiments. Since the coefficients on η_{t-1}^2 and $\sigma_{\eta,t-1}^2$ sum to 0·9, the data display the substantial serial correlation in the conditional variance of $e_t - e_{t-1}$ that is suggested by such studies. When the coefficient on η_{t-1}^2 is 0·05 (experiment B1), the formulas in Bollerslev (1986) indicate that h_t has finite fourth moments, as is required by our theory. When this coefficient is 0·30 (experiment B2), h_t has finite second but not third moments; this is inconsistent with our theory but seemed worth studying since some empirical estimates do indeed imply that such moments do not exist.[8] The variances of e_t and η_t were chosen so that the implied unconditional variances of $e_t - e_{t-1}$ and $e_t - f_{t-13}$ matched those of weekly data for the Deutschmark-dollar, 1971–1991.

For both B1 and B2, two sample sizes were used: $T = 300$ is roughly that of Hansen and Hodrick (1980), $T = 1000$ roughly that currently available to a researcher using weekly data from the current floating exchange rate era. To generate a data set, the initial $\sigma_{\eta 0}^2$ was set to the unconditional variance of $e_t - e_{t-1}$ and η_0 was drawn from a $N(0, \sigma_{\eta 0}^2)$ distribution. 1100 observations were then generated, the first 100 of which were thrown away. Observations 101–400 were used when $T = 300$, observations 101–1100 when $T = 1000$. Once again, the population autocovariances of $w'h_t$ and of h_t were computed analytically.

C. *Simulation Results*

Table IV has sizes of nominal 1, 5 and 10% tests for experiments A1 and A2. All the feasible kernels over-reject, with sizes of nominal 5% tests, for example, ranging from

7. That this follows under the indicated data-generating process may be seen by beginning with $e_t = e_{t-1} + \eta_t + \varepsilon_{t-1}$ and then recursively substituting out for e_{t-1}, then $e_{t-2}, \ldots$, then e_{t-12}, yielding $e_t = e_{t-13} + \varepsilon_{t-13} + \eta_t + \sum_{i=1}^{12} (\eta_{t-13+i} + \varepsilon_{t-13+i}) \Rightarrow E_{t-13}e_t = e_{t-13} + \varepsilon_{t-13}$.

8. See Hansen (1992) on consistency of covariance matrix estimators when fourth moments do not exist. As Hansen (1992, p. 969) notes, the asymptotic mean squared error criterion we use is not obviously applicable when such moments do not exist.

TABLE IV

Sizes of nominal 1, 5, *and* 10% *Tests, Experiment A*

(1)	(2)	(3a)	(3b)	(4a)	(4b)	(4c)	(5a)	(5b)	(5c)
		Bandwidth m or lag selection parameter n		Experiment A1 Size			Experiment A2 Size		
Kernel	PW?	A1	A2	1·0	5·0	10·0	1·0	5·0	10·0
1. Bartlett		4	5	4·5	12·4	18·6	2·7	8·1	14·0
2. QS-AR(1)		n.a.	n.a.	3·6	11·6	17·4	3·0	8·5	15·1
3. QS-AR(1)	y	n.a.	n.a.	3·9	11·2	17·3	2·3	7·3	13·4
4. Truncated		4	5	9·0	18·3	24·6	4·8	11·8	18·2
5. Population S		n.a.	n.a.	0·3	1·4	3·8	0·5	2·5	5·8
6. Bartlett, population γ		8	13	5·9	14·9	20·9	4·5	10·7	16·7
7. Bartlett		4	5	4·6	13·1	19·7	3·5	9·3	15·0
8. Bartlett		12	15	11·0	19·7	27·0	5·2	11·6	17·9
9. Bartlett	y	3	3	4·8	12·9	19·1	3·0	8·5	14·1
10. Parzen		4	4	5·5	15·3	21·7	3·3	9·8	15·4
11. Parzen		12	15	17·0	25·6	34·1	7·9	14·7	21·5
12. Parzen	y	3	3	5·1	13·4	20·8	3·3	9·8	15·5
13. QS		4	4	5·0	14·0	21·4	3·1	9·1	15·3

Notes:

1. Column 1 gives the kernel. See Table I.
2. In column 2, a "y" indicates that VAR(1) prewhitening was done prior to kernel-based estimation of the spectral density.
3. For m and n defined in the notes to Table I, columns 3a and 3b, rows 1, 4 and 6 give the bandwidth m; other non-zero entries give lag selection parameter n; $T=100$ in Experiment A1, $T=300$ in Experiment A2. The two experiments differ not only in sample size but in other dimensions as well; see text. The rule used to choose the lag selection parameter n is presented in Table II.
4. In columns (4a), (4b), and (4c), rows 1–13 give the actual sizes of nominal 1, 5 and 10% tests, for experiment A1. Columns (5a), (5b), and (5c) do the same for experiment A2.

about 10 to about 25%. So, too, does the non-feasible Bartlett kernel that uses the population bandwidth (row (6)). That the (non-feasible) kernel using the population S under- rather than over-rejects (row (5)) indicates that error in estimation of the autocovariances is in part responsible for the over-rejections. Comparison of rows 2 and 3, 7 and 9, and 10 and 12 indicates that prewhitening usually leads to a small improvement.

Table IV suggests to us the following. First, in this experiment, there are no firm grounds for preferring our procedure over the others we consider; indeed, to a referee and perhaps others it suggests that the QS–AR(1) procedure, which in its prewhitened form (line (3)) was most accurately sized of all feasible estimators in all columns but (4a), should be used. Second, in samples as small as that in these experiments ($T=100$ and $T=300$), our procedure performs less well with a relatively large lag selection parameter: compare lines 7 and 8, and lines 10 and 11. Third, within the class of estimators that use our procedure, the experiment suggests no particular grounds for preferring one kernel over another. The Bartlett, Parzen and QS kernels in lines 7, 10 and 13, each of which use a lag selection parameter of 4 or 5 (i.e., $4 \leqq n \leqq 5$), all perform comparably.

This last point will apply in our second set of experiments, so we pause here to suggest a theoretical rationale for it. Consider an analytical expression for the finite-sample mean squared error in estimating $s \equiv w'Sw$, obtained by dividing an asymptotic biased squared and an asymptotic variance by appropriate functions of sample size. The bias squared and variance are those for a hypothetical estimator such as that in Lemma 1 that uses cross-products of regressors and unobservable disturbances rather than cross-products of regressors and OLS residuals. That is, for a given kernel, use the population values of $s^{(q)}$, s

and γ to compute

$$T^{-2q/(2q+1)}(s^{2q}s^{(q)} \cdot \text{constant})^{2/(2q+1)} \tag{4.2}$$

where $T=100$ or $T=300$, $q=1$ for the Bartlett kernel, $q=2$ for Parzen and QS kernels, and the constant varies from kernel to kernel. It may be shown that apart from the factor of $T^{-2q/(2q+1)}$, (4.2) is the mean squared error given by Lemma 1.

