TECHNICAL WORKING PAPER SERIES

A PRACTITIONER'S GUIDE TO ROBUST COVARIANCE MATRIX ESTIMATION

Wouter J. den Haan
Andrew T. Levin

Technical Working Paper 197

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 1996

We appreciate comments and suggestions from Jeff Campbell, Larry Christiano, Graham Elliot, Rob Engle, Neil Ericsson, Ron Gallant, Clive Granger, G.S. Maddala, Masao Ogaki, Adrian Pagan, Peter Phillips, Jim Stock, P.A.V.B. Swamy, George Tauchen, and Hal White. This project is supported by NSF grant SBR-9514813. This paper is part of NBER's research program in Economic Fluctuations and Growth. Any opinions expressed are those of the authors and not those of the Board of Governors of the Federal Reserve System, other members of its staff, or the National Bureau of Economic Research.

© 1996 by Wouter J. den Haan and Andrew T. Levin. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

NBER Technical Working Paper 197
June 1996

A PRACTITIONER'S GUIDE TO ROBUST COVARIANCE MATRIX ESTIMATION

ABSTRACT

This chapter analyzes kernel-based and parametric spectral estimation procedures for constructing heteroskedasticity and autocorrelation consistent (HAC) covariance matrices, and provides guidelines for effective implementation of these procedures. To implement a kernel-based procedure, the practitioner must choose a particular kernel, a bandwidth selection method, and a prewhitening filter. To implement a parametric procedure, the practitioner must choose a class of admissible models and a criterion to select a particular model within this class. Simulation experiments indicate that these choices can have important implications for the accuracy of inferences based on the estimated HAC covariance matrix. Thus, rather than viewing any of these procedures as fully "automatic," a combination of diagnostic statistics and common sense should be regarded as essential in practical applications.

Wouter J. den Haan
Department of Economics
University of California
9500 Gilman Drive
La Jolla, CA 92093-0508
and NBER

Andrew T. Levin
International Finance Division
Federal Reserve Board of Governors
Mailstop 24
Washington, DC 20551

1. INTRODUCTION.

In many structural economic or time-series models, the errors may have heterogeneity and temporal dependence of unknown form. Thus, to draw more accurate inferences from estimated parameters, it has become increasingly common to construct test statistics using a heteroskedasticity and autocorrelation consistent (HAC) or "robust" covariance matrix. Since the estimated covariance matrix approaches a constant value as the sample length becomes arbitrarily large, the test statistic typically has a standard normal or chi-squared limiting distribution, which is used in constructing confidence intervals and performing hypothesis tests.

However, to the extent that the estimated HAC covariance matrix exhibits substantial mean-squared error (MSE) in finite samples, the resulting inferences may be severely distorted. For example, substantial variation in the estimated standard error generally causes a t-statistic to take large values (in absolute terms) more frequently than predicted by the limiting standard normal distribution, thereby leading to a tendency to over-reject the null hypothesis in a two-sided test. Other distortions in inference can result when the standard error exhibits bias, skewness, and/or correlation with the estimated model parameter.

The key step in constructing a HAC covariance matrix is to estimate the spectral density matrix at frequency zero of a vector of residual terms. In some empirical problems, the regression residuals are assumed to be generated by a specific parametric model. In a rational expectations model, for example, the Euler equation residuals typically follow a specific moving-average (MA) process of known finite order. For these cases, the practitioner can utilize the spectral estimation procedures of Eichenbaum, Hansen, and Singleton (1988) and West (1994) to obtain a consistent estimate of the covariance matrix. In the more general case where the regression residuals can possess heteroskedasticity and temporal dependence of unknown form, existing results in the spectral density estimation literature (cf. Parzen 1957; Priestley 1982) have contributed to the rapid development of HAC covariance matrix estimation procedures (e.g., White 1984; Gallant 1987; Newey and West 1987, 1994; Gallant and White 1988; Andrews 1991; Robinson 1991; Andrews and Monahan 1992, Den Haan and Levin 1994; and Lee and Phillips 1994).

These HAC covariance matrix estimation procedures may be classified into two broad categories: non-parametric kernel-based procedures, and parametric procedures. Each kernel-based procedure uses a weighted sum of the autocovariances to estimate the spectral density at frequency zero, where the weights are determined by the kernel and the bandwidth parameter. Each parametric procedure estimates a time-series model and then constructs the spectral density at frequency zero that is implied by this model. As shown in Den Haan and Levin (1994), a parametric spectral estimator is consistent under very general conditions similar to those used in the literature to prove consistency of kernel-based estimators. Furthermore, when the sequence of autocovariances satisfy a standard invertibility condition, the parametric VAR estimator converges at a faster rate than any positive semi-definite kernel-based estimator.

To implement a kernel-based procedure, the practitioner must choose a particular kernel, a bandwidth selection method, and a prewhitening filter. To implement a parametric procedure, the practitioner must choose a class of admissible models and a criterion to select a particular model within this class. Simulation experiments indicate that these choices can have important implications for the accuracy of inferences based on the estimated HAC covariance matrix. Thus, rather than viewing any of these procedures as fully "automatic," a combination of diagnostic statistics and common sense should be regarded as essential in practical applications.

Although we focus in this paper on the properties of HAC estimators for conducting inferences, the spectral estimators discussed in this paper are used in many other econometric procedures. For example, the Phillips-Perron unit root test requires a HAC estimator of the spectral density of the first difference. A HAC spectral density estimator is also needed to construct efficient GMM parameter estimates in the case of overidentifying assumptions. For these exercises one only needs the spectral density at frequency zero. The techniques discussed in this paper, however, can easily be used to estimate the spectral density at other frequencies.[1]

[1] See Robinson (1991) for an alternative procedure to estimate the spectrum over a range of frequencies and for econometric problems that require estimates of the spectrum over a range of frequencies.

The remainder of this paper is organized as follows. Section 2 gives step-by-step descriptions of five HAC covariance matrix estimation procedures: the kernel-based procedures proposed by Andrews and Monahan (1992) and Newey and West (1994); the parametric estimators proposed by Den Haan and Levin (1994) and Lee and Phillips (1994); and the non-smoothed non-parametric estimator proposed by Robinson (1995). Section 3 compares the asymptotic properties of kernel-based and parametric estimation procedures. Sections 4 and 5 analyse the choices faced by a researcher in implementing kernel-based procedures and parametric procedures, respectively. Section 6 provides some concluding remarks.

2. HAC COVARIANCE MATRIX ESTIMATORS STEP BY STEP.

In many estimation problems, a parameter estimate $\hat{\psi}_T$ for a $p\times 1$ vector ψ_0 is obtained from the sample analog of a set of moment conditions, such as $\mathrm{E}\, V_t(\psi_0) = 0$, where $V_t(\psi_0)$ is an $N\times 1$ vector of residual terms with $N \geq p$. This orthogonality condition is often used to motivate the following estimator of ψ_0:

$$\hat{\psi}_T = \operatorname{argmin}_{\psi} V_T' F_T V_T, \tag{2.1}$$

where $V_T = \sum_{t=1}^{T} V_t(\psi) / T$ is the vector of sample moments of $V_t(\psi)$ and F_T is an $N\times N$ (possibly) random, symmetric weighting matrix (cf. Hansen 1982). When $N = p$, then the results are invariant to the choice of the weighting matrix F_T. In this case,[2] the parameter $\hat{\psi}_T$, under regularity conditions, has the following limiting distribution:

$$\left[D^{-1} S D'^{-1}\right]^{-1/2} T^{1/2} \left(\hat{\psi}_T - \psi_0\right) \rightarrow \mathrm{N}(0, I_N) \tag{2.2}$$

as the sample size $T \rightarrow \infty$, where S is the spectral density at frequency zero of $V(\psi_0)$, I_N is the $N\times N$ identity matrix, and the $N\times p$ matrix D is defined as follows:

$$D = \mathrm{E}\left[\left.\frac{\partial V_t(\psi)}{\partial \psi'}\right|_{\psi=\psi_0}\right], \tag{2.3}$$

Usually, D is estimated by its sample analog $D_T(\hat{\psi}_T)$ and $D_T(\hat{\psi}_T)\text{-}D \rightarrow 0$ in probability as $T\rightarrow\infty$.

Two different approaches have been followed in the literature to estimate the spectral density of an $N\times 1$ random vector V_t. Non-parametric or kernel-based estimators have the following form:

$$\tilde{S}_T^{np} = \sum_{j=-T+1}^{T-1} \kappa\left(\frac{j}{\xi_T}\right) \tilde{\Gamma}_j, \tag{2.4}$$

where $\kappa(\cdot)$ is a weighting function (kernel) and ξ is a bandwidth parameter. Also,

---

[2] The general formula for the asymptotic covariance matrix is $(D_T' F_T D_T)^{-1} D_T' F_T S_T F_T D_T (D_T' F_T D_T)^{-1}$.

(2.5) $$\tilde{\Gamma}_j = \frac{1}{T}\sum_{t=1}^{T-j} V_t V'_{t+j}, \quad j = 0, \cdots, T\text{-}1,$$

and

$$\tilde{\Gamma}_j = \tilde{\Gamma}'_{-j}, \quad j = -1,-2, \cdots, -T+1.$$

Two widely-used kernels are defined as follows:

Bartlett Kernel: $$\kappa(x) = \begin{cases} 1-|x| & \text{for } |x| \le 1 \\ 0 & \text{otherwise} \end{cases}$$

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

In contrast, parametric estimators use the spectral density implied by a particular time-series model for V_t. For example, suppose that V_t is modeled as a VARMA(p,q) process. Let $\bar{A}_k$ be the matrix of k-th order AR coefficients, and let $\bar{B}_k$ be the matrix of k-th order MA coefficients. Define the $N\times 1$ vector $\bar{e}_t$ as $V_t - \sum_{k=1}^{p}\bar{A}_k - \sum_{k=1}^{q}\bar{B}_k$, and let $\tilde{\Sigma}_T = \sum_{t=k+1}^{T}\bar{e}_t\bar{e}'_t$ be the innovation variance. Then the parametric spectral estimator is given by:

(2.6) $$\tilde{S}_T^{par} = \left[I_N - \sum_{k=1}^{p}\bar{A}_k\right]^{-1}\left[I_N + \sum_{k=1}^{q}\bar{B}_k\right]\tilde{\Sigma}_T\left[I_N + \sum_{k=1}^{q}\bar{B}'_k\right]\left[I_N - \sum_{k=1}^{p}\bar{A}'_k\right]^{-1}.$$

In this section, we give a step-by-step description of five procedures to estimate the spectral density at frequency zero of $V_t(\psi_0)$ using a time series of the estimated residuals $V_t(\hat{\psi}_T)$ of length T. The five procedures are: (1) **QS-PW**, the kernel-based estimator of Andrews and Monahan (1992); (2) **NW-PW**, the kernel-based estimator of Newey and West (1994); (3) **VARHAC**, the parametric VAR estimator of Den Haan and Levin (1994); (4) **PL**, the estimator of Lee and Phillips (1994); and (5) **R95**, the non-parametric estimator of Robinson (1995), R95. Lee and Phillips (1994) consider the case where $V_t(\psi_0)$ is a scalar, while the other four papers consider the vector case. In this section, we describe how these estimators have been used by the authors who proposed

them. With the exception of the R95 estimator, some choices are required to implement each procedure. The implications of these choices will be analyzed in Sections 4 and 5.

2.1 The QS-PW estimator.

The QS-PW estimator from Andrews and Monahan (1992) applies a prewhitening AR filter of order b before the kernel-based estimator from Andrews (1991) is used. When b is set equal to zero in the first step, then the estimator is identical to the Andrews (1991) estimator.

Step 1: Obtain estimates for the "prewhitened" residuals. The following model is estimated with least-squares:

$$V_t(\hat{\psi}_T) = \sum_{k=1}^{b} \hat{A}_k \, V_{t-k}(\hat{\psi}_T) + \hat{e}_t \qquad \text{for} \quad t = b+1, \cdots, T. \tag{2.7}$$

Andrews and Monahan (1992) only consider fixed values for b. In their Monte Carlo experiments, b is set equal to zero or one for each element of $V_t(\hat{\psi}_T)$. If b is set equal to zero, then $\hat{e}_t \equiv V_t(\hat{\psi}_T)$, and the estimator is equal to the estimator from Andrews (1991). Note that we have placed the term "prewhitened" in quotation marks, because no correction for serial correlation would be needed if the residuals were truly prewhitened.

Step 2: Choose a weighting matrix. Under certain regularity conditions, it is possible to derive the bandwidth parameter growth rate that minimizes the asymptotic MSE of the spectral estimator (cf. Priestley 1982; Andrews 1991). The optimal bandwidth parameter sequence for a given kernel depends on an $N^2 \times N^2$ weighting matrix W and on the smoothness properties of the kernel, as indicated by the characteristic exponent, q (cf. the discussion in Section 3.2 below). For the Bartlett kernel, $q = 1$; and for the QS kernel, $q = 2$. For a given kernel with characteristic exponent q, the asymptotically optimal bandwidth parameter sequence is given by:

$$\xi_T^* = c(q)\left[\alpha(q)T\right]^{1/(2q+1)}. \tag{2.8}$$

Here

$$\alpha(q) = \frac{2\,\mathrm{vec}(S^{(q)})' W \,\mathrm{vec}(S^{(q)})}{\mathrm{tr}\left(W(I+K)(S \otimes S)\right)}, \tag{2.9}$$

$$(2.10) \qquad c(q) = \begin{cases} 1.1447 & \text{for } q = 1 \\ 1.3221 & \text{for } q = 2 \end{cases},$$

K is the $N^2 \times N^2$ commutation matrix that transforms vec(B) into vec(B'). $S^{(q)}$ indicates the q-th generalized derivative of the spectral density at frequency zero, which is defined as follows (cf. the discussion in Section 3.2 below):

$$(2.11) \qquad S^{(q)} = \sum_{j=-\infty}^{\infty} |j|^q C_j .$$

Andrews (1991) and Andrews and Monahan (1992) only assign positive weight to the N diagonal elements of S and $S^{(q)}$. Denote the n-th weight by ω_n. In a least-squares estimation problem, Andrews and Monahan (1992) set all weights ω_n corresponding to the slope coefficients equal to unity, and the element corresponding to the regression intercept equal to zero. However, as discussed in Section 4.3.3, these weights make the bandwidth parameter sensitive to the scaling of the variables, which can lead to highly unsatisfactory results in practical applications. A straightforward way to avoid this problem is to set ω_n equal to the inverse of the variance of $V_{nt}(\hat{\psi}_T)$.

Step 3: Calculate the data-dependent bandwidth parameter. Andrews (1991) and Andrews and Monahan (1992) propose that a parametric model be used to provide initial estimates of S and $S^{(q)}$, which are then plugged into equation (2.9). In simulation experiments, these authors estimate univariate AR(1) representations for each of the N elements of $\hat{e}_t$. Denote the resulting parameter estimates by $(\hat{\rho}_n, \hat{\sigma}_n^2)$, $n = 1, \cdots, N$. For this parametric model, estimates of $\alpha(q)$ as constructed as follows:[3]

$$(2.12) \qquad \hat{\alpha}(1) = \frac{\sum_{n=1}^{N} \omega_n \dfrac{4\hat{\rho}_n^2 \hat{\sigma}_n^4}{(1-\hat{\rho})^6 (1+\hat{\rho})^2}}{\sum_{n=1}^{N} \omega_n \dfrac{\hat{\sigma}_n^4}{(1-\hat{\rho})^4}},$$

[3] When a more general weighting matrix was chosen in step 2, then a parametric model that provides estimates for the off-diagonal element of S and $S^{(q)}$ must be used.

$$\hat{\alpha}(2) = \frac{\sum_{n=1}^{N} \omega_n \frac{4\hat{\rho}_n^2 \hat{\sigma}_n^4}{(1-\hat{\rho})^8}}{\sum_{n=1}^{N} \omega_n \frac{\hat{\sigma}_n^4}{(1-\hat{\rho})^4}}. \tag{2.13}$$

Finally, we obtain the following data-dependent bandwidth parameter for the Bartlett kernel:

$$\hat{\xi}_T^* = 1.1447\left[\hat{\alpha}(1)T\right]^{1/3}. \tag{2.14}$$

For the QS kernel, the data-dependent bandwidth parameter is given by:

$$\hat{\xi}_T^* = 1.3221\left[\hat{\alpha}(2)T\right]^{1/5}. \tag{2.15}$$

For any positive semi-definite kernel, the bandwidth parameter must grow arbitarily large with increasing sample size to ensure the consistency of the spectral estimator. Thus, even when the data are known *a priori* to be generated by a finite-order MA(q) process, the kernel estimator may exhibit very poor properties if the bandwidth parameter is simply set equal to q (cf. Ogaki 1992). Furthermore, if the kernel estimator is calculated under the restriction that the autocovariances beyond q are zero, then the modified estimator is not necessarily positive semi-definite. These considerations highlight the advantages of using the parametric procedures proposed by Eichenbaum, Hansen and Singleton (1988) or West (1994) for a MA process of known finite order.

Step 4: Calculate the spectral density of the "prewhitened" residuals.

The spectral density at frequency zero of the "prewhitened" residuals is given by:

$$\begin{aligned} \hat{\Sigma}_T^{QS-PW} &= \sum_{j=1-T}^{T-1} \kappa\left(\frac{j}{\hat{\xi}_T}\right) \hat{\Gamma}_T(j), \quad \text{where} \\ \hat{\Gamma}_T(j) &= \frac{1}{T}\sum_{t=1}^{T-j} \hat{e}_t \hat{e}'_{t+j} \qquad \text{for } j \geq 0 \quad \text{and}, \\ \hat{\Gamma}_T(j) &= \hat{\Gamma}'_T(-j) \qquad \text{for } j < 0. \end{aligned} \tag{2.16}$$

Step 5: Calculate the HAC estimate of the spectral density. The estimate of the spectral density at frequency zero is given by

$$\hat{S}_T^{QS-PW}(\hat{\psi}_T) = \left[I_N - \sum_{k=1}^{b} \hat{A}_k\right]^{-1} \hat{\Sigma}_T^{QS-PW} \left[I_N - \sum_{k=1}^{b} \hat{A}'_k\right]^{-1} \tag{2.17}$$

2.2 The NW-PW estimator.

The NW-PW estimator by Newey and West (1994) is similar to the QS-PW estimator. The main difference lies in the procedure used to obtain initial estimates for S and $S^{(q)}$ in equation (2.9). Whereas Andrews (1991) uses a parametric model to obtain these initial estimates, Newey and West (1994) propose the use of a non-parametric method.

Step 1: *Obtain estimates for the "prewhitened" residuals.* Same as for QS-PW.

Step 2: *Choose a weighting matrix.* Newey and West (1994) assign positive weight to the diagonal and off-diagonal elements of S and $S^{(q)}$. In particular, given an $N\times1$ vector w, the $N^2\times N^2$ weighting matrix W in equation (2.9) is specified as a diagonal matrix, where the i-th diagonal element is equal to the i-th element of vec($w\,w'$). This specification simplifies the formula for $\hat{\alpha}(q)$ in equation (2.9) considerably. In Monte Carlo simulation experiments, Newey and West (1994) set all elements of w corresponding to the slope coefficients equal to one, and the element corresponding to the regression intercept equal to zero. However, this choice of weights makes the bandwidth parameter sensitive to the scaling of the variables. As discussed in Section 4.3.3, a straightforward way to avoid this problem is to set w_n equal to the inverse of the standard deviation of $V_{nt}(\hat{\psi}_T)$.

Step 3: *Calculate the data-dependent bandwidth parameter.* When W is as described in step 2, then equation (2.9) can be expressed as follows:

$$\alpha(q) = \left[\frac{w' S^{(q)} w}{w' S w}\right]^2 \tag{2.18}$$

Newey and West (1994) propose that $\alpha(q)$ be estimated non-parametrically as follows:

$$
\begin{aligned}
\hat{\alpha}(q) &= \left[\frac{w'\hat{S}^{(q)}w}{w'\hat{S}^{(0)}w}\right]^2 \quad q=0,1,2, \quad \text{where} \\
\hat{S}^{(q)} &= \sum_{j=-l}^{l}|j|^q\hat{\Gamma}_j, \\
l &= \beta_1\left(\frac{T}{100}\right)^{2/9} \quad \text{for the Bartlett kernel and,} \\
l &= \beta_2\left(\frac{T}{100}\right)^{2/25} \quad \text{for the QS Kernel,}
\end{aligned}
\tag{2.19}
$$

where $\hat{\Gamma}_j$ is defined as in equation (2.16). Newey and West (1994) consider the values 4 and 12 for β_1, and the values 3 and 4 for β_2. The characteristic exponent of the kernel determines the rate at which the truncation parameter l increases with sample length T. Using the estimate of $\alpha(q)$ given in equation (2.19), the data-dependent bandwidth parameter is determined by equation (2.14) for the Bartlett kernel, and by equation (2.15) for the QS kernel.

Step 4: Calculate the spectral density of the "prewhitened" residuals. Same as for QS-PW (cf. equation 2.16), using the bandwidth parameter given by Step 3. The spectral estimator of the vector of "prewhitened" residuals is denoted by $\hat{\Sigma}_T^{NW-PW}$.

Step 5: Calculate the HAC estimate of the spectral density. Same as for QS-PW (cf. equation 2.17), using the results of Step 4:

$$
\hat{S}_T^{NW-PW}(\hat{\psi}_T) = \left[I_N - \sum_{k=1}^{b}\hat{A}_k\right]^{-1}\hat{\Sigma}_T^{NW-PW}\left[I_N - \sum_{k=1}^{b}\hat{A}'_k\right]^{-1} \tag{2.20}
$$

2.3 The VARHAC estimator.[4]

The VARHAC estimator of Den Haan and Levin (1994) estimates a VAR representation for $V_t(\hat{\psi}_T)$, and then constructs the spectral density at frequency zero implied by this model.

[4] GAUSS, RATS, and Fortran programs to calculate the VARHAC estimator can be found on the web-site: http://weber.ucsd.edu/~wdenhaan.

Step 1: Lag order selection for each VAR equation. For the n^{th} element V_{nt} of the vector $V_t(\hat{\psi}_T)$ ($n = 1, \cdots, N$) and for each lag order $\kappa = 1, \cdots, \overline{K}$, the following model is estimated by ordinary least squares:

$$V_{nt} = \sum_{j=1}^{N} \sum_{k=1}^{\kappa} \hat{\alpha}_{njk}(\kappa)\, V_{j,t-k} + \hat{e}_{nt}(\kappa) \quad \text{for } t = \overline{K}+1, \cdots, T. \tag{2.21}$$

Equation (2.21) represents the regression of each component of V_t on its own lags and the lags of the other components. For lag order 0, we set $\hat{e}_{nt}(\kappa) \equiv V_{nt}$. Next, the model selection criterion is calculated for each lag order $\kappa = 0, \cdots, \overline{K}$. In this case, Akaike's 1973) information criterion is given by:

$$\text{AIC}(\kappa; n) = \log\left(\frac{\sum_{t=\overline{K}+1}^{T} \hat{e}_{nt}^2(\kappa)}{T}\right) + \frac{2\kappa N}{T}. \tag{2.22}$$

Schwarz' (1978) Bayesian information criterion is given by:

$$\text{BIC}(\kappa; n) = \log\left(\frac{\sum_{t=\overline{K}+1}^{T} \hat{e}_{nt}^2(\kappa)}{T}\right) + \log(T)\frac{\kappa N}{T}. \tag{2.23}$$

For each element of $V_t(\hat{\psi}_T)$, the optimal lag order κ_n is chosen as the value of κ that minimizes AIC(κ;n) or BIC(κ;n). Den Haan and Levin (1994) show that setting $\overline{K}$ equal to $T^{1/3}$ leads to a consistent covariance matrix estimator. Note that the only specifications that are considered are the ones in which all elements of V_t enter with the same number of lags in the regression equation for V_{nt}. This constraint can easily be relaxed, but at a substantial computational cost when the dimension N is large.

Step 2: Calculate the spectral density of the prewhitened residuals. Let $\hat{K}_T$ be the largest lag-order chosen by the model selection criterion for the N elements of $V_t(\hat{\psi}_T)$. Using the results of step 1, the restricted VAR can be expressed as:

$$V_t(\hat{\psi}_T) = \sum_{k=1}^{\hat{K}_T} \hat{A}_k^{VAR} V_{t-k}(\hat{\psi}_T) + \hat{e}_t, \tag{2.24}$$

where $\hat{e}_t$ is an $N\times 1$ vector with typical element $\hat{e}_{nt}(\kappa_n)$. The (n,j) element of $\hat{A}_k^{VAR}$ is equal to zero if $k > \kappa_n$ and it is equal to $\hat{\alpha}_{njk}(\kappa_n)$ if $k \le \kappa_n$. The innovation covariance matrix $\hat{\Sigma}_T^{VARHAC}$ is estimated as follows:

$$\hat{\Sigma}_T^{VARHAC} = \frac{\sum_{t=\bar{K}+1}^{T} \hat{e}_t \hat{e}'_t}{T}. \tag{2.25}$$

Alternatively, seemingly unrelated regression (SUR) methods could be used to obtain joint estimates of the restricted VAR parameters and the innovation covariance matrix, which would yield more efficient parameter estimates if the innovation covariance matrix contains significant off-diagonal elements.[5]

Step 3: Calculate the HAC estimate of the spectral density. Using the results of steps 1 and 2, the spectral density matrix at frequency zero is estimated by:

$$\hat{S}_T^{VARHAC}(\hat{\psi}_T) = \left[I_N - \sum_{k=1}^{\hat{K}_T} \hat{A}_k^{VAR} \right]^{-1} \hat{\Sigma}_T^{VARHAC} \left[I_N - \sum_{k=1}^{\hat{K}_T} \hat{A}_k^{VAR} \right]^{-1} \tag{2.26}$$

2.4 The PL estimator.[6]

The estimator of Lee and Phillips (1994) combines elements of the procedures described above. Note that in this section, V_t is assumed to be a scalar process.

Step 1: Lag order selection using an ARMA specification. Lee and Phillips (1994) propose that the Hannan-Rissanen recursion (cf. Hannan and Rissanen 1982) be used to determine the order and estimated coefficients of an ARMA representation of the data. In the first stage, an AR specification for $V_t(\hat{\psi}_T)$ is selected using AIC as the model selection criterion. The estimated residuals from this regression are denoted by $\hat{e}_t$. In the second stage of the algorithm, $V_t(\hat{\psi}_T)$ is regressed on lagged values of $V_t(\hat{\psi}_T)$ and $\hat{e}_t$. That is,

$$V_t(\hat{\psi}_T) = \sum_{k=1}^{p} \tilde{a}_k V_{t-k}(\hat{\psi}_T) + \sum_{k=1}^{q} \tilde{b}_k \hat{e}_{t-k} + \tilde{e}_t. \tag{2.27}$$

Then $\hat{p}$ and $\hat{q}$ are selected as the order estimates that minimize the BIC criterion.

---

[5] Efficiency gains can also be achieved in small samples by reestimating the equations using observations before $\bar{K}$, whenever possible.

[6] A GAUSS subroutine library is available from Predicta Software Inc. (phone: 203-432-3695).

Let the estimates for $\tilde{a}_k$ and $\tilde{b}_k$ using the ARMA($\hat{p},\hat{q}$) specification be denoted by $\hat{a}_k$ and $\hat{b}_k$, respectively. Then the estimated residuals from this model are given by:

$$\hat{\tilde{e}}_t = V_t(\hat{\psi}_T) - \sum_{k=1}^{\hat{p}} \hat{a}_k V_{t-k}(\hat{\psi}_T) - \sum_{k=1}^{\hat{q}} \hat{b}_k \hat{e}_{t-k} . \tag{2.28}$$

Step 2: Calculate the spectral density of the "prewhitened" residuals.
The procedure of Andrews (1991) is used to obtain an estimate for the spectral density at frequency zero of the "prewhitened" residuals $\hat{\tilde{e}}_t$ (as described in Steps 2 to 4 of Section 2.1 above). As in Lee and Phillips (1994), we use $\hat{\Sigma}_T^{PL}$ to denote the spectral estimator at frequency zero of the process $\hat{\tilde{e}}_t$.

Step 3: Calculate the HAC estimate of the spectral density. The spectral density at frequency zero of the process $V_t(\hat{\psi}_T)$ is estimated by:

$$\hat{S}_T^{PL}(\hat{\psi}_T) = \frac{\left[1 + \sum_{k=1}^{\hat{q}} \hat{b}_k\right]^2 \hat{\Sigma}_T^{PL}}{\left[1 - \sum_{k=1}^{\hat{p}} \hat{a}_k\right]^2} . \tag{2.29}$$

2.5 The R95 estimator.

Robinson (1995) has recently proposed a non-parametric estimator of the spectral density of $u_t \otimes x_t$. This non-parametric estimator does not require the use of a kernel. The R95 estimator is given by:

$$\begin{aligned} \hat{S}_T^R &= \sum_{j=1-T}^{T-1} \hat{\Gamma}_T^u(j)\hat{\Gamma}_T^x(j), && \text{where for } z = u, x \\ \hat{\Gamma}_T^z(j) &= \frac{1}{T}\sum_{t=1}^{T-j}(z_t - \bar{z})(z_{t+j} - \bar{z})' && \text{for } j \geq 0, \text{ and} \\ \hat{\Gamma}_T^z(j) &= \hat{\Gamma}_T'^z(-j) && \text{for } j < 0, \quad \text{and} \\ \bar{z} &= \frac{1}{T}\sum_{t=1}^{T} z_t . \end{aligned} \tag{2.30}$$

An interesting feature of this estimator is that no choices are required, making it the simplest HAC estimator discussed in this chapter. However, the R95 estimator has an important disadvantage. Consistency requires that the following condition is satisfied:

(2.31) $$E(u_0 \otimes x_0)\,(u_j' \otimes x_j') = E(u_0 u_j') \otimes E(x_0 x_j').$$

This condition rules out any form of heteroskedasticity. Moreover, as noted by Robinson (1995), both u_t and x_t must be random processes, so the estimator cannot be used for scalar processes. This would occur when V_t contains two elements, one of which is a constant term. In this case, the R95 estimator is identical to the sample periodogram, which is not a consistent estimator of the spectral density (cf. Priestley 1982).