Expressed as a ratio to the figure for QS, the resulting figures are as follows. Bartlett: 0·76 (experiment A1), 0·88 (A2); Parzen: 1·09 (A1 and A2). It may appear surprising that the MSE for Bartlett is smaller than that for QS, since the QS is asymptotically optimal by our mean squared error criterion among all kernels that generate positive semidefinite estimates (Andrews (1991)). To understand why this theoretical figure is lower for the Bartlett, note first that the MSE for the Bartlett is computed as (asymptotic MSE)/$T^{2/3}$, for the QS as (asymptotic MSE)/$T^{4/5}$; with a big enough sample, QS will be more efficient. But for our data-generating processes, $T=300$ evidently is not sufficiently big. The key feature of our data-generating process that makes the Bartlett MSE relatively small is that $s^{(1)} \equiv 2\sum_{j=1}^{\infty} j\sigma_j$ (which is relevant for the Bartlett) is small relative to $s^{(2)} \equiv 2\sum_{j=1}^{\infty} j^2\sigma_j$ (which is relevant for the QS kernel). This in turn will tend to be true if the autocorrelations of h_t are positive and die out slowly, even if, as indicated in panel A of Table III, all but a few autocorrelations are small enough that they might be ignored in traditional Box–Jenkins analysis. As noted in the previous section, Cochrane (1988) argues that this is a possibility with economic data.

Now, the small magnitude of the 1·09 figure for the relative MSE of the Parzen, and similarly small figures for some other kernels for which $q=2$, led Priestley (1981, p. 574) to suggest that which of such kernels one uses is of secondary importance; as argued above, for economic data, which often seem to have high-order autocorrelations that are of the same (positive) sign, it is our view that as long as one chooses the bandwidth optimally, choice of kernel may well be of secondary importance even if one considers kernels with $q=1$ as well.

In the actual experiments, the MSE's of the estimators in lines 7 and 10 (again, relative to that of QS in line 13) were as follows. Bartlett: 1·01 (A1), 0·99 (A2); Parzen: 1·03 (A1), 0·99 (A2). Thus in practice as well as according to (4.2) QS did not uniformly dominate other estimators in terms of MSE. However, here and for other kernels, the very simple asymptotic approximation (4.2) did not perfectly predict the ordering nor the dispersion of the relative MSE's. This indicates that sampling error in estimation of the regression vector, which is ignored in (4.2), is important in practice. More generally, it illustrates what to us is disappointingly small guidance of the asymptotic theory for the behaviour of the estimators in experiment A.

The theory, however, is more useful in experiment B. Table V has results analogous to those reported in Table IV. As in experiment A, all feasible kernels over-reject, as does the Bartlett estimator that uses the population γ (line (6)). Interestingly, each kernel performs about as well for B1 (for which the asymptotic theory has been shown to apply) as for B2 (for which it has not).

Nonetheless, the asymptotic theory is useful here. First, rejection frequencies are closer to nominal levels for sample sizes of 1000 than 300 (compare column (5) to (4) and column (7) to (6)). Second, the Bartlett estimators that use the data-dependent bandwidth (lines (7) and (8)) perform markedly better than the Bartlett estimator that does not (line (1)), while the QS estimator that chooses the bandwidth optimally (line (13)) provides a (very) modest improvement over one that does not (line (2)). Third, with the larger sample

TABLE V

Sizes of nominal 1, 5, *and* 10% *tests, Experiment B*

(1)	(2)	(3a)	(3b)	(4a)	(4b)	(4c)	(5a)	(5b)	(5c)	(6a)	(6b)	(6c)	(7a)	(7b)	(7c)
						Experiment B1						Experiment B2			
Kernel	PW?	Bandwidth or lag selection parameter			$T=300$ Size			$T=1000$ Size			$T=300$ Size			$T=1000$ Size	
		$T=300$	$T=1000$	1·0	5·0	10·0	1·0	5·0	10·0	1·0	5·0	10·0	1·0	5·0	10·0
1. Bartlett		4	5	17·8	31·1	41·5	10·7	23·0	30·9	17·3	32·3	42·2	10·7	23·6	33·3
2. QS–AR(1)		n.a.	n.a.	16·2	27·5	36·3	6·3	14·4	24·2	16·5	27·2	36·6	6·6	15·9	24·6
3. QS–AR(1)	*y*	n.a.	n.a.	10·4	20·0	26·0	4·0	9·8	15·6	10·7	20·7	27·6	5·4	12·2	20·4
4. Truncated		12	12	14·7	23·9	30·9	3·5	12·2	18·3	14·7	24·7	32·3	4·8	13·0	21·5
5. Population S		n.a.	n.a.	2·5	6·8	11·2	1·5	5·9	11·5	0·9	1·7	2·6	0·7	2·3	3·3
6. Bartlett, population γ		16	24	11·5	21·4	30·6	5·0	14·9	22·4	12·6	23·5	31·1	5·5	14·9	23·8
7. Bartlett		5	6	11·1	20·9	30·7	5·0	15·2	22·0	11·5	22·5	31·2	5·5	15·6	23·9
8. Bartlett		12	12	14·8	24·0	32·5	5·7	15·6	22·9	15·1	26·3	34·4	5·6	16·4	25·0
9. Bartlett	*y*	3	5	9·2	17·3	24·2	2·5	8·1	14·0	9·2	19·0	24·9	4·1	9·1	16·5
10. Parzen		4	5	11·3	20·0	28·8	4·4	14·6	21·0	11·3	21·4	30·9	4·8	13·6	22·5
11. Parzen		12	12	14·4	23·1	31·4	4·7	14·3	21·6	14·4	23·8	33·2	5·6	14·5	23·6
12. Parzen	*y*	3	4	8·8	16·8	24·0	2·6	7·6	13·6	8·7	18·3	23·6	3·5	8·7	16·0
13. QS		12	12	14·1	23·0	31·9	4·8	15·0	21·9	14·4	24·2	33·3	5·6	14·3	23·8

Notes:
1. See notes to Table IV.

size, there is not much sensitivity to how the lag selection parameter n is chosen (in columns 5 and 7, compare lines 7 and 8, 10 and 11).

In this experiment, the VAR(1) prewhitening results in a distinct improvement in the accuracy of test sizes. Compare line 2 versus 3, lines 7 and 8 versus 9, or lines 10 and 11 versus 12. In fact, among the feasible kernels, the two prewhitened kernels with optimal bandwidths are invariably the most accurately sized (lines 9 and 12), and the prewhitened QS–AR(1) (line 3) is usually but not always the third most accurate.