3. ASYMPTOTIC PROPERTIES

In this section, we discuss the asymptotic properties of HAC robust covariance matrix estimation procedures. In particular, we discuss consistency and the rates at which the estimators converge to the population values. For each of the estimation procedures reviewed in Section 2, the specific assumptions and methods of proof used to verify these asymptotic properties can be found in the references cited there. Therefore, in this section we focus on the broader issues concerning the large-sample performance of these estimators. Nevertheless, this section is more technical than the other section in this paper. However, the reader does not have to read this section to be able to follow Sections 4 and 5. In Section 3.1, we give an overview of the issues discussed in this section.

3.1 General Considerations.

The estimated HAC covariance matrix is typically used to construct test statistics based on the limiting distribution of the regression parameters. Given that the true limiting covariance matrix is constant, the test statistic typically has a standard normal or chi-squared limiting distribution. To the extent that the estimated HAC covariance matrix is not constant due to sampling variation, the test statistic will tend to deviate from its limiting distribution and thereby generate distorted inferences.

Based on these considerations, the key asymptotic property to be determined is the rate at which the estimated HAC covariance matrix converges (in mean-squared) to its fixed limiting value. From equation (2.2), it can be seen that this rate depends on the convergence of the differential matrix, D_T , and the estimated spectral density matrix at frequency zero, S_T . The differential matrix D_T (defined in equation 2.3) typically converges at the rate $O_p(T^{-1/2})$, where the notation $O_p(\cdot)$ indicates convergence in probability.[7] However, to obtain a spectral estimator that captures general temporal dependence, it is necessary to increase the bandwidth parameter (for a kernel-based procedure) or the lag order (for a parametric procedure). Thus, the estimated spectral density matrix generally converges more slowly than $O_p(T^{-1/2})$, so that this becomes the rate-limiting step in constructing a HAC covariance matrix. Under certain regularity

[7] As indicated in footnote 2, in estimation problems where $N > p$, the asymptotic covariance matrix also depends on the limiting value of the weighting matrix, F_T . However, this matrix typically converges at rate $O_p(T^{-1/2})$, and may converge at an even faster rate if the weights are non-stochastic.

conditions, the use of estimated residuals rather than observed data has a negligible effect on the asymptotic properties (cf. Newey and West 1987; Andrews 1991; Den Haan and Levin 1994).

In light of these considerations, the asymptotic properties of alternative HAC covariance matrix estimators can be largely understood by analyzing the properties of the corresponding spectral density estimators. The asymptotic mean-squared error (MSE) of the spectral density estimator can be decomposed into a non-stochastic component, henceforth referred to as the asymptotic bias, and a stochastic component, henceforth referred to as the asymptotic variance. Sections 3.2 and 3.3 discuss these components for kernel-based spectral estimators, and Sections 3.4 and 3.5 consider these components for the VAR spectral estimator.

3.2 Asymptotic Bias of Kernel Estimators.

Kernel-based spectral estimators face three sources of bias. First, from equation (2.5), it can be seen that the sample autocovariances used by the kernel estimator divide by T and not by the actual number of observations used, so that each sample autocovariance $\tilde{\Gamma}_T(j)$ is biased by the factor $-j/(T-j)$. However, this source of bias will generally be asymptotically negligible. For example, the truncated, Bartlett, and Parzen kernels only assign non-zero weight to sample autocovariances of order $|j| < \xi_T$, so that the bandwidth parameter ξ_T may also be referred to as the lag truncation point for these kernels. For the truncated kernel, this bias will be $|\xi_T|/(T-|\xi_T|)$. For the Bartlett and Parzen kernels, the weight assigned to autocovariances $|j| < \xi_T$ declines at least linearly as a function of the lag order j, so that the maximum degrees-of-freedom bias is even smaller. Thus, as long as ξ_T grows sufficiently slowly as a function of the sample length T, this source of bias becomes asymptotically negligible. Similar considerations apply to the QS kernel, and to all other kernels that ensure a positive semi-definite spectral density matrix, even when the bandwidth parameter does not serve as a lag truncation point.

Second, kernel-based estimators of the spectral density incur bias due to assigning zero weight to autocovariances of lag orders longer than the sample length T. The true spectral density at frequency zero can be expressed as:

$$f(0) = \sum_{j=-\infty}^{+\infty} \Gamma(j). \tag{3.1}$$

Thus, the bias due to neglected autocovariances is equal to the sum of all autocovariances $\Gamma(j)$, summing over $T \leq |j| \leq +\infty$. This source of bias clearly diminishes with increasing sample length, but it is useful to quantify the rate at which the bias vanishes as $T \to \infty$. In particular, suppose that the absolute value of $\Gamma(j)$ shrinks geometrically at the rate $|j|^{r+\delta}$ for some r > 0 and some $\delta > 1$. Then it is not difficult to show (cf. Davidson 1994, pp. 31-32) that:

$$\sum_{j=-\infty}^{+\infty} |j|^{r} |\Gamma(j)| < +\infty. \tag{3.2}$$

In this case, the bias due to neglected autocovariances vanishes at the rate T^r. It is interesting to note that for even values of the parameter r, the left-hand side of equation (3.2) can be viewed as the r-th derivative of the spectral density at frequency zero. For $r > 0$, Parzen and subsequent authors have referred to this formula as the *generalized* derivative of the spectral density at frequency zero (cf. Priestley 1982, p. 459). Thus, the parameter r can be interpreted as the degree of smoothness of the spectral density at frequency zero; i.e., r indicates the highest order for which the derivative of the spectral density is well-defined. For finite-order ARMA processes, the autocovariances vanish at an exponential rate; in this case, the spectral density is infinitely differentiable at frequency zero, so that an arbitrarily large value of r may be chosen. If $r < 1$, then the spectral density displays a "cusp" (or kink) at frequency zero, and is not differentiable in the generalized sense.

The third and dominant source of bias faced by kernel estimators is incurred by placing weights less than unity on the autocovariances at lags shorter than the sample length. As seen in equation (3.1), the true spectral density at frequency zero assigns a weight of unity to all of the autocovariances. The sample periodogram at frequency zero places a weight of unity on all of the sample autocovariances, but it is easy to see that the variance of this estimator does not converge to zero. For example, the sample autocovariance of order T - 1 is always determined by the first and last observations, regardless of the sample length.

The truncated kernel is the simplest method that yields a consistent estimate of the spectral density. This estimator places weight of unity on all autocovariances up to the lag truncation point ξ_T . Thus, from equation (3.2), it can be seen that the bias of the truncated kernel vanishes at the rate ξ_T^r, a rate that may be very rapid if the spectral density is very smooth at frequency zero. In fact, if the data are generated by a finite-order moving-average (MA) process, this bias disappears once the lag truncation point reaches the MA order. Unfortunately, the truncated kernel does not necessarily yield a positive semi-definite spectral density matrix, which limits its usefulness in HAC covariance matrix estimation. Note that positive semi-definite kernels require that $\xi_T \to \infty$ as $T \to \infty$ to eliminate the bias, even when the data are generated by a finite-order moving average processes.

This source of bias is more severe for estimators in the class of kernels that ensure a positive semi-definite spectral density matrix. Kernels in this class must assign weights less than unity to all sample autocovariances (except the sample variance), and the weights must decline toward zero with increasing lag order j. For example, as seen below equation (2.5), the Bartlett kernel assigns linearly declining weights that reach zero at the lag truncation point ξ_T . The QS kernel assigns weights that decline non-linearly, reaching zero at a lag order of about 120 percent of the bandwidth parameter ξ_T and then oscillate around zero for higher lag orders up to the sample length T. For any particular kernel $\kappa(\cdot)$, this source of bias can be expressed as follows:

$$\text{Bias}_{\kappa}(T,\xi_T) = \sum_{j=-T+1}^{T-1} \left(1 - \kappa(j/\xi_T)\right)\Gamma(j). \tag{3.3}$$

Since $\kappa(z) < 1$ for $z \neq 0$, this formula indicates that the bandwidth parameter (lag truncation point) must increase with sample length to reduce this source of bias. Even if the data is generated by a finite-order MA process, so that true autocovariances are equal to zero beyond some maximum lag length, it is necessary for $\xi_T \to \infty$ to ensure the consistency of estimators in this class of kernels. In this case, the kernel argument j/ξ_T declines toward zero for each fixed value of j.

If the bandwidth parameter increases with the sample length, and the autocovariances vanish at a sufficiently rapid rate (i.e., the spectral density is sufficiently

smooth at frequency zero), we can expect that the asymptotic bias will ultimately be determined by the behavior of the kernel $\kappa(z)$ around $z = 0$ as the sample length grows arbitrarily large. For example, the weights assigned to the low-order autocovariances are closer to unity for the QS kernel than for the Bartlett kernel, so that we may expect this bias to vanish at a faster rate for the QS kernel. This property can be made more precise by considering the characteristic exponent q, which is defined as the largest positive integer such that $(1 - \kappa(z))/|z|^q$ has a finite, non-zero limit as $z \to 0$ (cf. Priestley 1982, p. 459). Thus, the characteristic exponent can be viewed as indicating the smoothness of the kernel $\kappa(z)$ at $z = 0$. It is easy to verify that $q = 1$ for the Bartlett kernel, and that $q = 2$ for the QS kernel. More generally, it can be shown that $q \leq 2$ for <u>every</u> kernel that ensures a positive semi-definite spectral density matrix (Priestley 1982, p. 568). The truncated kernel, which is infinitely differentiable at $z = 0$, obviously violates this condition.

Now we can quantify the asymptotic bias for the class of kernels that ensure a positive semi-definite matrix. For a given kernel $\kappa(\cdot)$ with characteristic exponent q, we can rewrite equation (3.3) as follows:

$$\text{Bias}_{\kappa}(T,\xi_T) = \xi_T^{-q} \sum_{j=-T+1}^{T-1} \left\{ \frac{1-\kappa(j/\xi_T)}{(|j|/\xi_T)^q} \right\} \left[|j|^q \Gamma(j) \right]. \tag{3.4}$$

If we assume that r, the largest generalized derivative of the spectral density at frequency zero, is at least as large as the characteristic exponent q, then the term in curly brackets is bounded and the term in square brackets is absolute summable. Thus, this source of bias vanishes at rate $O(\xi_T^{-q})$ as the bandwidth parameter $\xi_T \to +\infty$. If we also assume that the bandwidth parameter increases sufficiently slowly that $\xi_T^q / T \to 0$, then it can be shown that the bias indicated in equation (3.4) dominates the previous two sources of bias (cf. Priestley 1982, p. 459).

3.3 Asymptotic Variance and MSE of Kernel Estimators.

Since kernel-based spectral estimators are calculated from the sample autocovariances, it is clear that the variance of the kernel-based estimate will depend

on the higher moments and temporal dependence of the true data generating process. To analyze this issue further, let us consider a stochastic process $\{V_t\}$ which has stationary moments up to at least the fourth order. In this case, the fourth-order cumulants $K_4(t,t+j,t+m,t+n)$ measure the extent to which this process displays excess kurtosis relative to the fourth-order moments implied by a normally distributed process, $\tilde{V}_t$, with identical autocovariances (cf. Hannan 1970, p. 23; Priestley 1982, p. 58-59).

$$(3.5)\quad K_4(t,t+j,t+m,t+n) = \mathrm{E}(V_t - \mathrm{E}V_t)\,(V_{t+j} - \mathrm{E}V_{t+j})\,(V_{t+m} - \mathrm{E}V_{t+m})\,(V_{t+n} - \mathrm{E}V_{t+n}) - \mathrm{E}(\tilde{V}_t - \mathrm{E}\tilde{V}_t)\,(\tilde{V}_{t+j} - \mathrm{E}\tilde{V}_{t+j})\,(\tilde{V}_{t+m} - \mathrm{E}\tilde{V}_{t+m})\,(\tilde{V}_{t+n} - \mathrm{E}\tilde{V}_{t+n}).$$

Now suppose that this generalized form of excess kurtosis is not too large, so that the fourth-order cumulants are absolutely summable:

$$(3.6)\qquad \sum_{j=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} \sum_{n=-\infty}^{+\infty} K_4\,(t,t+j,t+m,t+n) \;<\; +\infty\ .$$

Under this condition, Bartlett (1946) obtained results that provided the foundation for all subsequent research on the sampling properties of spectral estimators. First, the variance of each sample autocovariance $\tilde{\Gamma}_T(j)$ vanishes at the rate $1/T$. Second, if we consider two sample autocovariances $\tilde{\Gamma}_T(m)$ and $\tilde{\Gamma}_T(n)$ at different lags $m \neq n$, then the covariance between $\tilde{\Gamma}_T(m)$ and $\tilde{\Gamma}_T(n)$ vanishes at the rate $1/T$; i.e., the sampling variation in sample autocovariances at different lags becomes uncorrelated as the sample length grows large.

These results are immediately applicable to any kernel-based spectral estimator, such as the truncated or Bartlett kernel, that can be expressed as a weighted average of the sample autocovariances for lags $0 \leq j \leq \xi_T$; i.e., any kernel that assigns zero weight to sample autocovariances beyond the lag truncation point ξ_T . In particular, these results indicate that such estimators will have asymptotic variance of $O_p\,(\xi_T\ /T)$ as long as the bandwidth parameter ξ_T grows at a slower rate than the sample length T. This result for the asymptotic variance can also be obtained for spectral estimators based on more general kernels, such as the QS kernel (cf. Priestley 1982, p. 457; Andrews 1988, 1991). Finally, these results can be extended to non-stationary processes, under certain conditions on temporal dependence and the existence of sufficiently high moments

(cf. Andrews 1991).[8]

Using this result and those of the previous section, we can now evaluate the asymptotic MSE for the class of kernels that yield a positive semi-definite spectral density matrix. In particular, by adding the squared asymptotic bias to the asymptotic variance for a given kernel $\kappa(\cdot)$ with characteristic exponent q, the asymptotic MSE can be expressed as follows:

$$\text{MSE}_{\kappa}(T, \xi_T) = O\left(\xi_T^{-2q}\right) + O_p\left(\xi_T / T\right). \tag{3.7}$$

This formula highlights the MSE tradeoff in choosing the bandwidth parameter ξ_T for a given sample of length T. On the one hand, using a higher bandwidth reduces the bias caused by the declining kernel weights. On the other hand, raising the bandwidth places larger weight on the high-order sample autocovariances that are relatively poorly estimated.

We can also use equation (3.7) to evaluate the optimal growth rate of the bandwidth parameter, ξ_T, and the corresponding minimum asymptotic MSE. By differentiating the right-hand-side of equation (3.7) with respect to ξ_T and setting the result to zero, we find that the asymptotic MSE is minimized for a kernel with characteristic exponent q when the bandwidth parameter grows at rate $O(T^{1/(2q+1)})$, and that the minimum asymptotic MSE vanishes at rate $O(T^{-2q/(2q+1)})$. Thus, as seen in equation (2.8), the optimal growth rate of the bandwidth parameter for the QS kernel is $O(T^{1/5})$. Using the optimal sequence of bandwidth parameters, the QS spectral estimator converges in mean-squared at rate $O_p\left(T^{-2/5}\right)$.

As discussed in the previous section, the weighting scheme of the Bartlett kernel imposes a higher degree of bias than the QS. Thus, as seen in equation (2.8), the Bartlett kernel utilizes a higher bandwidth parameter growth rate of $O(\text{T}^{1/3})$, which diminishes the influence of the bias, but at the cost of additional variance. Thus, the spectral estimator based on the Bartlett kernel converges in mean-squared at a somewhat slower rate of $O_p\left(T^{-1/3}\right)$.

---

[8] The consistency of kernel-based estimators has been demonstrated under even weaker conditions by Hansen (1992).

3.4 Asymptotic Bias of the VAR Spectral Estimator.

As seen in equation (2.24), the VARHAC estimator depends on the VAR coefficients and the estimated innovation covariance matrix. Since these can be expressed in terms of the sample autocovariances, the asymptotic properties of the VAR spectral estimator can be analyzed using essentially the same methods discussed in Sections 3.2 and 3.3 for kernel-based spectral estimators. In this discussion, we will consider a scalar process $\{V_t\}_{t=-\infty}^{\infty}$, as shown in Den Haan and Levin (1994), it is relatively straightforward to extend this analysis to multivariate processes.

Before analyzing the properties of AR approximation, it is useful to review the conditions under which the true autocovariance structure of a stochastic process can be represented by an infinite-order AR. These conditions are well-understood for weakly stationary processes: if a time series is linearly non-deterministic, then the process has an MA(∞) representation with white-noise (homoscedastic and orthogonal) innovations; if no linear combination of $\{V_t\}_{t=-\infty}^{\infty}$ has zero variance, then the process also has an AR(∞) representation. In the absence of weak stationarity, the stochastic process itself does not have an MA(∞) or AR(∞) representation with white-noise innovations. Nevertheless, under the same conditions that have been used to analyze kernel-based spectral estimators, Den Haan and Levin (1994) have demonstrated that the limiting population autocovariances have an MA(∞) representation. Furthermore, if no linear combination of the data has zero variance (a condition typically used to verify the temporal dependence conditions utilized for kernel-based estimators), then the limiting autocovariances also have an AR(∞) representation. Thus, to simplify the following discussion, we will focus on the case in which the stochastic process is strictly stationary.

To evaluate the asymptotic bias of the AR spectral estimator, it is useful to define the sequence of Toeplitz matrices G_h, , and the corresponding infinite-dimensional matrix G_∞. The autocovariance $\Gamma(j-i)$ comprises the (i,j)th element of G_h for $i,\ j = 1, \cdots, h$, and the (i,j)th element of G_∞ for $i, j = 1, 2, \cdots$. It is also useful to define the sequence of vectors g_h and the corresponding infinite-dimensional vector g_∞, where $\Gamma(j)$ comprises the j-th element of g_h for $j = 1, \cdots, h$, and the j-th element of g_∞ for $j = 1, 2, \cdots$.

Now if we assume that the spectral density function $f(\omega)$ is positive over $[0,\pi]$, then it can be shown that all eigenvalues of G_∞ are positive, and that all eigenvalues of G_h are positive for all $h \geq 1$ (Grenander and Szegö 1958). Thus, $\det(G_\infty) \neq 0$, and $\det(G_h) \neq 0$ for all $h \geq 1$, thereby ruling out cases in which some linear combination of the elements of $(V_t, \cdots, V_{t-h})$ has zero variance. In this case, the infinite-order Yule-Walker equations $G_\infty A_\infty = g_\infty$ are well-defined (cf. Hannan and Kavalieris 1983, 1986; Hannan and Deistler 1988). Since the inverse of G_∞ is also well-defined, the infinite-dimensional vector A_∞ of AR(∞) coefficients and the innovation variance Σ_∞ can be expressed as follows:

$$A_\infty = G_\infty^{-1} g_\infty \quad \text{and} \quad \Sigma_\infty = \Gamma(0) + g'_\infty A_\infty . \tag{3.8}$$

The spectral density at frequency zero can be expressed as follows:

$$f(0) = \Sigma_\infty \left[1 - \sum_{j=1}^{\infty} A_\infty(j) \right]^{-2} . \tag{3.9}$$

In this case, it can also be shown that the AR(∞) coefficients decline at the same rate as the autocovariances:

$$\text{If } \sum_{j=-\infty}^{+\infty} |j|^r |G(j)| < +\infty, \text{ then } \sum_{j=0}^{\infty} |j|^r |A_\infty(j)| < +\infty . \tag{3.10}$$

Now consider the AR(h) approximation, which is based on the true autocovariances $\Gamma(j)$ for $j = 0, \cdots, h$. Since $\det(G_h) > 0$ for all h, we can express the autoregressive coefficient vector A_h and the innovation variance Σ_h as follows:

$$A_h = G_h^{-1} g_h \quad \text{and} \quad \Sigma_h = \Gamma(0) + g'_h A_h . \tag{3.11}$$

The spectral density at frequency zero corresponding to the AR(h) approximation can be expressed as follows:

$$S_h^{ar} = \Sigma_h \left[1 - \sum_{j=1}^{h} A_h(j) \right]^{-2} . \tag{3.12}$$

Den Haan and Levin (1994) establish the asymptotic bias of the AR(h) spectral estimator as follows:

$$\text{Bias}_{ar}(h) = \left| S_h^{ar} - f(0) \right| = O(h^{-r}). \tag{3.13}$$

Thus, as seen from equation (3.2), the smoothness of the spectral density at frequency zero determines the asymptotic bias of the AR spectral estimator. Thus, unless the data are generated by a finite-order AR process, it will be required that the lag order $h \to \infty$ in order to capture the true autocovariance structure of the data.

From the discussion in Section 3.2, it can be seen that the bias of the AR(h) spectral estimator vanishes at the same rate as the bias of the truncated kernel estimator. As previously noted, however, the truncated kernel does not necessarily yield a positive semi-definite spectral density at frequency zero, whereas the AR(h) spectral estimator is ensured to be positive semi-definite by construction.

To understand this result further, it is useful to note that the AR spectral estimator can be expressed as $S_h^{ar} = \sum_{j=-\infty}^{\infty} \Gamma_h^*(j)$, where $\Gamma_h^*(j)$ are the autocovariances implied by the AR(h) model. From equation (3.11), it can be seen that the AR(h) coefficients are determined by the hth-order Yule-Walker equations, so that $\Gamma_h^*(j) = \Gamma(j)$ for $|j| \leq h$. Thus, the difference between the AR(h) and truncated spectral estimators can be expressed as $D_h^{ar} = \sum_{|j|>h} \Gamma_h^*(j)$. Furthermore, as with any stationary finite-order AR process, the implied higher-order autocovariances $\Gamma_h^*(j)$ decline exponentially toward zero as $j \to \infty$ (cf. Hamilton 1994, p. 266). This implies that D_h^{ar} vanishes at the same rate as the leading term $\Gamma_h^*(h+1) = O(h^{-q-1})$. Thus, by including these implied higher-order autocovariances, the VAR(h) estimator ensures a positive definite spectral density matrix with negligible effects on the asymptotic bias relative to the truncated kernel estimator.

3.5 Asymptotic Variance and MSE of the VAR Spectral Estimator.

To analyze the asymptotic variance of the AR spectral estimator, we define the sequence of sample Toeplitz matrices $\hat{G}_{Th}$, where the $(j\text{-}i)$-th sample autocovariance of $V_i(\hat{\psi}_T)$, $\hat{\Gamma}_T(j-i)$, comprises the (i,j)th element of $\hat{G}_{Th}$ for $i, j = 1, \cdots, h$; and we define the sequence of sample vectors $\hat{g}_{Th}$, where $\hat{\Gamma}_T(j)$ comprises the j-th element of $\hat{g}_{Th}$ for $j = 1, \cdots, h$. Then the estimated AR(h) coefficient vector $\hat{A}_{Th}$ and the estimated innovation variance $\hat{\Sigma}_{Th}$ can be expressed as follows:

$$\hat{A}_{Th} = \hat{G}_{Th}^{-1}\,\hat{g}_{Th} \quad \text{and} \quad \hat{\Sigma}_{Th} = \hat{\Gamma}_T(0) + \hat{g}_{Th}'\,\hat{A}_{Th}. \tag{3.14}$$

The spectral density estimator at frequency zero corresponding to the estimated AR (h) approximation can be expressed as follows:

$$\hat{S}_{Th}^{ar} = \hat{\Sigma}_{Th}\left[1 - \sum_{j=1}^{h} \hat{A}_{Th}(j)\right]^{-2}. \tag{3.15}$$

Now we can evaluate the rate at which $\hat{S}_{Th}^{ar}$ converges to S_h^{ar}. From equations (3.14) and (3.15), it is clear that the AR spectral estimator can be expressed in terms of the sample autocovariances. If the maximum lag order H_T is restricted to grow at rate $O(T^{1/3})$, then Den Haan and Levin (1994) demonstrate that $\hat{G}_{Th}^{-1}$ converges at rate $o(h/T)^{1/2}$ to G_h^{-1}, uniformly in $0 \le h \le H_T$. In this case, the asymptotic variance of $\hat{S}_{Th}^{ar}$ is dominated by the sum of elements of the vector $G_h^{-1}(\hat{g}_{Th} - g_h)$, which can be expressed as a weighted average of the sample covariance deviations $\hat{\Gamma}_T(j) - \Gamma(j)$. Thus, Bartlett's (1946) result (or its generalization to non-stationary processes) can be applied directly to this weighted average. Thus, we find the asymptotic variance of the AR spectral estimator to be $O(h/T)$, uniformly in $0 \le h \le H_T$. In other words, the asymptotic variance of the AR spectral estimator converges at the same rate as the asymptotic variance of kernel-based spectral estimators.

Combining this result with the asymptotic bias given in equation (3.13), we can evaluate the asymptotic MSE of $\hat{S}_{Th}^{ar}$ as follows:

$$\text{MSE}_{ar}(T, h_T) = O(h_T^{-2r}) + O_p(h_T / T). \tag{3.16}$$

uniformly in $0 \le h_T \le H_T = O(T^{1/3})$. This result reveals a MSE tradeoff in the choice of lag order h, similar to the MSE tradeoff in the choice of bandwidth parameter for kernel-based estimators: namely, a higher lag order reduces the asymptotic bias and increases the asymptotic variance. Since the optimal growth rate of the lag order depends on the smoothness of the spectral density at frequency zero, one might suppose that the optimal rate cannot be identified in practice.

In fact, however, we can approach arbitrarily closely to the optimal growth rate by using Schwarz' (1978) Bayesian Information Criterion (BIC) to select the lag order. The BIC penalty term, $h \log(T)/T$, is sufficiently large to dominate the sampling variation of the estimated innovation covariance matrix, so that $\hat{\Sigma}_{Th}$ can be used as a proxy for Σ_h, the covariance matrix implied by the true AR(h) approximation. Furthermore, Σ_h converges at rate $O(h^{-2r})$ to Σ_∞, the innovation covariance matrix implied by the AR(∞) representation. Thus, BIC provides a means of evaluating the tradeoff between asymptotic bias (by measuring the extent to which additional lags improve the goodness-of-fit) and asymptotic variance (by penalizing the use of additional parameters).

If the spectral density is differentiable at frequency zero (i.e., $r \ge 1$), the lag order chosen by BIC converges to $(T / \log(T))^{1/(2r+1)}$, so that the AR spectral estimator converges in probability at a geometric rate arbitrarily close to $T^{-r/(2r+1)}$. If the true autocovariances correspond to those of a finite-order ARMA process (i.e., $r \to +\infty$), then the lag order chosen by BIC grows at a logarithmic rate, and the AR spectral estimator converges in probability at a rate arbitrarily close to $T^{1/2}$. Finally, in the case where the spectral density is not differentiable at frequency zero (i.e., $0 < r < 1$), the lag order chosen by BIC approaches the maximum rate $H(T) = T^{1/3}$, and the AR spectral estimator converges in probability at the rate $T^{r/3}$.

As previously noted, the truncated kernel estimator also has asymptotic bias of $O(h^{-r})$ and asymptotic variance of $O_p(h/T)$. Thus, in principle, the truncated kernel estimator could converge at rate $T^{-r/(2r+1)}$ if the lag truncation point ξ_T could be chosen to grow at the optimal rate. In practice, however, a data-dependent bandwidth selection

procedure has not been developed for the truncated kernel estimator (cf. Priestley 1982, pp. 460-462; White 1984, p. 159; Andrews 1991, p. 834).

Finally, these asymptotic results indicate that the AR spectral estimator converges at a faster rate than any positive semi-definite kernel-based estimator for almost all autocovariance structures. If $q < r$, the positive definite kernel estimators lose efficiency by placing weight less than unity on the low-order autocovariances. The extreme case is one in which the autocovariances have the structure of a finite-order ARMA process, so that r is arbitrarily large. In this case, the AR spectral estimator converges at a rate approaching $O_p(T^{-1/2})$, whereas spectral estimators based on either the Parzen or QS kernel converge at the rate $O_p(T^{-2/5})$, and the spectral estimator based on the Bartlett kernel converges at the rate $O_p(T^{-1/3})$.

For $r < q$, positive definite kernel estimators with $q = 2$ are also less efficient than the AR spectral estimator, because the bandwidth parameter specified by Andrews (1991) grows too slowly. For example, in the case where $r = 1/2$, BIC will asymptotically select the maximum lag order $O(T^{1/3})$, so that the AR spectral estimator converges at rate $O_p(T^{-1/6})$. In contrast, the spectral estimators which are based on either the Parzen or QS kernel, and which utilize Andrews' (1991) bandwidth selection procedure, will converge at rate $O_p(T^{-1/10})$. Thus, the VAR spectral estimator converges at a faster rate than the QS or Parzen kernels except in the special case where r is exactly equal to 2. The AR spectral estimator converges at a faster rate than the Bartlett kernel estimator for $r > 1$. If $r \leq 1$, the bandwidth parameter of the Bartlett kernel and the VAR lag order both increase at rate $O(T^{1/3})$, so that both estimators converge in probability at the same rate $T^{-r/3}$ in this case.

4. CHOICES FOR KERNEL-BASED ESTIMATORS.

To implement a kernel-based procedure, the practitioner must choose a particular kernel and bandwidth parameter, as well as the order of a prewhitening filter, if any. To construct a data-dependent bandwidth parameter, as proposed by Andrews (1991) and Newey and West (1994), the practitioner must choose a weighting matrix and a method of providing initial estimates of the spectral density and its first or second derivative at frequency zero. In this section, we utilize simulation experiments to highlight the implications of these choices for the finite-sample behavior of the data-dependent bandwidth parameter, the estimated HAC covariance matrix, and the resulting accuracy of inferences on linear regression parameters. This analysis also provides some useful guidelines to aid a practitioner in the effective implementation of these procedures.

4.1 Prewhitening.

Andrews and Monahan (1992) considered the benefits of applying an AR(1) prewhitening filter to the vector of residuals before using a kernel-based estimator (cf. Priestley 1982, pp. 556-557). The AR(1) filter has provided improved inference properties in many Monte Carlo simulation experiments, some of which have considered data generating processes resembling actual economic time series (cf. Andrews and Monahan 1992; Newey and West 1994; Christiano and Den Haan 1996; and Burnside and Eichenbaum 1996).

In the absence of a prewhitening filter, kernel-based spectral estimators tend to exhibit substantial bias in cases where the autocovariances decline gradually toward zero. First, kernel-based procedures assign zero (or approximately zero) weight to autocovariances at lags higher than the bandwidth parameter. Second, to ensure a positive semi-definite estimator, kernel-based procedures assign weights less than unity to autocovariances at lags less than the bandwidth parameter. The rate at which these weights decline toward zero also depends on the bandwidth parameter: i.e., the auto-covariance at a given lag receives less weight when the bandwidth parameter is small.