Once again, it seems that choice of kernel is secondary, given our rule for selecting the bandwidth. Compare (a) lines 7, 8, 10, 11 and 13, or (b) lines (9) and (12).

Nonetheless, all the feasible kernels have substantial size distortions. This is troubling, but perhaps unsurprising, in light of the results of Monte Carlo experiments performed by other authors. Table III indicates that the first-order autocorrelation of these data is 0·88 (B1) or 0·87 (B2). For data-generating processes whose first-order autocorrelations are of comparable magnitude and sample sizes roughly those of experiments B1 and B2, Andrews (1991), Keener, Kmenta and Weber (1991), Andrews and Monahan (1992), and Christiano and den Haan (1993) all find comparable tendencies to over-reject in their preferred procedures for estimating a variance-covariance matrix.

V. CONCLUSIONS

We have proposed a computationally convenient procedure for automatically selecting the number of lags to use in computing a heteroskedasticity and autocorrelation consistent variance–covariance matrix. Monte Carlo experiments provide some support for use of the procedure, and suggest that careful selection of the number of lags may be more important than choice of kernel. They also indicate that more accurate test statistics result if prewhitening (Andrews and Monahan (1992)) is combined with our procedure. Nonetheless, substantial size distortions remain. An important task for future research is refining or extending ours or others' procedures to get estimators whose actual size is closer to nominal size. Because of the relatively good performance of the prewhitened estimators, a priority is theoretical and empirical investigation of autoregressive or autoregressive-moving average spectral estimators (e.g., Berk (1974)).

APPENDIX

Using arguments such as those below, it is straightforward to show that $\hat{S}$ and $\bar{S}$ are $O_p(1)$. By Assumption 4, then, $T^{q/(2q+1)}(w_T'\hat{S}w_T - w'\hat{S}w) \xrightarrow{p} 0$, $T^{q/(2q+1)}(w_T'\bar{S}w_T - w'\bar{S}w) \xrightarrow{p} 0$. In the proofs of Theorems 1 to 3, it therefore suffices to consider the scalars $\hat{s} \equiv w'\hat{S}w$ and $\bar{s} \equiv w'\bar{S}w$, and we redefine $\hat{\sigma}_j \equiv w'\hat{\Omega}_j w$ where $\hat{\sigma}_j$ was originally defined in (3.9). Let $s \equiv w'Sw$,

$$\tilde{s} = w'\tilde{S}w = \sum_{j=-T+1}^{T-1} k_j w'\tilde{\Omega}_j w \equiv \sum_{j=-T+1}^{T-1} k_j \tilde{\sigma}_j, \tag{A.1}$$

$$\tilde{s}^{(q)} = \sum_{j=-n}^{n} |j|^q \tilde{\sigma}_j, \tag{A.2}$$

where $\tilde{S}$ and $\tilde{\Omega}_j$ are defined in (3.6). For notational simplicity, we assume that h_t is a scalar (i.e., we do not distinguish between $w'h_t$ and h_t). In the proofs, c or c_i is a generic constant, not necessarily the same from equation to equation.

To conserve space, we omit the proof of Theorem 2, which is similar to that of Theorem 1. A sketch of the proof is available on request.

Lemma A1. *If $n \to \infty$ as $T \to \infty$ such that $n^{2q+1}/T \to 0$, (T/n^{2q+1}) var $(\tilde{s}^{(q)}) = O(1)$.*

Proof. The case for $q = 0$ is in Anderson (1971, p. 531). Therefore, assume $q \neq 0$, in which case *var* $(\tilde{s}^{(q)}) = \text{var}\,(2\sum_{j=1}^{n} j^q \tilde{\sigma}_j) = 4\sum\sum_{i,j=1}^{n} (ij)^q \text{ cov}\,(\tilde{\sigma}_i, \tilde{\sigma}_j)$.

Let $\kappa(\cdot,\cdot,\cdot)$ be the fourth cumulant of h_t. From Anderson (1971, p. 527), $\operatorname{cov}(\tilde{\sigma}_i, \tilde{\sigma}_j) = T^{-2}\sum_{t=j+1}^{T}\sum_{s=i+1}^{T}[\sigma_{t-s}\sigma_{t-j-s+i}+\sigma_{t-s+i}\sigma_{t-j-s}+\kappa(-i, s-t, s-t+j)]$. We therefore have

$$\begin{aligned}\operatorname{var}(\tilde{s}^{(q)}) &= 4T^{-2}\sum\sum_{i,j=1}^{n}(ij)^q\sum_{t=j+1}^{T}\sum_{s=i+1}^{T}(\sigma_{t-s}\sigma_{t-j-s+i})\\ &\quad+4T^{-2}\sum\sum_{i,j=1}^{n}(ij)^q\sum_{t=j+1}^{T}\sum_{s=i+1}^{T}(\sigma_{t-s+i}\sigma_{t-j-s})\\ &\quad+4T^{-2}\sum\sum_{i,j=1}^{n}(ij)^q\sum_{t=j+1}^{T}\sum_{s=i+1}^{T}\kappa(-i, s-t, s-t+j)\\ &\equiv 4V_1+4V_2+4V_3 \qquad \text{(A.3)}\end{aligned}$$

The desired result will follow if it can be shown that V_1, V_2 and V_3 are each $O(T^{-1}n^{2q+1})$. Consider first V_1. Making the change of variable $k=t-s$, we get

$$\begin{aligned}|V_1| &= T^{-2}\Big|\sum\sum_{i,j=1}^{n}(ij)^q\sum_{t=j+1}^{T}\sum_{k=i+1-t}^{T-t}\sigma_k\sigma_{k+i-j}\Big|\\ &\leqq T^{-2}\sum\sum_{i,j=1}^{n}(ij)^q\sum_{t=1}^{T}\sum_{k=-T+1}^{T-1}|\sigma_k\sigma_{k+i-j}|\\ &= T^{-1}\sum\sum_{i,j=1}^{n}(ij)^q\sum_{k=-T+1}^{T-1}|\sigma_k\sigma_{k+i-j}|\\ &= T^{-1}\sum_{v=-n+1}^{n-1}\sum_{r=|v|+1}^{n}[r(r-|v|)]^q\sum_{k=-T+1}^{T-1}|\sigma_k\sigma_{k+v}|\\ &\leqq T^{-1}\sum_{v=-n+1}^{n-1}\sum_{r=1}^{n}r^{2q}\sum_{k=-T+1}^{T-1}|\sigma_k\sigma_{k+v}|\\ &= T^{-1}\Big(\sum_{r=1}^{n}r^{2q}\Big)\Big(\sum_{v=-n+1}^{n-1}\sum_{k=-T+1}^{T-1}|\sigma_k\sigma_{k+v}|\Big)\\ &\leqq T^{-1}\Big(\sum_{r=1}^{n}r^{2q}\Big)\Big(\sum_{k=-T-n}^{T+n}|\sigma_k|\Big)^2\\ &\leqq T^{-1}\Big(\sum_{r=1}^{n}r^{2q}\Big)\Big(\sum_{k=-\infty}^{\infty}|\sigma_k|\Big)^2\\ &= O(T^{-1}n^{2q+1})\Big(\sum_{k=-\infty}^{\infty}|\sigma_k|\Big)^2 = O(T^{-1}n^{2q+1}). \qquad \text{(A.4)}\end{aligned}$$

That $V_2 = O(T^{-1}n^{2q+1})$ can be shown similarly. So consider V_3 in equation (A.3). From Anderson (1971, p. 530) we see that $V_3 = T^{-1}\sum\sum_{i,j=1}^{n}(ij)^q\sum_{-T+1}^{T-1}\kappa(i,-t,j-t)\phi_T(t,i,j)$, where ϕ_T is a certain nonstochastic function satisfying $0\leqq\phi_T\leqq 1$. Thus

$$\begin{aligned}|V_3| &\leqq T^{-1}\sum\sum_{i,j=1}^{n}(ij)^q\sum_{t=-T+1}^{T-1}|\kappa(i,-t,j-t)|\\ &\leqq T^{-1}n^{2q}\sum\sum_{i,j=1}^{n}\sum_{t=-T+1}^{T-1}|\kappa(i,-t,j-t)|\\ &\leqq T^{-1}n^{2q}\sum\sum\sum_{t,t',t''=-\infty}^{\infty}|\kappa(t,t',t'')|\\ &= O(T^{-1}n^{2q}). \quad \|\end{aligned}$$

Lemma A2. *Let $n\to\infty$ as $T\to\infty$ in such a fashion that for some d, $0<d\leqq 1/2$.*

$$T^{(1-2d)/[2(p-q)]}n^{-1}\to 0, \qquad \text{(A.5a)}$$

$$T^{-2d/(2q+1)}n\to 0. \qquad \text{(A.5b)}$$

Then

$$T^{(1/2)-d}(\tilde{s}^{(q)}-s^{(q)})\xrightarrow{p}0.$$

Proof. We have $T^{1-2d}\operatorname{var}(\tilde{s}^{(q)}) = (T^{-2d}n^{2q+1})(T/n^{2q+1})\operatorname{var}(\tilde{s}^{(q)})\to 0$ by Lemma A1, since (A.5b) states that $T^{1-2d}(n^{2q+1}/T)\to 0$ and thus $n^{2q+1}/T\to 0$. The result therefore will follow if $|T^{0.5-d}(s^{(q)}-E\tilde{s}^{(q)})|^2\to 0$. Since $E\tilde{\sigma}_j = [(T-j)/T]\sigma_j$, $\sigma_j - E\tilde{\sigma}_j = T^{-1}j\sigma_j$ and

$$\begin{aligned}&|T^{0.5-d}(s^{(q)}-E\tilde{s}^{(q)})|\\ &\quad= 2T^{0.5-d}\Big|\sum_{j=n+1}^{\infty}j^q\sigma_j+T^{-1}\sum_{j=1}^{n}j^{q+1}\sigma_j\Big|\\ &\quad\leqq 2T^{0.5-d}\sum_{j=n+1}^{\infty}j^q|\sigma_j|+2T^{-1.5-d}\sum_{j=1}^{n}j^{0.5}|j^{q+0.5}\sigma_j|\\ &\quad\leqq 2T^{0.5-d}(n^{-(p-q)})\sum_{j=n+1}^{\infty}j^p|\sigma_j|+2T^{-1.5-d}n^{0.5}\sum_{j=1}^{\infty}|j^{q+0.5}\sigma_j|\to 0,\end{aligned}$$

since $\sum_{j=1}^{\infty}j^p|\sigma_j|<\infty$ for some $p>q+0.5$ by assumption 3(a), $T^{0.5-d}(n^{-(p-q)})\to 0$ by (A.5a), and $n/T\to 0$ by (A.5b). $\|$

Lemma A3. *For $0<d<1/2$, if $n\to\infty$ as $T\to\infty$ such that (A.5b) holds, $T^{(1/2)-d}(\hat{s}^{(q)}-\tilde{s}^{(q)})\xrightarrow{p}0$.*

Proof. A second-order Taylor series expansion of $\hat{h}_t$ around h_t yields $\hat{h}_t=h_t+h_{t\theta}(\hat{\theta}-\theta_0)+0{\cdot}5\overset{*}{h}_{t\theta\theta}(\hat{\theta}-\theta_0)^2$, where $\overset{*}{h}_{t\theta\theta}$ is evaluated at a point between $\hat{\theta}$ and θ_0. Let $\bar{h}_\theta\equiv Eh_{t\theta}$. Then for $j\geqq 0$, $\hat{\sigma}_j-\tilde{\sigma}_j=(\hat{\theta}-\theta_0)R_{1j}+\bar{h}_\theta(\hat{\theta}-\theta_0)R_{2j}+(\hat{\theta}-\theta_0)^2R_{3j}+(\hat{\theta}-\theta_0)^3R_{4j}+(\hat{\theta}-\theta_0)^4R_{5j}$, where

$$\begin{aligned}
R_{1j}&\equiv T^{-1}\textstyle\sum_{t=j+1}^{T}[h_t(h_{t-j,\theta}-\bar{h}_\theta)+h_{t-j}(h_{t\theta}-\bar{h}_\theta)],\\
R_{2j}&\equiv T^{-1}\textstyle\sum_{t=j+1}^{T}(h_t+h_{t-j}),\\
R_{3j}&\equiv T^{-1}\textstyle\sum_{t=j+1}^{T}[h_{t\theta}h_{t-j,\theta}+0{\cdot}5h_t\overset{*}{h}_{t-j,\theta\theta}+0{\cdot}5h_{t-j}\overset{*}{h}_{t\theta\theta}],\\
R_{4j}&\equiv 0{\cdot}5T^{-1}\textstyle\sum_{t=j+1}^{T}[h_{t\theta}\overset{*}{h}_{t-j,\theta\theta}+h_{t-j,\theta}\overset{*}{h}_{t\theta\theta}],\\
R_{5j}&\equiv 0{\cdot}25T^{-1}\textstyle\sum_{t=j+1}^{T}\overset{*}{h}_{t\theta\theta}\overset{*}{h}_{t-j,\theta\theta}.
\end{aligned}\tag{A.6}$$