The AR(1) filter estimates the value of an autoregressive root based on the first-order autocovariance. After the filtering of this autoregressive root, the autocovariances of the prewhitened residuals may decline more rapidly toward zero, thereby reducing the

bias of the kernel-based estimator. Thus, AR(1) prewhitening can provide finite-sample benefits even when the true *dgp* is not a low-order VAR process. For example, Andrews and Monahan (1992, Table V) find that the AR(1) filter yields improved inference properties even when the residuals are MA(q) processes.

It should also be noted that the AR(1) prewhitening filter is a special case of parametric estimators which determine the autoregressive order using a data-dependent model selection criterion. Lee and Phillips (1994) consider the use of BIC to choose an ARMA process to prewhiten the data, and then apply a kernel-based estimator to the prewhitened residuals. In the case where the true data generating process is a finite-order ARMA with i.i.d. innovations, Lee and Phillips (1994) have demonstrated that the optimal bandwidth parameter grows very slowly, so that the kernel has negligible asymptotic influence on the spectral estimate. The asymptotic analysis of Den Haan and Levin (1994) indicates that this result holds under much more general conditions: as the sample length increases, the data becomes truly prewhitened by the parametric procedure, so that no additional benefits can be derived from applying a kernel-based procedure to the prewhitened data. In small samples, of course, the parametric procedure does not completely prewhiten the data, so that applying a kernel estimator to the parametric residuals may provide improved inferences under certain conditions. In future research, this possibility should be explored using Monte Carlo simulation experiments.

4.2 Choice of the kernel.

Many different kernels have been considered in the literature. The truncated kernel assigns unit weight to all sample autocovariances up to the bandwidth parameter, also referred to as the lag truncation point (cf. White 1984). Nevertheless, the truncated kernel does not ensure a positive semi-definite covariance matrix, and no method is currently available for determining the optimal lag truncation point. In contrast, to ensure a positive semi-definite spectral estimate, the Bartlett, Parzen, and QS kernels assign weights less than unity to these sample autocovariances, with the weights declining toward zero as the autocovariance lag increases. Within the class of kernels that ensure a positive semi-definite spectral estimate, the QS kernel minimizes the asymptotic MSE (cf. Priestley 1982; Andrews 1991). However, several simulation studies indicate that all

kernels within this class have fairly similar finite-sample properties (cf. Andrews 1991; Newey and West 1994; Christiano and Den Haan 1996).

4.3 Optimal Bandwidth Procedure.

The choice of the bandwidth parameter is crucial for the behavior of a kernel-based estimator. Increasing the bandwidth parameter reduces the bias while increasing the variance of the estimated covariance matrix. The sensitivity of inferences to the value of the bandwidth parameter motivated the derivation of data-dependent bandwidth parameter methods proposed by Andrews (1991) and Newey and West (1994). Although these methods are sometimes referred to as "automatic," the practitioner should be aware of several important issues which arise in obtaining a data-dependent bandwidth parameter. Section 4.3.1 discusses the optimality criterion used in deriving these methods. Section 4.3.2 reviews the calculation of preliminary spectral estimates required to implement these methods. Section 4.3.3 considers the determination of the weighting matrix in multivariate settings, and highlights the restriction that a single bandwidth must be used for all elements to ensure a positive semi-definite HAC covariance matrix.

4.3.1 The Optimality criterion.

Andrews (1991) and Newey and West (1994) used the asymptotic (truncated) MSE as the optimality criterion in obtaining the bandwidth parameter formula given in equation (2.8) above. Thus, for a given kernel, the data-dependent bandwidth parameter formula only expresses the rate at which the bandwidth parameter should grow as a function of the sample size, and cannot indicate the optimal value of the bandwidth parameter for any particular finite sample. More precisely, for any fixed integer M, the bandwidth parameter $\xi_T^{**} = \xi_T^{*} + M$ meets the same asymptotic optimality criterion as the bandwidth parameter ξ_T^{*} defined in equation (2.8). Unfortunately, while ξ_T^{*} and ξ_T^{**} may yield dramatically different results in a particular finite sample, there is no *a priori* basis upon which to choose one bandwidth parameter over the other.

This non-uniqueness property may appear similar to other uses of asymptotic optimality criteria in the literature. For example, if the OLS estimator $\hat{\beta}_T$ is consistent,

then $\hat{\beta}_T + M/T$ is also consistent for any fixed value of M. The essential difference is that the OLS estimator also satisfies a sensible finite-sample estimation criterion (namely, minimizing the sum of squared residuals of the regression model), whereas current bandwidth selection procedures do not satisfy any particular finite-sample criterion.

Although the data-dependent bandwidth parameter formula given in equation (2.8) does not have a specific finite-sample justification, several simulation studies indicate that this formula performs reasonably well in finite samples, *if* reasonably good initial spectral density estimates can be plugged into this formula. The question of how to obtain such initial estimates will be discussed in Sections 4.3.2 and 4.3.3.

4.3.2 Implementing the Optimal Bandwidth Procedure.

The data-dependent bandwidth parameter formula given in equation (2.8) depends on S and $S^{(q)}$, the spectral density and its q-th generalized derivative at frequency zero. Thus, preliminary estimates $\hat{S}_T$ and $\hat{S}_T^{(q)}$ are required to obtain an estimate of the data-dependent bandwidth parameter $\hat{\xi}_T^*$, which is then used to obtain the final kernel-based spectral estimator. As indicated in Section 2.1 above, Andrews (1991) and Andrews and Monahan (1992) obtain these preliminary estimates of S and $S^{(q)}$ using a parametric approach, namely, fitting a univariate AR(1) model to each element of the residual vector $V(\hat{\psi}_T)$. As indicated in Section 2.2 above, Newey and West (1994) obtain these initial estimates using a non-parametric approach, based on truncated sums of the sample autocovariances.[9]

The key difference between these two methods is that the procedure of Andrews (1991) and Andrews and Monahan (1992) only considers the first-order autocorrelation of each element of the residual vector, whereas the procedure of Newey and West (1994) considers several autocovariances and cross-covariances. The following Monte Carlo experiment illustrates the extent to which this distinction can be important in practice.

Consider the problem of estimating the mean of the following scalar process:

---

[9] That is, they calculate these statistics using the truncated kernel. The estimated bandwidth will always be positive, since these statistics are squared in the formula for the optimal bandwidth.

(4.1) $$Y_t = \varepsilon_t + \nu\,\varepsilon_{t-1} + \mu\,\varepsilon_{t-q},\ q \in \{2,3\}, \quad \text{and} \quad \hat{\psi}_T = \frac{\sum_{t=1}^{T} Y_t}{T}$$

where ε_t is an i.i.d. normally distributed random variable with zero mean and unit variance. The parameters are chosen in such a way that the first-order autocorrelation coefficient of the prewhitened series is small or equal to zero, but higher-order autocorrelation coefficients are substantially larger.

Several empirical cases suggest that such a time series process for Y_t is not unrealistic. First, Fama and French (1988) documented that for stock returns, autocorrelations are small for short horizons, but relatively large for large horizons. For instance, the average first-order autocorrelation across industries is equal to -0.03 for one-year returns, but equal to -0.34 for four-year returns. Second, Christiano and Den Haan (1994) used a *dgp* resembling that of US quarterly GNP, and found that some prewhitened residuals had a very low first-order MA coefficient, but substantial higher-order serial correlation. This example will be discussed further in Section 4.4.

Table 1: The ability of QS-PW and QS-NW to detect serial correlation patterns.

				QS-PW		Average		NW-PW		Average
q	ν	μ	99%	95%	90%	$\hat{\xi}_T$	95%	95%	90%	$\hat{\xi}_T$
2	0.0	-0.3	100.0	99.6	98.3	0.81	97.5	93.1	87.9	8.83
2	-0.1	-0.3	100.0	99.8	99.1	0.92	97.2	92.7	88.2	10.10
2	0.0	0.3	95.1	87.4	80.3	0.95	97.4	91.2	84.8	4.39
2	0.1	0.3	95.9	88.6	81.7	1.02	97.5	91.3	85.2	4.44
3	0.0	-0.3	100.0	99.3	98.0	0.62	97.0	92.1	87.7	11.52
3	-0.1	-0.3	100.0	99.6	98.7	0.66	96.8	91.8	87.3	12.97
3	0.0	0.3	95.5	87.5	80.9	0.62	96.9	90.7	84.7	5.09
3	0.1	0.3	95.7	88.1	81.5	0.64	96.9	90.8	84.8	4.83

Note: This table reports the coverage probabilities of the t-statistic that tests whether the mean of y_t is equal to its true value. The following *dgp* is used to generate the data: $Y_t = \varepsilon_t + \nu\,\varepsilon_{t-1} + \mu\,\varepsilon_{t-q}$, $q = 2,3$, where ε_t is an i.i.d standard normal random variable. $\hat{\xi}_T$ indicates the estimated bandwidth parameter. $T = 128$ and the results are based on 10,000 replications. The results for VARHAC are given in table 5.

Table 1 reports the average bandwidth parameter obtained by these methods, and the resulting confidence interval for a t-statistic to test whether the true mean is equal to zero. It can be seen that the Newey-West procedure is better able to detect the higher-order serial correlation, chooses a higher bandwidth parameter, and consequently has better inference properties. Of course, the Andrews (1991) method might yield superior properties in an example where the autocovariances decline gradually and monotonically. In practice, of course, the properties of the true autocovariances are unknown, so that it is probably unwise to rely on an arbitrary time-series model to determine the bandwidth parameter used to obtain an estimated HAC covariance matrix. In particular, it seems doubtful that the data-dependent bandwidth parameter should depend exclusively on the first-order autocorrelations of the prewhitened residuals, when the residual vector has already been prewhitened by an AR(1) filter.

An alternative to these methods would be to use a formal procedure to select the best parametric model for $V(\hat{\psi}_T)$, and then to use the estimates of S_T and $S_T^{(q)}$ implied by this model. In this case, however, one might consider simply using the parametric estimator of S_T in constructing the HAC covariance matrix, rather than trying to determine the data-dependent bandwidth parameter and then using a kernel-based procedure. This issue will be discussed further in Section 5.

Finally, this simulation experiment highlights the danger of viewing any particular data-dependent bandwidth selection procedure as being fully "automatic". As documented in Table 1, the average bandwidth parameter chosen by QS-PW is less than one. When such a low bandwidth parameter is obtained for a sample of 128 observations, it would be useful to check whether the resulting inferences are sensitive to an increase in the bandwidth parameter. Even with a sample of this length, it should be possible to estimate more than one autocovariance with reasonable accuracy.

4.3.3 The Choice of W and the Costs of Imposing a Single Bandwidth Parameter.

As documented in equation (2.9), the optimality criterion used to derive the optimal bandwidth parameter formula depends on a weighting matrix W. The weighting matrix is very important for the following reason. To ensure that the estimated covariance matrix is positive semi-definite, a single bandwidth parameter must be chosen for the

entire vector $V(\hat{\psi}_T)$. Thus, the data-dependent bandwidth parameter must compromise in evaluating the serial correlation properties of the various elements of $V(\hat{\psi}_T)$. In particular, assigning more weight to specific elements of $V(\hat{\psi}_T)$ influences the estimated bandwidth parameter $\hat{\xi}_T$.

Unfortunately, Andrews (1991), Andrews and Monahan (1992), and Newey and West (1994) do not provide much guidance in choosing the weighting matrix W. In simulation experiments, Andrews (1991) and Andrews and Monahan (1992) choose W such that a unit weight is given to the N-1 diagonal elements of S and $S^{(q)}$ that correspond to the N-1 slope coefficients. All other elements of W are set equal to zero. The simulation experiments of Newey and West (1994) assign unit weights to all diagonal and off-diagonal elements of S and $S^{(q)}$ that do not correspond to the intercept in the regression model. In both cases, these weighting schemes work reasonably well, because the elements of $V(\hat{\psi}_T)$ have reasonably similar variance and autocorrelation properties.

In practice, however, using fixed equal weights can have very undesirable consequences. Since the optimal bandwidth formula is designed to minimize the asymptotic MSE, the elements of $V(\hat{\psi}_T)$ with the highest variance have the most influence in determining the data-dependent bandwidth parameter. Thus, if a particular regressor is rescaled, its sample variance will change, and the autocorrelation properties of that variable will receive a different weight in determining the bandwidth parameter. We illustrate this point with the following Monte Carlo experiment. Consider the ordinary least-squares estimator for the following linear model:

$$
\begin{aligned}
y_t &= \alpha + \beta z_t + \varepsilon_t, \\
(1 - 0.9L)\,\varepsilon_t &= e_{1,t} \\
x_t &= e_{2,t} \\
z_t &= \lambda x_t,
\end{aligned}
\tag{4.2}
$$

where $\alpha = \beta = 0$, $e_{1,t}$ and $e_{2,t}$ are i.i.d. normally distributed random variables. The parameter λ scales the explanatory variable. The unconditional variance of ε_t and x_t is equal to 1. The two elements of the vector V_t are ε_t and $\lambda\varepsilon_t x_t$. Thus the first element is a first-order AR process, and the second element is serially uncorrelated. Varying the

scale coefficient λ is equivalent to expressing the explanatory variable in different measurement units.

To highlight the fundamental point, we do not use the prewhitening option, since first-order prewhitening would make both components close to white noise. For higher-order processes of ε_t, the kernel-based estimators would encounter the same limitations as those discussed here. However, the discussion would be complicated by the misspecification bias of the AR(1) coefficient in the prewhitening regression. Also, because this issue does not depend on the procedure to estimate $\alpha(q)$, we only report the results for the QS estimator of Andrews (1991).

Theory suggests that the choice of a smaller bandwidth parameter in this experiment should improve the finite-sample behavior of the standard error for the slope coefficient, while a larger bandwidth parameter will tend to improve the accuracy of inferences concerning the regression intercept. As indicated in Table 2, the results for the QS estimator are highly sensitive to the value of λ. For the QS procedure, choosing a larger value of λ raises the weight on the second element of V_t, reduces the average bandwidth parameter chosen, and diminishes the accuracy of the estimated standard error of the regression intercept. The average bandwidth parameter across Monte Carlo replications is equal to 23.4, 2.3, and 1.7 for values of λ equal to 1, 100, and 1000, respectively. As expected, a larger value of λ reduces the bandwidth parameter and improves the behavior of the estimated standard error for the slope coefficient.

Table 2: The Limitations of a Single Bandwidth Parameter (QS kernel).

a: unit weight assigned to both diagonal elements.

	α			β			Average
λ	99%	95%	90%	99%	95%	90%	$\hat{\xi}_T$
1	87.8	78.7	72.0	92.7	84.6	77.0	23.35
100	62.4	51.1	43.7	98.7	94.2	88.4	2.32
1000	56.6	45.2	38.4	98.7	94.4	88.9	1.70

b: unit weight assigned only to the diagonal element of the corresponding parameter.

α			β			Average $\hat{\xi}_T$	
99%	95%	90%	99%	95%	90%	α	β
88.3	79.0	71.5	98.6	94.1	88.6	23.26	1.70

Note: These tables report the 99%, 95%, and 90% confidence intervals for the t-statistics that test whether the least-squares estimates for the constant α and the slope β are equal to its true value. The *dgp* is given in equation 4.2. The parameter λ is a scaling variable. A higher value for λ means that the variance of the independent variable increases. $\hat{\xi}_T$ indicates the estimated bandwidth parameter. $T = 128$ and the results are based on 10,000 replications. The results for the VARHAC estimator are reported in table 7.