Then

$$\begin{aligned}
T^{(1/2)-d}(\hat{s}^{(q)}-\tilde{s}^{(q)})&=T^{(1/2)-d}(\hat{\sigma}_0-\tilde{\sigma}_0)+2(\hat{\theta}-\theta_0)[T^{(1/2)-d}\textstyle\sum_{j=1}^{n}j^qR_{1j}]\\
&\quad+2(\hat{\theta}-\theta_0)\bar{h}_\theta[T^{(1/2)-d}\textstyle\sum_{j=1}^{n}j^qR_{2j}]+2T(\hat{\theta}-\theta_0)^2[T^{-(1/2)-d}\textstyle\sum_{j=1}^{n}j^qR_{3j}]\\
&\quad+2T(\hat{\theta}-\theta_0)^3[T^{-(1/2)-d}\textstyle\sum_{j=1}^{n}j^qR_{4j}]+2T(\hat{\theta}-\theta_0)^4[T^{-(1/2)-d}\textstyle\sum_{j=1}^{n}j^qR_{5j}].
\end{aligned}\tag{A.7}$$

We have $T^{(1/2)-d}(\hat{\sigma}_0-\tilde{\sigma}_0)\xrightarrow{p}0$. The proof of Lemma A2 is easily adapted to show that $\sum_{j=1}^{n}j^qR_{1j}$ is $O_p(1)$ (since $\sum_{j=1}^{n}j^qR_{1j}$ is just $\tilde{s}^{(q)}$ with cross-moments of h_t and $h_{t\theta}$ replacing own moments of h_t); absolute summability of the autocovariances of h_t implies that $\sum_{j=1}^{n}j^qR_{2j}$ converges in mean square and thus in probability to zero. Since $T^{(1/2)-d}(\hat{\theta}-\theta_0)\xrightarrow{p}0$ for any $d>0$, the terms involving R_{1j} and R_{2j} are $o_p(1)$. For $i=3,4,5$ Assumption 2b implies that $E|R_{ij}|\leqq 2D\Rightarrow T^{-(1/2)-d}E|\sum_{j=1}^{n}j^qR_{ij}|\leqq T^{-(1/2)-d}O(n^{q+1})=O((n^{q+(1/2)}/T^d)(n/T)^{(1/2)})\to 0$ by (A.5)$\Rightarrow T^{-(1/2)-\delta}\sum_{j=1}^{n}j^qR_{ij}\xrightarrow{p}0$ by Markov's inequality. Since $T(\hat{\theta}-\theta_0)^2=O_p(1)$ by Assumption 2, the summations in (A.7) involving R_{3j}, R_{4j}, and R_{5j} are each $o_p(1)$. ‖

Lemma A4. *If $n\to\infty$ such that (A.5) holds for some d, $\boldsymbol{d}<d<1/2$, $\boldsymbol{d}$ defined in Assumption 3(b), then $T^{(1/2)-d}(\hat{s}^{(q)}-s^{(q)})\xrightarrow{p}0$.*

Proof. Follows from Lemmas A2 and A3. ‖

Lemma A5. *If $s^{(q)}\neq 0$, then under the conditions of Lemma A4, $T^{(1/2)-d}(\hat{\gamma}^{-1}-\gamma^{-1})\xrightarrow{p}0$.*

Proof. Write $\hat{\gamma}^{-1}=g(\hat{s}^{(q)},\hat{s}^{(0)})$. A mean-value expansion of $\hat{\gamma}^{-1}$ around γ yields $\hat{\gamma}^{-1}=\gamma^{-1}+\overset{*}{g}_1(\hat{s}^{(q)}-s^{(q)})+\overset{*}{g}_2(\hat{s}^{(0)}-s^{(0)})$, where $g_1\equiv\partial g/\partial s^{(q)}$, $g_2\equiv\partial g/\partial s^{(0)}$, and the "*" indicates that the derivatives are evaluated at points between $\hat{s}^{(q)}$ and $s^{(q)}$, and between $\hat{s}^{(0)}$ and $s^{(0)}$. Since $s^{(q)}, s^{(0)}\neq 0$, $\partial g/\partial s^{(q)}(s^{(q)},s^{(0)})\equiv g_1<\infty$, $\partial g/\partial s^{(0)}(s^{(q)},s^{(0)})\equiv g_2<\infty$. Then $T^{(1/2)-d}(\hat{\gamma}^{-1}-\gamma^{-1})=\overset{*}{g}_1[T^{(1/2)-d}(\hat{s}^{(q)}-s^{(q)})]+\overset{*}{g}_2[T^{(1/2)-d}(\hat{s}^{(0)}-s^{(0)})]$. By Lemma A3, $\hat{s}^{(q)}\xrightarrow{p}s^{(q)}$, $\hat{s}^{(0)}\xrightarrow{p}s^{(0)}\Rightarrow\overset{*}{g}_1\xrightarrow{p}g_1$, $\overset{*}{g}_2\xrightarrow{p}g_2$. The conclusion then follows from Lemma A4. ‖

Lemma A6. *Under the conditions of Theorem 1, $T^{q/(2q+1)}[\sum_{j=-T+1}^{T-1}(\hat{k}_j-k_j)(\tilde{\sigma}_j-E\tilde{\sigma}_j)]\xrightarrow{p}0$.*

Proof. Choose v and define a so that

$$1+(1/(2b-2))<v<1+(q/2)-\varepsilon;\ a=[T^{v/(2q+1)}].\tag{A.8}$$

(The assumption on b in (3.17) guarantees that this can be done.) As in the proof of a similar proposition in Andrews (1991), we divide the sum that is to be shown to be $o_p(1)$ into two parts, for $j\leqq a$ using the Lipschitz condition in Assumption 1(b), for $j>a$ using the bound on $|k(x)|$ in Assumption 3(b).

Since $\hat{k}_0-k_0=0$, $\hat{k}_{-j}=\hat{k}_j$, and $k_{-j}=k_j$,

$$\begin{aligned}
&T^{q/(2q+1)}\textstyle\sum_{j=-T+1}^{T-1}(\hat{k}_j-k_j)(\tilde{\sigma}_j-E\tilde{\sigma}_j)\\
&=2T^{q/(2q+1)}\textstyle\sum_{j=1}^{a}(\hat{k}_j-k_j)(\tilde{\sigma}_j-E\tilde{\sigma}_j)+2T^{q/(2q+1)}\sum_{j=a+1}^{T-1}\hat{k}_j(\tilde{\sigma}_j-E\tilde{\sigma}_j)\\
&\quad-2T^{q/(2q+1)}\textstyle\sum_{j=a+1}^{T-1}k_j(\tilde{\sigma}_j-E\tilde{\sigma}_j)\\
&\equiv 2A_1+2A_2-2A_3.
\end{aligned}\tag{A.9}$$

Consider A_1 first. We have

$$\begin{aligned}|A_1| &\leqq T^{q/(2q+1)} \textstyle\sum_{j=1}^{a} |\hat{k}_j - k_j||\tilde{\sigma}_j - E\tilde{\sigma}_j| \\ &\leqq c_1 T^{q/(2q+1)} \textstyle\sum_{j=1}^{a} |(j/\hat{\gamma}T^{1/(2q+1)}) - (j/\gamma T^{1/(2q+1)})||\tilde{\sigma}_j - E\tilde{\sigma}_j| \\ &= c_1[T^{(1/2)-2\varepsilon/(2q+1)}|\hat{\gamma}^{-1} - \gamma^{-1}|][T^{(q-1)/(2q+1)}T^{-(1/2)+2\varepsilon/(2q+1)}T^{-1/2}] \textstyle\sum_{j=1}^{a} T^{1/2} j|\tilde{\sigma}_j - E\tilde{\sigma}_j|,\end{aligned}$$

where the second inequality follows from the Lipschitz condition. In the final expression, the first term in brackets is $O_p(1)$ by Lemma A5 because $(q+0{\cdot}5)d < \varepsilon \Leftrightarrow 2\varepsilon/(2q+1) > d$. Since $E|\tilde{\sigma}_j - E\tilde{\sigma}_j| \leqq (E|\tilde{\sigma}_j - E\tilde{\sigma}_j|^2)^{1/2} = \text{var}\,(\tilde{\sigma}_j)^{1/2}$, and $\text{var}\,(T^{1/2}\tilde{\sigma}_j) \leqq c_2$ for some c_2 that depends on neither j nor T, the last term will be $o_p(1)$ by Markov's inequality if $T^{(q-1)/(2q+1)}T^{-1/2+2\varepsilon/(2q+1)}T^{-1/2}\sum_{j=1}^{a} j \to 0$. But this does indeed hold, since $\sum_{j=1}^{a} j = O(a^2) = O(T^{2v/(2q+1)})$, and $(q-1-(2q+1)+2\varepsilon+2v) < 0$ by assumption.

Now,

$$\begin{aligned}|A_2| &\leqq T^{q/(2q+1)} \textstyle\sum_{j=a+1}^{T-1} |\hat{k}_j||\tilde{\sigma}_j - E\tilde{\sigma}_j| \\ &\leqq cT^{q/(2q+1)} \textstyle\sum_{j=a+1}^{T-1} |j/\hat{\gamma}T^{1/(2q+1)}|^{-b}|\tilde{\sigma}_j - E\tilde{\sigma}_j| \\ &= c\hat{\gamma}^b T^{(q+b)/(2q+1)}T^{-1/2} \textstyle\sum_{j=a+1}^{T-1} j^{-b}T^{1/2}|\tilde{\sigma}_j - E\tilde{\sigma}_j|\end{aligned}$$

Since $\hat{\gamma}^b \xrightarrow{p} \gamma^b$, by the logic used in considering A_1, it suffices to show $T^{(q+b)/(2q+1)}T^{-1/2}\sum_{j=a+1}^{T-1} j^{-b} \to 0$. This holds since the summation is $O(T^{(1-b)v/(2q+1)})$ and by assumption $q+b-((2q+1)/2)+(1-b)v < 0$. A similar argument shows $A_3 \xrightarrow{p} 0$. ‖

Lemma A7. *If* $\hat{\gamma} \xrightarrow{p} \gamma \neq 0$, $T^{q/(2q+1)}[\sum_{j=-T+1}^{T-1} (\hat{k}_j - k_j)E\tilde{\sigma}_j] \xrightarrow{p} 0$.

Proof. For $0 < \hat{x}_j < x$ (x defined in Assumption 1c), expand $\hat{k}_j \equiv k(\hat{x}_j)$ around $k(0)$:

$$\hat{k}_j = k(0) + \cdots + k^{([q])}(0)\hat{x}_j^{[q]}/[q]! + \hat{k}_j^{([q]+1)}(\overset{*}{x}_j)\hat{x}_j^{[q]+1}/([q]+1)!,$$

where $k^{(n)}$ is the n'th derivative of k and $\overset{*}{x}_j$ lies between 0 and $\hat{x}_j$. Since $\lim_{|x|\to 0}(1-k(x))/|x|^q < \infty$, $k^{(n)}(0) = 0$ for $n < q$. After a similar expansion of k_j, also around 0, we therefore get

$$\begin{aligned}\hat{k}_j - k_j = (k^{[q]}/[q]!)(\hat{x}_j^{[q]} - x_j^{[q]}) &+ (1/([q]+1)!)\hat{k}_j^{([q]+1)}(\overset{*}{x}_j)x_j^{[q]+1} \\ &- (1/([q]+1)!)k_j^{([q]+1)}(\overset{**}{x}_j)x_j^{[q]+1},\end{aligned}$$

where $\overset{**}{x}_j$ lies between 0 and x_j. This Taylor-series expansion is valid when $\hat{x}_j \equiv j/(\hat{\gamma}T^{1/(2q+1)}) < x$, $x_j \equiv j/(\gamma T^{1/(2q+1)}) < x$, i.e., if

$$j \leqq \bar{j} \equiv \min\{T-1, [x\hat{\gamma}T^{1/(2q+1)}], [x\gamma T^{1/(2q+1)}]\}.$$

Write $T^{q/(2q+1)}[\sum_{j=-T+1}^{T-1} (\hat{k}_j - k_j)E\tilde{\sigma}_j]$ as

$$\begin{aligned}&2T^{q/(2q+1)} \textstyle\sum_{j=1}^{\bar{j}} (\hat{k}_j - k_j)E\tilde{\sigma}_j + 2T^{q/(2q+1)} \sum_{j=\bar{j}+1}^{T-1} \hat{k}_j E\tilde{\sigma}_j - 2T^{(q/2q+1)} \sum_{j=\bar{j}+1}^{T-1} k_j E\tilde{\sigma}_j \\ &\quad \equiv 2B_1 + 2B_2 + 2B_3.\end{aligned}$$

Consider B_1:

$$\begin{aligned}B_1 = {} & T^{q/(2q+1)}(k^{([q])}(0)/[q]!) \textstyle\sum_{j=1}^{\bar{j}} [(j/(\hat{\gamma}T^{1/(2q+1)}))^{[q]} - (j/(\gamma T^{1/(2q+1)}))^{[q]}]E\tilde{\sigma}_j \\ &+ T^{q/(2q+1)}(1/([q]+1)!) \textstyle\sum_{j=1}^{\bar{j}} \hat{k}_j^{([q]+1)}(\overset{*}{x}_j)(j/(\hat{\gamma}T^{1/(2q+1)})^{[q]+1}E\tilde{\sigma}_j \\ &- T^{q/(2q+1)}(1/([q]+1)!) \textstyle\sum_{j=1}^{j} k_j^{([q]+1)}(\overset{**}{x}_j)(j/(\gamma T^{1/(2q+1)})^{[q]+1}E\tilde{\sigma}_j \\ \equiv {} & B_{11} + B_{12} + B_{13}.\end{aligned}$$