From this example, it is clear that a minimal requirement for the choice of W is that it should make the optimal bandwidth parameter scale-independent. However, it is not clear how to do this. Den Haan and Levin (1994) consider the use of the inverse of the unconditional covariance matrix and the inverse of the spectral density at frequency zero for the choice of W.[10] It becomes somewhat more difficult to evaluate the optimal bandwidth formula in equations (2.8) and (2.9) if a general weighting matrix is specified instead of a vector of weights. More importantly, while this approach resolves the scaling problem, it cannot resolve the limitation that a single bandwidth parameter must be chosen for the entire vector $V(\hat{\psi}_T)$ to ensure a positive semi-definite HAC covariance matrix.

Now suppose that the practitioner wishes to make inferences concerning a single parameter in a linear regression problem. In this case, the weighting matrix W can be constructed with unit weight assigned to the appropriate element of $V(\hat{\psi}_T)$, and zero weight assigned to all other elements. The results of this approach are reported in panel b of Table 2. As documented in the table, this procedure improves the results drastically.

However, it is clear that the approach of assigning positive weight to only one element of W cannot always resolve the limitation of using a single bandwidth parameter. For example, when standard errors are calculated for non-linear problems, the standard error of each parameter typically depends on the entire spectral density matrix, including both diagonal and off-diagonal elements. Similar considerations apply when restrictions involving several parameters are tested in a linear regression framework. Finally, when the estimated spectral density matrix is used to construct an optimal weighting matrix to

[10] To implement the second suggestion, a preliminary estimate has to be constructed for the spectral density at frequency zero.

obtain efficient GMM estimates, all elements of the spectral density matrix are used, so that zero weight should not be assigned to any particular element. In general, therefore, when the elements of $V(\hat{\psi}_T)$ have different serial correlation properties, the resulting data-dependent bandwidth parameter and HAC covariance matrix will inevitably reflect a somewhat unpleasant compromise.

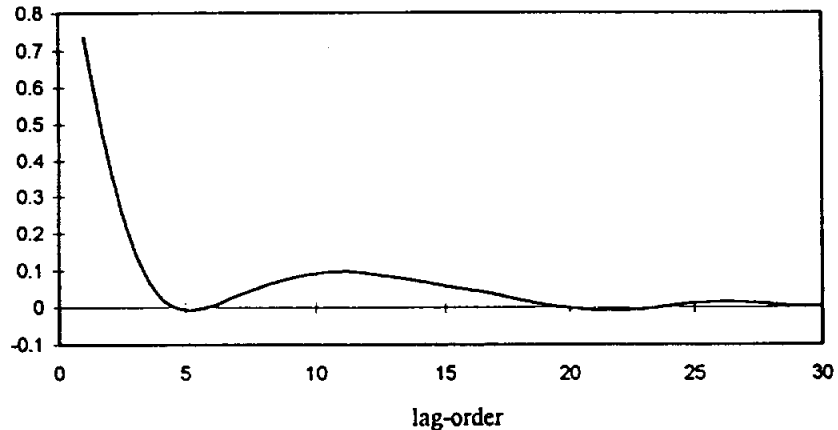
4.4 Complicated serial correlation patterns and kernel-based estimators.

To illustrate several of the topics discussed in this section, we summarize the results of a Monte Carlo experiment performed by Christiano and Den Haan (1996). In this experiment, we consider the following *dgp*:

$$
\begin{aligned}
z_t &= 0.4\, z_{t-1} + e_t, \\
y_t &= y_{t-1} + z_t, \qquad \text{and} \\
y_t^{hp} &= HP(L)\, y_t,
\end{aligned}
\tag{4.3}
$$

where e_t is an i.i.d. normally distributed variable with zero mean. $HP(L)$ stands for the Hodrick-Prescott filter, which is an approximate high-pass filter that removes spectral components with cycles greater than 32 periods. Thus, the HP filter is commonly applied to quarterly macroeconomic data to study the properties of business cycles.[11]

Figure 1: Autocorrelation coefficients of $\left(y_t^{hp}\right)^2$.



Note: This graph plots the autocorrelation function of $(y_t^{hp})^2$. The law of motion for y_t^{hp} is given in equation (4.3.)

We analyze the confidence intervals of the t-statistic that tests whether the standard deviation of $HP(L)y_t$ is equal to its population value. Thus,

[11] See King and Rebelo (1993) and Christiano and Den Haan (1996) for a detailed discussion on the HP filter.

$V_t(\psi_0) = (y_t^{hp})^2 - (\psi_0)^2$. As seen in Figure 1, the serial correlation properties of $V_t(\psi_0)$ are quite complicated.

Table 3: Inference in the presence of complicated serial correlation(QS-PW and NW-PW).

a: $T = 128$.

bandwidth procedure	kernel	prewhitening order	5%	10%	90%	95%	average $\hat{\xi}_T$
Andrews	QS	0	18.6	22.9	15.0	9.4	10.0
Andrews	Bartlett	0	19.0	23.2	16.4	9.8	10.7
NW	Bartlett	0	20.9	24.4	18.0	11.7	5.0
Andrews	QS	1	12.0	17.3	5.0	1.3	2.96
Andrews	Bartlett	1	12.1	17.5	5.5	1.6	3.26
NW	Bartlett	1	16.7	20.5	9.3	5.1	13.05
Andrews	QS	2	18.8	22.2	15.7	9.7	0.95
Andrews	Bartlett	2	18.8	22.3	15.5	9.6	0.71
NW	Bartlett	2	18.8	22.1	15.5	9.6	3.18

b: $T = 1000$.

bandwidth procedure	kernel	prewhitening order	5%	10%	90%	95%	average $\hat{\xi}_T$
Andrews	QS	0	8.6	15.4	12.6	6.8	17.30
Andrews	Bartlett	0	9.0	15.5	13.0	7.1	24.26
NW	Bartlett	0	11.1	17.7	15.2	9.0	11.79
Andrews	QS	1	5.0	9.3	6.8	3.1	4.67
Andrews	Bartlett	1	5.7	10.7	7.7	3.9	6.91
NW	Bartlett	1	7.5	13.6	11.1	5.3	40.70
Andrews	QS	2	10.8	17.3	15.1	9.1	0.98
Andrews	Bartlett	2	10.8	17.4	15.0	9.2	0.74
NW	Bartlett	2	10.7	17.3	15.0	9.1	5.86

Note: These tables report the coverage probabilities of the t-statistic that tests whether the standard deviation of y_t^{hp} is equal to its true value. The 5% (95%) and 10% (90%) columns report the frequency the t-statistic is less (higher) than the lower (upper) 5% and 10% critical value. The *dgp* for y_t^{hp} is given by equation (4.3). $\hat{\xi}_T$ indicates the estimated bandwidth parameter. The results are based on 1,000 replications. The corresponding results for VARHAC are reported in table 6.

The methods of Andrews (1991) and Newey and West (1994) are used to determine the data-dependent bandwidth parameter for the Bartlett and QS kernels, with the use of an autoregressive prewhitening filter of order 0, 1, or 2. Table 3 summarizes

the results. From Table 3, we can make the following observations. First, as mentioned above, the results for the QS and Bartlett kernel are very similar. Second, the distribution of the t-statistic is highly skewed. In fact, analyzing two-sided confidence intervals can give a misleading picture of the deviation of the t-statistic from its limiting distribution. For example, when first-order prewhitening is used with the data-dependent bandwidth method of Andrews (1991), the two-sided t test has an empirical size of 13.3% when the nominal size is 10%. However, this empirical size consists of 12.0% in the left tail and 1.3% in the right tail. Christiano and Den Haan (1996) document that this skewness is caused by the correlation between the estimated standard deviation and the spectral estimate. This reveals one weakness of using MSE as the underlying optimality criterion. The practitioner who calculates a HAC covariance matrix is typically interested in drawing accurate inferences about regression parameters rather than in the covariance matrix itself.

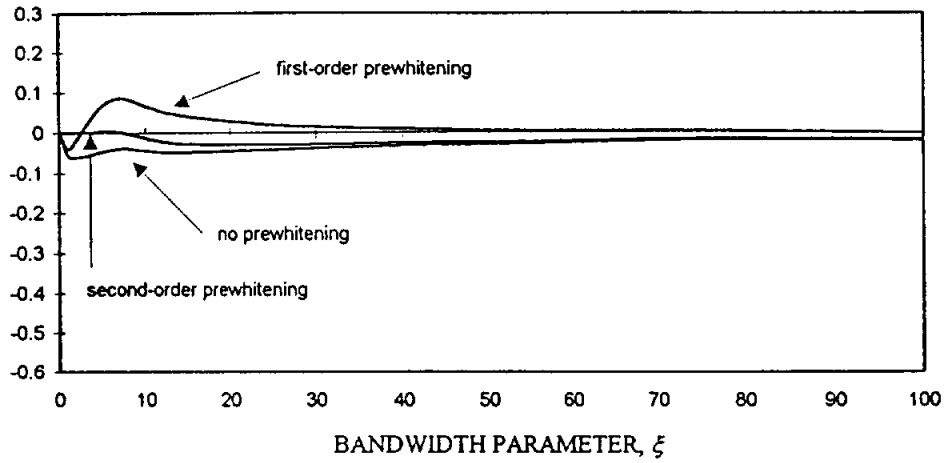
Table 3 also contains some rather surprising results, which provide some useful insight into the characteristics of kernel-based methods. First, compared with the absence of prewhitening, the inference accuracy for two-sided tests improves dramatically with the use of an AR(1) filter. Given the complicated pattern of serial correlation, one would expect second-order prewhitening to yield further improvements in performance, or at least to provide about the same performance as first-order prewhitening. In fact, however, inferences associated with the AR(2) filter are much less accurate than those associated with the AR(1) filter, and are only slightly better than no prewhitening at all.

Second, the AR(1) filter yields a larger improvement in inference accuracy when using Andrews' bandwidth selection method compared with the Newey-West method. This result is surprising because the AR(1) prewhitened residuals have relatively low first-order autocorrelation but continue to have complicated higher-order autocorrelation. As discussed in Section 4.3.2, we would expect the Newey-West method to detect the higher-order serial correlation more effectively than Andrews' method, which only considers the first-order autocorrelation.

Some insight into these findings can be obtained by constructing each kernel-based estimator using the true autocovariances of $HP(L)y_t$. At any given value of the bandwidth parameter ξ, Figure 2a confirms that the Bartlett and QS kernels yield very similar

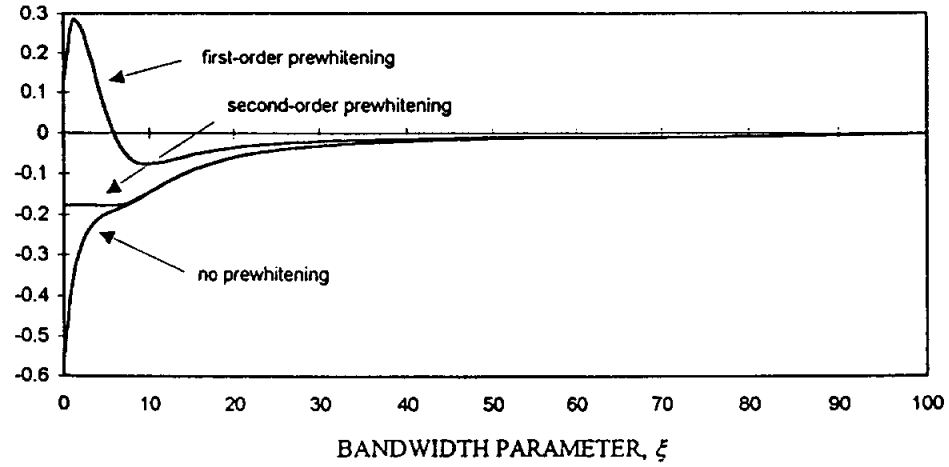
approximations to the true spectral density at frequency zero: i.e., the deviation between $S_B(\xi)$ and $S_{QS}(\xi)$ is always less than 10 percent of the value of S.

Figure 2a: Comparison of Bartlett and QS spectral estimators.



Note: This graph plots the difference between the Bartlett spectral estimator with bandwidth parameter ξ and the QS spectral estimator with bandwidth parameter ξ as a fraction of the true spectral density. The Bartlett and QS spectral estimator are calculated using the true autocovariances of the GMM residual $(y_t^{hp})^2$.

Figure 2b: Relative Bias of QS Spectral Estimator.



Note: The relative bias of the QS spectral estimator is defined as $(S_{QS}(\xi) - S) / S$. $S_{QS}(\xi)$ is the approximate spectral density (using the true autocovariances) of the GMM residual $(y_t^{hp})^2$ based on the indicated prewhitening filter, the QS kernel, and the bandwidth parameter ξ. As $\xi \to \infty$, $S(\xi) \to S$, and the relative bias shrinks to zero.

Figure 2b indicates that the prewhitening order and the choice of bandwidth parameter dramatically influence the relative bias, $(S_{QS}(\xi) - S) / S$. In the absence of prewhitening, the QS kernel generally underestimates the true spectral density, and a fairly

large bandwidth parameter (higher than 20) is required to achieve relative bias of less than 10 percent. Thus, the severe size distortions in the first three rows of Table 3a can be partly explained by the use of a relatively low bandwidth parameter, with an average value of about 10 for Andrews' (1991) method and only 5 for the Newey-West (1994) method. As seen in Table 3b, increasing the sample length from T=128 to T=1000 causes a doubling of the average bandwidth parameter, thereby reducing the bias of the kernel estimator and improving the accuracy of subsequent inferences. Similar results may be observed when second-order prewhitening is performed.

In contrast, first-order prewhitening induces a very different pattern of bias. When the bandwidth parameter is less than about 5, the QS kernel approximation (based on the true autocovariances) *over-estimates* the true spectral density by up to 35 percent. For larger bandwidth parameters, the relative bias is always less than 10 percent in absolute value. The middle three rows of Table 3a indicate that Andrews' (1991) bandwidth selection procedure yields an average bandwidth parameter of about 3, whereas the Newey-West (1994) method yields a much higher average bandwidth parameter of about 13. Thus, one would expect the Newey-West estimator to yield more accurate inferences than the Andrews estimator, but in fact, the opposite is true. Christiano and Den Haan (1996) have shown that the sample autocovariances of $HP(L)y_t$ exhibit substantial downward bias for T=128, which coincidentally offsets the upward bias induced by a low bandwidth parameter (as chosen by Andrews' method), and exacerbates the downward bias induced by a high bandwidth parameter (as chosen by the Newey-West method). This result is clearly rather specific to this particular *dgp*, but is useful for illustrating the factors that can affect the finite-sample performance of alternative spectral estimators.

4.5 Non-parametric estimation without a kernel.

We conclude this section on non-parametric procedures by discussing the R95 estimator proposed by Robinson (1995). Recall that the R95 estimator calculates the spectral density of a vector V_t that can be written as $u_t \otimes x_t$. To analyze the small sample properties of this estimator in conducting inferences, we estimate the covariance of u_t and x_t when the data are generated by the following *dgp*:

$$
\begin{aligned}
z_t^x &= 0.4\, z_{t-1}^x + e_t^x, \\
x_t &= B^i(L)\, z_t^x \ , \quad i = \{HP, \Delta\} \\
z_t^u &= 0.4\, z_{t-1}^u + e_t^u, \\
u_t &= B^i(L)\, z_t^u \left|z_t^x\right|^j \quad j \in \{0,1\},\ i = \{HP, \Delta\} \\
B^{HP} &= \frac{HP(L)}{1-L}, \\
B^{\Delta} &= 1,
\end{aligned}
\tag{4.4}
$$

where e_t^x and e_t^u are i.i.d N(0,1) random variables. Note that when j is equal to one, the distribution of u_t is heteroskedastic (condition (2.31) is not satisfied) and when j is equal to zero, then u_t is homoskedastic (condition (2.31) is satisfied). When the B^{Δ} filter is used, the product V_t displays a fairly simple pattern of serial correlation pattern, whereas the B^{HP} filter generates relatively complicated serial correlation.