If $[q] < q$, then since $\lim_{|x|\to 0}((1-k(x))/(|x|^q)) < \infty$, $k^{[q]}(0) = 0 \Rightarrow B_{11} = 0$. So assume $[q] = q$. Then $B_{11} = (k^{(q)}(0)/[q]!)(\hat{\gamma}^{-q} - \gamma^{-q})\sum_{j=1}^{\bar{j}} j^q E\tilde{\sigma}_j \xrightarrow{p} 0$ since $\hat{\gamma}^{-q} - \gamma^{-q} \xrightarrow{p} 0$, and $|\sum_{j=1}^{\bar{j}} j^q E\tilde{\sigma}_j| \leqq \sum_{j=1}^{\bar{j}} j^q|\sigma_j| \leqq \sum_{j=1}^{\infty} j^q|\sigma_j| < \infty$. Now consider B_{12}. Since the $[q]+1$-derivative of k is bounded on $[0, x]$, $|E\tilde{\sigma}_j| \leqq |\sigma_j|$, and $\bar{j} \leqq [x\gamma T^{1/2q+1}]$, $|B_{12}| \leqq c\hat{\gamma}^{-[q]-1}T^{(q-[q]-1)/(2q+1)}\sum_{j=1}^{[x\hat{\gamma}T^{1/(2q+1)}]} j^{[q]+1}|\sigma_j|$ since $\sum_{j=1}^{\infty} j^{q+0{\cdot}5}|\sigma_j| < \infty$, $|\sigma_j| \leqq c_2 j^{-q-1{\cdot}5} \Rightarrow j^{[q]+1}|\sigma_j| \leqq c_2 j^{[q]-q-0{\cdot}5} \Rightarrow$ the final sum in the above inequality for $|B_{12}|$ is $O_p(\hat{\gamma}^{[q]-q+0{\cdot}5}T^{([q]-q+0{\cdot}5)/(2q+1)})$, from which it follows that $|B_{12}| \xrightarrow{p} 0$. A similar argument shows that $B_{13} \xrightarrow{p} 0 \Rightarrow B_1 \xrightarrow{p} 0$.

Now consider B_2 defined above. We have

$$|B_2| \leqq c_1 \hat{\gamma}^b T^{(q+b)/(2q+1)} \sum_{j=\bar{j}+1}^{\infty} j^{-b} |E\tilde{\sigma}_j|$$
$$\leqq c_2 \hat{\gamma}^b T^{(q+b)/(2q+1)} \sum_{j=\bar{j}+1}^{\infty} j^{-b-q-1\cdot 5}$$
$$\leqq c_2 \hat{\gamma}^b T^{(q+b)/(2q+1)} j^{-b-q-0\cdot 5}|_{\min[T-1, x\hat{\gamma}T^{1/(2q+1)}, x\gamma T^{1/(2q+1)}]} \xrightarrow{p} 0$$

since the sum approaches zero for any of the three lower bounds. That $B_3 \xrightarrow{p} 0$ can be established by exactly the same argument. ‖

Lemma A8. *If* $\hat{\gamma} \xrightarrow{p} \gamma \neq 0$, $T^{q/(2q+1)}[\sum_{j=-T+1}^{T-1} \hat{k}_j(\hat{\sigma}_j - \tilde{\sigma}_j)] \xrightarrow{p} 0$.

Proof. Write $T^{q/(2q+1)} \sum_{j=-T+1}^{T-1} \hat{k}_j(\hat{\sigma}_j - \tilde{\sigma}_j)$ as

$$T^{q/(2q+1)}(\hat{\sigma}_0 - \tilde{\sigma}_0) + [2T^{q/(2q+1)} \sum_{j=1}^{T-1} \hat{k}_j(\hat{\sigma}_j - \tilde{\sigma}_j)] = o_p(1) + [2T^{1/2}(\hat{\theta} - \theta_0)G_1 + 2\bar{h}_\theta T^{1/2}(\hat{\theta} - \theta_0)G_2$$
$$+ 2T(\hat{\theta} - \theta_0)^2 G_3 + 2T(\hat{\theta} - \theta_0)^3 G_4 + 2T(\hat{\theta} - \theta_0)^4 G_5],$$
$$G_i = T^{(-1/2)/(2q+1)} \sum_{j=1}^{T-1} \hat{k}_j R_{ij}, \; i = 1, 2,$$
$$G_i = T^{(-q-1)/(2q+1)} \sum_{j=1}^{T-1} \hat{k}_j R_{ij}, \; i = 3, 4, 5;$$

see A.6. Since $T^{1/2}(\hat{\theta} - \theta_0) = O_p(1)$, it suffices to show $G_i \xrightarrow{p} 0$, $i = 1, \ldots, 5$.

For $i = 1, 2$, $G_i = T^{(-1/2)/(2q+1)} \sum_{j=1}^{T-1} (\hat{k}_j - k_j) R_{ij} + T^{(-1/2)/(2q+1)} \sum_{j=1}^{T-1} k_j R_{ij}$. The proofs of Lemma A6 and A7 are easily adapted to show that the first summation is $o_p(1)$ (since $\sum_{j=1}^{T-1} (\hat{k}_j - k_j) R_{1j}$ is just $\sum_{j=1}^{T-1} (\hat{k}_j - k_j)\tilde{\sigma}_j$ with cross-moments of h_t and h_{t0} replacing own-moments of h_t). A standard proof of the consistency of kernel estimators (Andrews (1991)) shows that the second summation is $O_p(1)$, and thus for $i = 1, 2$ $G_i \xrightarrow{p} 0$. For G_i, $i = 3, 4, 5$: since $\hat{k}_j$ is bounded and $|k(x)| \leqq c_2|x|^{-b}$,

$$|G_i| \leqq c_1 T^{(-q-1)/(2q+1)} \sum_{j=1}^{[T^{1/(2q+1)}]} |R_{ij}| + c_2 \hat{\gamma}^b T^{(-q-1+b)/(2q+1)} \sum_{j=[T^{1/(2q+1)}]+1}^{T-1} j^{-b} |R_{ij}| \xrightarrow{p} 0$$

by Markov's inequality, since, $\hat{\gamma}^b \xrightarrow{p} \gamma^b$ and, by Assumption 2b, $E|R_{ij}| \leqq 2D$. ‖