Table 4 reports the confidence intervals obtained for a test of the null hypothesis of no covariance between u_t and x_t. The R95 estimator is compared with the QS estimator of Andrews (1991) without prewhitening. For this example, the results for the estimator without prewhitening turned out to be somewhat better than the results with first-order or second-order prewhitening.

First, consider the case with no heteroskedasticity. Both estimators provide reasonably accurate inferences when the degree of serial correlation is relatively limited (for the B^{Δ} filtered data) or when the sample is relatively large ($T = 1000$). However, when the sample is relatively small ($T = 128$) and the data display the complicated serial correlation pattern induced by the B^{HP} filter, the R95 estimator clearly outperforms the kernel-based estimator. For example, the R95 estimator yields a 10.8% empirical size for a two-sided test with a 10% nominal size, compared with the 23.2% empirical size of the kernel-based estimator.

In contrast, when the data exhibit heteroskedasticity, the R95 estimator yields much less accurate inferences, whereas the inference accuracy of the kernel-based estimator is not affected very much. When the B^{HP} filter is used, the ability of the R95 to capture complicated patterns of serial correlation is offset by its inability to adjust for heteroskedasticity. When the B^{Δ} filter is used, the accuracy of inferences is dominated by

the effects of the heteroskedasticity. Unfortunately, in the presence of heteroskedasticity, the inference accuracy of the R95 estimator does not seem to improve in larger samples: the confidence intervals for the R95 estimator are as distorted for T = 1,000 as for T= 128.

Table 4: Non-parametric estimation without a kernel.

a: Without Heteroskedasticity.

T	serial correlation	estimation procedure	kernel	prewhitening order	5%	10%	90%	95%	average $\hat{\xi}_T$
128	B^{Δ}	Andrews	QS	0	6.1	11.8	11.7	5.9	2.4
128	B^{Δ}	Robinson	-	-	5.1	9.9	10.4	5.4	-
129	B^{HP}	Andrews	QS	0	10.1	15.9	19.5	13.1	10.4
128	B^{HP}	Robinson	-	-	4.6	10.1	12.9	6.2	-
1000	B^{Δ}	Andrews	QS	0	5.4	10.0	10.2	4.7	3.8
1000	B^{Δ}	Robinson	-	-	5.1	9.3	9.6	4.3	
1000	B^{HP}	Andrews	QS	0	5.9	10.9	15.6	9.0	17.3
1000	B^{HP}	Robinson	-	-	3.7	8.2	11.7	6.0	

b: With Heteroskedasticity.

T	serial correlation	estimation procedure	kernel	prewhitening order	5%	10%	90%	95%	average $\hat{\xi}_T$
128	B^{Δ}	Andrews	QS	0	6.3	12.0	11.3	5.9	2.4
128	B^{Δ}	Robinson	-	-	18.9	25.1	23.8	17.5	-
128	B^{HP}	Andrews	QS	0	10.6	16.8	19.2	12.6	11.1
128	B^{HP}	Robinson	-	-	10.2	16.4	18.5	11.6	-
1000	B^{Δ}	Andrews	QS	0	5.4	10.2	10.5	5.5	3.5
1000	B^{Δ}	Robinson	-	-	18.0	23.4	23.9	18.5	
1000	B^{HP}	Andrews	QS	0	6.0	10.9	13.9	8.3	18.9
1000	B^{HP}	Robinson	-	-	10.2	16.0	18.3	11.9	

Note: This tables reports the coverage probabilities of the t-statistic that tests whether the covariance of u_t and x_t is equal to its true value of zero. The 5% (95%) and 10% (90%) columns report the frequency the t-statistic is less (higher) than the lower (upper) 5% and 10% critical value. The *dgp* for y_t^{Δ} is given by equation 4.4. $\hat{\xi}_T$ indicates the estimated bandwidth parameter. The results are based on 3,000 replications

5. CHOICES FOR PARAMETRIC ESTIMATORS.

In this section, we analyze the choices required to implement a parametric spectral estimator. Section 5.1 considers the choice of a class of parametric models. Section 5.2 evaluates the properties of alternative model selection criteria. Section 5.3 documents the advantages of being able to select a different lag-order for each element of V_t. Finally, Section 5.4 considers the potential benefits and pitfalls of applying a kernel-based spectral estimator to the residuals of a parametric model that has been chosen by a model selection criterion, as proposed by Lee and Phillips (1994).

5.1 The Class of Admissible Models.

In some empirical problems, the regression residuals are assumed to be generated by a specific parametric model. In a rational expectations model, for example, the Euler equation residuals typically follow a specific moving-average (MA) process of known finite order. For these cases, the practitioner can utilize the procedures of Eichenbaum, Hansen, and Singleton (1988) and West (1994). These procedures yield consistent covariance matrix estimates when the regression residuals are generated by an MA(q) process for which the finite order q is known *a priori*. Furthermore, West's (1994) estimator converges at the rate $T^{-1/2}$, and in contrast to the truncated kernel estimator, is guaranteed to be positive semi-definite.

In general, however, the *dgp* of the regression residuals is not known *a priori*. In this case, the practitioner must use some criterion to select a particular model from a prespecified class of parametric models. Ideally, one would like to search within the class of finite-order ARMA models, as Lee and Phillips (1994) consider in estimating the spectral density of a scalar process. In the multivariate context, however, vector ARMA estimation and model selection is typically highly computationally intensive and often subject to convergence failure or other numerical problems.

In contrast, VAR estimation and model selection can usually be implemented fairly easily at low computational cost. Den Haan and Levin (1994) have shown that VAR approximation yields a consistent covariance matrix estimate under very general conditions. For example, the regression residuals do not have to follow a finite-order vector ARMA process, or even be covariance stationary. Furthermore, as discussed in

Sections 3.4 and 3.5 above, the VAR spectral estimator converges at a faster rate than any positive semi-definite kernel-based estimator. In particular, if the residual vector does follow a finite-order MA or ARMA process, the VAR spectral estimator converges at a geometric rate arbitrarily close to $T^{-1/2}$. Thus, restricting consideration to the class of VAR models rather than the more general class of vector ARMA models has an asymptotically negligible cost in MSE.

Even when consideration is limited to the class of VAR processes, the number of admissible models can still be very large. In estimating each VAR equation, one can allow a different lag order for each variable. However, this approach requires the estimation of $(\overline{K}+1)^N$ alternative formulations of the equation, which is only computationally feasible if the dimension N and the maximum lag order $\overline{K}$ are fairly small. For each equation, these computational requirements can be reduced by imposing the same lag order for all variables, or by imposing a single lag order for all variables except the lagged dependent variable. As shown in the next subsection, allowing the lag order to vary across equations can yield substantial benefits in finite samples. In relatively high-dimensional systems, however, one may wish to restrict attention to the class of VAR models in which a single lag order is used for the entire system.

5.2 Model Selection Criteria.

As outlined in Judge et al. (1985, pp. 240-247), a number of different model selection criteria can be expressed in the following form:

$$\Omega_{K,T} = \Omega\left(\Sigma_{T,K}, K\right), \tag{5.1}$$

where $\Sigma_{T,K}$ is the estimated innovation variance of the model with K free parameters.

For example, Aikaike's (1973) Information Criterion (AIC) sets $\Omega_{K,T} = \log(\Sigma_{T,K}) + 2K/T$. If the true *dgp* is an AR(p_o) process for some finite p_o, then asymptotically AIC will select a lag order $p_o \leq p \leq p_o + c$ with probability 1 for some positive constant c. Shibata (1976) has demonstrated that AIC is not a consistent model selection criterion, but overestimates the true lag order with positive probability, even as the sample length grows arbitrarily large. However, the probability of choosing an order $p > p_o$ decreases rapidly with p (cf. Shibata 1976; Lütkepohl 1985). Furthermore, for lags greater than p_o,

the AR coefficients converge in probability to zero as the sample length grows large. Thus, the inconsistency of AIC uses up a finite number of extra degrees of freedom, but does not affect the consistency or convergence rate of the AR spectral estimator. Finally, if the true *dgp* is an AR(∞) process with i.i.d. Gaussian innovations, then Shibata (1980, 1981) finds that AIC selects the asymptotically optimal growth rate for the AR lag order.

Schwarz' (1989) Bayesian Information Criterion (BIC) sets $\Omega_{K,T} = \log(\Sigma_{T,K}) + 2K\log(T)/T$. Thus, BIC assigns a higher penalty than AIC for additional parameters, so that the lag order chosen by BIC is always less than or equal to that chosen by AIC. BIC has been shown to be a consistent model selection criterion when the true *dgp* is a finite-order AR or finite-order ARMA process. Furthermore, in simulation experiments comparing a variety of model selection criteria, Lütkepohl (1985) reports that BIC achieves the best performance in choosing the correct AR order and minimizing the mean-squared forecasting error. As discussed in Section 3.5, there is also some asymptotic justification for using BIC rather than AIC in AR spectral estimation, especially for *dgp*s with unknown heteroskedasticity and temporal dependence. Nevertheless, simulation experiments performed by Den Haan and Levin (1994) indicate that parametric HAC covariance matrix estimates based on either AIC or BIC yield relatively similar inference properties for a wide variety of *dgp*s.

More generally, the optimality criterion in equation (5.1) is designed to capture the tradeoff between parsimony and goodness-of-fit in finite samples. Nevertheless, this criterion focuses on minimizing the innovation variance, which is not the only sample statistic which is relevant for spectral estimation. A parametric spectral estimator also requires an accurate estimate of the sum of AR coefficients (and an estimate of the sum of MA coefficients for ARMA spectral estimation). Thus, a model selection criterion which efficiently chooses the correct order or minimizes the innovation variance does not necessarily yield the best spectral estimate.

5.2.1 AR approximation of a finite-order MA processes.

Several of these issues can be illustrated using the experimental design considered in Section 4.3.2, in which we estimate the mean of the following scalar process:

$$(5.2) \qquad Y_t = \varepsilon_t + \sum_{q=1}^{Q} \nu_q \varepsilon_{t-q} \quad \text{and} \quad \hat{\psi}_T = \frac{\sum_{t=1}^{T} Y_t}{T},$$

where ε_t is an i.i.d. normally distributed random variable with zero mean and unit variance. This experimental design sheds light on the extent to which an AR model can be used to capture a finite-order MA process, and provides a useful comparison of AIC and BIC. Table 5 reports the average lag order chosen by AIC and BIC for this *dgp*, and the implied confidence intervals for the test statistic of the null hypothesis of a zero mean, using the parametric variance estimator constructed using each model selection criterion. In particular, AR[AIC] refers to the VARHAC estimator constructed using the lag order chosen by AIC, while AR[BIC] refers to the estimator constructed using BIC.

Table 5: VARHAC inferences for finite-order MA processes.

			AR [AIC]			Average	AR [BIC]			Average
q	ν	μ	99%	95%	90%	$\hat{K}_T$	99%	95%	90%	$\hat{K}_T$
2	0.0	-0.3	98.7	94.9	90.3	2.55	99.4	96.9	93.8	1.25
2	-0.1	-0.3	99.0	95.7	91.4	2.70	99.6	97.7	95.1	1.39
2	0.0	0.3	97.9	92.9	87.8	2.52	97.4	91.9	86.5	1.12
2	0.1	0.3	97.7	92.8	87.8	2.60	96.9	91.7	86.2	1.20
3	0.0	-0.3	99.1	96.0	91.9	2.90	99.5	97.9	95.2	1.10
3	-0.1	-0.3	99.2	96.7	93.3	3.05	99.6	98.4	96.4	1.25
3	0.0	0.3	97.8	92.8	88.4	2.82	96.5	90.1	84.2	0.91
3	0.1	0.3	97.8	92.8	88.1	2.95	96.0	89.0	83.3	1.00

Note: This table reports the coverage probabilities of the t-statistic that tests whether the mean of y_t is equal to its true value. The following *dgp* is used to generate the data: $Y_t = \varepsilon_t + \nu\varepsilon_{t-1} + \mu\varepsilon_{t-q}$, $q = 2,3$, where ε_t is an i.i.d standard normal random variable. The sample length $T = 128$, and the results are based on 10,000 replications. The maximum AR lag order is equal to 5. AR[AIC] refers to the VARHAC estimator constructed using the lag order chosen by AIC, while AR[BIC] refers to the estimator constructed using BIC. $\hat{K}_T$ indicates the chosen lag order. The corresponding results for QS-PW and NW-PW are reported in table 1.

Since the true autocovariances vanish beyond lag 2 or 3, one might expect that

a kernel-based spectral estimator would outperform the parametric AR estimator, regardless of the choice of model selection criterion. In fact, a comparison of Tables 1 and 5 demonstrates that both the AR[AIC] and AR[BIC] estimators yield more accurate confidence intervals than the kernel-based estimator of Andrews (1991). Compared with the kernel-based procedure of Newey and West (1994), the inference accuracy associated with the AR[AIC] estimator is clearly superior, while the accuracy of the AR[BIC] estimator is roughly similar. In effect, this experiment reveals the cost of ensuring a positive semi-definite kernel estimate, as discussed in Sections 3.2 and 3.4: by assigning weights substantially less than unity to the second-order and third-order autocovariances, the kernel-based estimators exhibit substantially more bias than the AR[AIC] estimator.

Nevertheless, the choice of model selection criterion has a substantial impact on the behavior of the AR spectral estimator in this experiment. As seen in Table 5, the average AR order chosen by AIC is generally one or two lags higher than the average AR order chosen by BIC. For example, consider the case when $\nu = 0$, $\mu = -0.3$, and $q = 2$. In this case, AIC chooses an AR order less than two in about 6 percent of the simulations, whereas BIC chooses a zero lag order in about 40 percent of the simulations, and almost never chooses a lag order equal to one. This experiment reveals the finite-sample consequences of achieving consistent lag order selection: due to its relatively high penalty term, BIC often selects an AR lag order which is too low to achieve a satisfactory approximation of a low-order MA process.

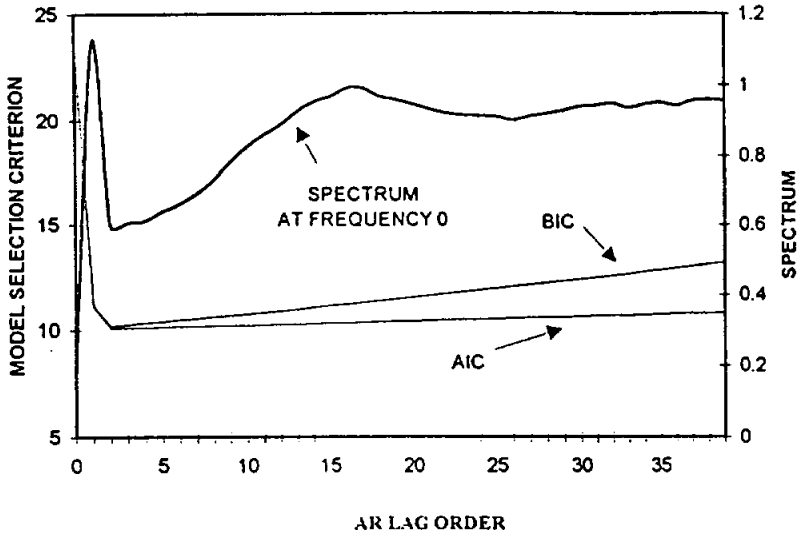
5.2.2 AR approximation of a process with complicated serial correlation.