Proof of Theorem 1. Follows from Lemmas A6, A7 and A8. ‖

Proof of Theorem 3. Since $\hat{k}_0\hat{\sigma}_0 = \hat{\sigma}_0 \xrightarrow{p} \sigma_0$, it suffices to show $2\sum_{j=1}^{T-1} \hat{k}_j \hat{\sigma}_j \xrightarrow{p} 0$. We have

$$\sum_{j=1}^{T-1} \hat{k}_j \hat{\sigma}_j = \sum_{j=1}^{T-1} \hat{k}_j \tilde{\sigma}_j + \sum_{j=1}^{T-1} \hat{k}_j(\hat{\sigma}_j - \tilde{\sigma}_j) \equiv H_1 + H_2.$$

Let

$$b' = \min(b, 2q+1).$$

Then $|k(x)| \leqq c|x|^{-b'}$: for $|x| \leqq 1$, $|k(x)| \leqq c|x|^{-b'}$ for arbitrary $b' \geqq 1$; for $|x| > 1$, $|k(x)| \leqq c_2|x|^{-b} \Rightarrow |k(x)| \leqq c|x|^{-b'}$ for any b' between 0 and b. Let ε be as in Theorem 1. Let η be chosen so that $0 < \eta < 1/2$, $(q/2) - \varepsilon > \eta(2q+1)/2b' + 0\cdot 25\{1 - [(2q+1)/b']\}$; this can always be done since $(q/2) - \varepsilon > 0$ by (3.17) and $0\cdot 25\{1 - [(2q+1)/b']\} \leqq 0$. Then using $|\hat{k}_j| \leqq c|\hat{x}_j|^{-b'} = cj^{-b'}\hat{\gamma}^{b'}T^{b'/(2q+1)}$, we get

$$|H_1| \leqq cT^{b'/(2q+1)} T^{-1/2+\eta} T^{-b'/2 + [2\varepsilon b'/(2q+1)]} (T^{(1/2)-2\varepsilon/(2q+1)}\hat{\gamma})^{b'} \sum_{j=1}^{T-1} j^{-b'} T^{1/2-\eta} |\tilde{\sigma}_j|.$$

The term in $\hat{\gamma}$ is $o_p(1)$ by Lemma A5, since $(q + 0\cdot 5)d < \varepsilon \Rightarrow 2\varepsilon/(2q+1) > d$. The final summation is $o_p(1)$ by Markov's inequality. And by assumption, η is chosen so that

$$0 > b'/(2q+1) - (1/2) + \eta - (b'/2) + [2\varepsilon b'/(2q+1)] \Rightarrow H_1 \xrightarrow{p} 0.$$

Now consider H_2. As in the proof of Lemma A8,

$$\sum_{j=1}^{T-1} \hat{k}_j(\hat{\sigma}_j - \tilde{\sigma}_j) = T^{1/2}(\hat{\theta} - \theta_0)H_{21} + \bar{h}_\theta T^{1/2}(\hat{\theta} - \theta_0)H_{22}$$
$$+ T(\hat{\theta} - \theta_0)^2 H_{23} + T(\hat{\theta} - \theta_0)^3 H_{24} + T(\hat{\theta} - \theta_0)^4 H_{25},$$
$$H_{2i} \equiv T^{-1/2} \sum_{j=1}^{T-1} \hat{k}_j R_{ij} \qquad i = 1, 2,$$
$$H_{2i} \equiv T^{-1} \sum_{j=1}^{T-1} \hat{k}_j R_{ij} \qquad i = 3, 4, 5,$$

with R_{ij} defined in A.6. The proof of Lemma A8 shows that $H_{2i} \xrightarrow{p} 0$, $i = 3, 4, 5$. That $H_{2i} \xrightarrow{p} 0$ for $i = 1, 2$ follows by writing $(R_{ij} = (R_{ij} - ER_{ij}) + ER_{ij}$; the logic used to show $H_1 \xrightarrow{p} 0$ shows that the summations involving $(R_{ij} - ER_{ij})$ are $o_p(1)$; absolute summability of autocovariances maintained in Assumptions 2c and 3a are easily shown to imply that the summations involving ER_{ij} are $o_p(1)$. ‖

Acknowledgements. We thank Dongchul Cho, John Hulbert and Ka-Fu Wong for excellent research assistance, two anonymous referees, Dongchul Cho, Blake LeBaron, Mico Loretan, Ka-Fu Wong and participants in seminars at the Federal Reserve Board of Governors and the National Bureau of Economic Research for helpful comments, and the National Science Foundation, the Sloan Foundation and the University of Wisconsin Graduate School for financial support.

REFERENCES

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

ANDREWS, D. W. K. (1991), "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation", *Econometrica*, **59**, 817–858.

ANDREWS, D. W. K. and MONAHAN, J. C. (1992), "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, **60**, 953–966.

BERK, N. (1974), "Consistent Autoregressive Spectral Estimates", *Annals of Statistics*, **2**, 489–502.

BOLLERSLEV, T. (1986), "Generalized Conditional Heteroskedasticity", *Journal of Econometrics*, **31**, 307–327.

CHRISTIANO, L. J. and DEN HAAN, W. J. (1993), "Small Sample Properties of GMM for Business Cycle Analysis" (manuscript, Northwestern University).

COCHRANE, J. (1988), "How Big is the Random Walk in GNP?", *Journal of Political Economy*, **96**, 893–920.

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

HANSEN, B. E. (1992), "Consistent Covariance Matrix Estimation for Dependent Heterogeneous Processes", *Econometrica*, **60**, 967–972.

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

HANSEN, L. P. and HODRICK, R. J. (1980), "Forward Exchange Rates as Optimal Predictors of Future Spot Rates: An Econometric Analysis", *Journal of Political Economy*, **88**, 829–853.

KEENER, R. W., KMENTA, J. and WEBER, N. C. (1991), "Estimation of the Covariance Matrix of the Least Squares Regression Coefficients When the Disturbance Covariance Matrix is of Unknown Form", *Econometric Theory*, **7**, 22–45.

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

PRIESTLEY, J. B. (1981) *Spectral Analysis and Time Series* (New York: Academic Press).

ROBINSON, P. M. (1991), "Automatic Frequency Domain Inference on Semiparametric and Nonparametric Models", *Econometrica*, **59**, 1329–1363.

SCHWERT, G. W. (1987), "Test for Unit Roots: A Monte Carlo Investigation" (National Bureau of Economic Research Working Paper No. 73).

SILVERMAN, B. W. (1986) *Density Estimation for Statistics and Data Analysis* (London: Chapman and Hall).

SIMS, C. A. (1972), "Money, Income and Causality", *American Economic Review*, **62**, 540–552.

WEST, K. D. and CHO, D. (1994), "The Predictive Ability of Several Models of Exchange Rate Volatility", *Journal of Econometrics* (forthcoming).