The experimental design considered in Section 4.4 can also be used to compare the properties of AIC and BIC. Recall from Figure 1 that the autocorrelations of the process $(y_t^{hp})^2$ decline slowly and non-monotonically, so that this *dgp* is useful for analyzing the extent to which an AR approximation provides a reasonable spectral density estimate for a process with general temporal dependence.

The bold curve in Figure 3 depicts the AR(h) approximation (for $h = 1,...,40$)

of the spectral density at frequency zero of $(y_t^{hp})^2$, where the population moments are used to calculate the AR(h) coefficients and the innovation variance.[12] Figure 3 also plots the values of AIC and BIC using the innovation variance implied by the AR(h) approximation. The penalty terms are based on a sample length of 1000 observations.

Figure 3: Model Selection Criteria and Autoregressive Spectra.



Note: This figure indicates the population value of the AR spectral density approximation using different lag-orders. It also depicts the population values of AIC and BIC using the innovation variance implied by an AR(h), again using population moments. The penalty terms are based on a sample length of 1000 observations. The population moments are computed using estimated values from a sample of 100,000 observations. The underlying series is the square of the random variable specified in equation 4.3.

As documented in Figure 3, neither AIC nor BIC is an optimal finite-sample lag order selection criterion in estimating the spectral density at frequency zero. Both model selection criteria reach a minimum when the AR lag order is equal to two. Such a low lag order, however, leads to a strong downward bias for the spectral density estimate. For this particular *dgp*, $1 - \sum_{k=1}^{\infty} A_k$ is relatively small, and the square of this term shows up in the denominator of the definition of the spectral estimate. Therefore, small changes in the sum of the AR coefficients have a large influence on the spectral density estimate.

Table 6 reports the results of using the AR spectral estimator to provide inferences about the standard deviation of y_t^{hp}. Comparison with Table 3 indicates that the confidence intervals implied by the AR spectral estimator are quite similar to those implied

[12] The population moments are computed using estimated values from a sample of 100,000 observations.

by the kernel-based estimators: i.e., the distribution of the test statistic is skewed, and the tails are too thick relative to the limiting distribution. Thus, even for sample lengths up to $T = 1000$, neither parametric nor kernel-based spectral estimators appear to be very successful in capturing the complicated pattern of temporal dependence.

Table 6: Inference in the presence of complicated serial correlation (VARHAC).

a: $T = 128$.

model selection	maximum lag order	5%	10%	90%	95%	Average $\hat{K}_T$
AIC	4	16.2	20.6	15.0	8.7	2.35
AIC	8	16.10	20.9	14.7	8.6	2.78
BIC	4	16.5	20.3	14.9	8.8	1.92
BIC	8	16.6	20.9	15.5	9.2	1.93

b: $T = 1000$.

model selection	maximum lag order	5%	10%	90%	95%	Average $\hat{K}_T$
AIC	10	10.0	14.5	12.5	7.2	4.48
AIC	20	9.9	14.4	12.0	6.7	5.58
BIC	10	11.0	16.2	14.8	8.7	2.05
BIC	20	11.0	16.5	15.3	8.9	2.05

Note: These tables report the coverage probabilities of the t-statistic that tests whether the standard deviation of y_t^{hp} is equal to its true value. The 5% (95%) and 10% (90%) columns report the frequency the t-statistic is less (higher) than the lower (upper) 5% and 10% critical value. The *dgp* for y_t^{hp} is given by equation 4.3. $\hat{K}_T$ indicates the chosen lag order. The results are based on 1,000 replications The corresponding results for QS-PW and NW-PW are given in Table 3.

Finally, it should be noted that the results in Table 6 are not sensitive to the choice of the maximum lag order. This is important, since no criterion is currently available to select the maximum lag order in a finite sample: the asymptotic theory simply prescribes a maximum rate at which it can grow as a function of the sample length. Adding a constant to the maximum lag order does not change any of the asymptotic properties and, at least in this simulation experiment, has little influence on the empirical distribution of the t-statistic.

5.3. The Advantages of Different Lag Orders.

Section 4 highlighted the finite-sample limitations of imposing a single bandwidth for the entire vector of residuals V_t. Recall that the only reason for this restriction is to ensure that the estimated covariance matrix is positive semi-definite. In contrast, the spectral density matrix of a parametric estimator is positive semi-definite by construction. Thus, parametric estimators do not have to compromise in evaluating the serial correlation properties of the elements of V_t, but a model selection criterion can be used to determine the appropriate lag order for each individual element of V_t. That is, if the model selection criterion detects high-order autocorrelation in an element of V_t, then a high lag order will be chosen for that particular element.

We illustrate this advantage of parametric spectral estimators using the experimental design presented in Section 4.3.3. The implied confidence intervals are shown in Table 6. Compared with the kernel-based results reported in Table 2a, it can be seen that the VARHAC procedure yields much more accurate confidence intervals, especially for the slope coefficient. Table 6b indicates that both AIC and BIC almost never choose a zero lag order for the equation corresponding to the regression intercept, where the dependent variable is highly persistent. In contrast, AIC and BIC choose a zero lag order in about 88 and 50 percent of replications, respectively, for the equation corresponding to the slope coefficient, where the dependent variable is white noise.

Table 7: The Benefits of Using Different Lag Orders (VARHAC).

a: confidence intervals.

	BIC			AIC		
parameter	99%	95%	90%	99%	95%	90%
intercept	93.4	86.5	80.5	93.4	86.3	80.4
slope	98.8	95.0	90.1	98.5	94.5	89.3

b: frequency autoregressive lag orders chosen (percentages).

element of V_t	BIC					AIC				
corresponding to	0	1	2	3	4	0	1	2	3	4
intercept	0	98.50	1.42	0.07	0.01	0	78.44	12.75	5.29	3.52
slope	88.41	10.24	1.14	0.19	0.02	49.56	25.17	12.05	6.96	6.26

Note: Panel a reports the 99%, 95%, and 90% confidence intervals constructed using the VARHAC estimator for the t-statistics that test whether the least-squares estimates for the intercept α and the slope coefficient β are equal to their true values. The *dgp* is given in equation 4.2. The sample length $T = 128$, and the results are based on 10,000 replications. $\hat{K}_T$ indicates the chosen lag order. The maximum lag order is equal to 4. Panel b reports the lag orders chosen by the indicated model selection criterion. The corresponding results for kernel-based estimators re reported in Table 2.

5.4 Applying a Kernel-Based Spectral Estimator to Prewhitened Residuals.

In this section, we outline some important issues which have been stimulated by the work of Lee and Phillips (1994), and which deserve to be examined in greater detail in subsequent research. It is useful to consider the potential benefits and pitfalls of applying a kernel-based estimator to the residuals of a parametric model, as outlined for the PL procedure discussed in Section 2.4. If the parametric lag order is high enough to remove all serial correlation, then a non-parametric correction for serial correlation of the prewhitened residuals must simply increase the variance of the final spectral estimate. However, if the residuals display negligible serial correlation, then the data-dependent bandwidth selection procedure of Andrews (1991) may be expected to yield a relatively low bandwidth parameter, so that the kernel-based spectral estimator is nearly identical to the estimated innovation variance of the parametric model. In this case, applying a kernel-based procedure to the prewhitened residuals would tend to have negligible influence on the MSE of the final spectral estimate. In contrast, when the parametric model is not very effective in removing serial correlation, applying a kernel-based estimator to the prewhitened residuals may yield substantial benefits.

Thus, the class of admissible models and the criterion used to select a particular model are likely to be important in determining the benefits of applying a kernel-based estimator to the residuals. As seen in the simulation experiment reported in Section 5.2.2, the AR lag order chosen by AIC is reasonably effective in approximating a low-order MA process, whereas the lag order chosen by BIC tends to be too conservative. Thus, at least in this case, applying a kernel-based spectral estimator to the prewhitened residuals may be more advantageous when the parametric model is chosen by BIC rather than AIC.

When the practitioner applies a kernel-based spectral estimator to the prewhitened residuals, particular care should be given to the method of determining the bandwidth parameter. For example, when an ARMA model is used to prewhiten the data, any remaining serial correlation will typically be exhibited at relatively long lag lengths. The data-dependent bandwidth selection procedure of Andrews (1991), which only considers the first-order autocorrelation, would appear to be unlikely to detect this form of serial correlation. As discussed in Section 4.3.2, the bandwidth selection procedure of Newey and West (1994) considers a larger number of autocovariances, making it

somewhat more effective in detecting higher-order serial correlation. However, when the parametric model has successfully prewhitened the residuals, this feature may generate a high bandwidth parameter and induce excessive sampling variation.

6. CONCLUDING COMMENTS.

Kernel-based and parametric covariance matrix estimation procedures are both consistent under fairly general conditions of heteroskedasticity and serial correlation. Nevertheless, each procedure requires the practitioner to make choices which have important implications in finite samples. Since the estimated HAC covariance matrix can be very sensitive to the method of determining the bandwidth parameter (for a kernel-based procedure) or the lag order (for a parametric procedure), it would generally be appropriate to utilize more than one approach in estimating the covariance matrix. Fortunately, as seen in Section 2, a number of alternative procedures are available for this purpose. However, if only a single HAC covariance matrix estimation procedure is to be used, we would recommend the parametric approach for the following reasons:

(1) The parametric VAR or ARMA estimation procedures can utilize a measure of the goodness-of-fit in determining the appropriate lag order in finite samples. In particular, a model selection criterion can be used to evaluate the tradeoff between parsimony and goodness-of-fit. Such criteria do not necessarily yield the optimal lag order, but seem to avoid the most egregious errors in practical applications.

In contrast, data-dependent bandwidth selection methods require the calculation of initial estimates of the spectral density and its first or second derivative at frequency zero. As documented in Section 4.3, poor initial spectral estimates can lead to rather absurd values for the bandwidth parameter, inducing excessive bias and/or variance of the kernel-based covariance matrix estimate, and severe distortions in subsequent inference.

(2) Kernel-based estimators incur substantial bias to ensure a positive semi-definite covariance matrix: weights less than unity are assigned to autocovariances at lags less than the bandwidth parameter, with the weights declining toward zero as the autocovariance lag increases. In contrast, the VARHAC estimator exhibits essentially the same bias as the truncated kernel estimator, which places unit weight on all autocovariances up to the bandwidth parameter. However, the truncated kernel does not

ensure a positive semi-definite covariance matrix, whereas the VARHAC estimator is positive semi-definite by construction. Thus, as discussed in Section 3, the VARHAC estimator converges to the true covariance matrix at a faster rate than any positive semi-definite kernel-based estimator. This bias differential is also evident in the simulation experiments reported in Sections 4.3.2 and 5.2.1: even for a low-order MA process, the AR spectral estimator provides a better approximation than the kernel-based estimators.

(3) To ensure that the estimated covariance matrix is positive semi-definite, kernel-based procedures must utilize a single bandwidth parameter in calculating all elements of the spectral density matrix at frequency zero. If some components of the vector of residuals exhibit high-order autocorrelation, while other components are close to white noise, then imposing the same bandwidth for both sets of variables tends to generate very ill-behaved estimates of the spectral density matrix at frequency zero.
In contrast, parametric estimators do not face such an unpleasant compromise: a different lag order can be chosen for each component of the residual vector, since the parametric estimator of the spectral density matrix is positive semi-definite by construction.

REFERENCES

Akaike, H., 1973, *Information Theory and an Extension of the Maximum Likelihood Principle, in Second International Symposium on Information Theory*, B.N. Petrov and F.Csaki, eds., Akademia Kiado (Budapest), pp. 267-281.

Andrews, D.W.K., 1991, Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica* 59, pp. 817-858.

Andrews, D.W.K., and J.C. Monahan, 1992, An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator, *Econometrica* 60, pp. 953-966.

Bartlett, M. S., 1946, On the theoretical specification of sampling properties of autocorrelated time series, *Journal of the Royal Statistical society supplement* 8, pp. 27-41.

Berk, K.N., 1974, Consistent autoregressive spectral estimates, *The Annals of Statistics* 2, pp. 489-502.

Burnside, C., and M. Eichenbaum, 1994, Small sample properties of generalized method of moments based Wald tests, manuscript. University of Pittsburgh and Northwestern University.

Christiano, L.J., and Wouter J. den Haan, 1996, Small sample properties of GMM for business cycle analysis, *Journal of Business and Economic Statistics*, forthcoming.

Davidson, J., 1995, *Stochastic Limit Theory*, Oxford University Press.

Den Haan, W.J., and A. Levin, 1994, Inferences from parametric and non-parametric covariance matrix estimation procedures, manuscript, UCSD and Federal Reserve Board.

Eichenbaum, M.S., L.P. Hansen, and K.J. Singleton, 1988, A time series analysis of representative agent models of consumption and leisure choice under uncertainty, *Quarterly Journal of Economics* CIII, pp. 51-78.

Fama, E.F., and K.R. French, 1988, Permanent and temporary components of stock prices, *Journal of Political Economy* 96, pp. 246-273.

Gallant, A.R., 1987, *Nonlinear Statistical Models*, Wiley Press (New York).

Gallant, A.R., and H. White, 1988, *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Basil Blackwell (New York).

Grenander, U., and G. Szegö, 1958, *Toeplitz Forms and Their Applications*, University of California Press.

Hamilton, J. D., 1994, *Time Series Analysis*, Princeton Univ. Press.

Hannan, E.J., 1970, *Multiple Time Series*, Wiley Press.

Hannan, E.J., and M. Deistler, 1988, *The Statistical Theory of Linear Systems*, Wiley Press.

Hannan, E.J., and L. Kavalieris, 1983, The convergence of autocorrelations and autoregressions, *Australian Journal of Statistics* 25(2), pp. 287-297.

Hannan, E.J., and L. Kavalieris, 1986, Regression and autoregression models, *Journal of Time Series Analysis* 7, pp. 27-49.

Hannan, E.J., and J. Rissanen, 1982, Recursive estimation of ARMA order, *Biometrika* 69, pp. 81-94.

Hansen, B. E. 1992, Consistent covariance matrix estimation for dependent heterogeneous processes, *Econometrica* 60, pp. 967-972.

Hansen, L.P., 1982, Large sample properties of generalized method of moments estimators, *Econometrica* 50, pp. 1029-2054.

Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl, and T.C. Lee, 1985, *The Theory and Practice of Econometrics*, 2nd edition. Wiley Press.

King, R.G., and S. Rebelo, 1993, Low frequency filtering and real business cycles, *Journal of Economic Dynamics and Control* 17, pp. 207-231.

Lee, C.C., and P.C.B. Phillips, 1994, An ARMA prewhitened long-run variance estimator, manuscript, Yale University.

Lütkepohl, H., 1985, Comparison of criteria for estimating the order of a vector autoregressive process, *Journal of Time Series Analysis* 6, pp. 35-52.

Newey, W.K., and K.D. West, 1987, A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, pp. 703-708.

Newey, W.K., and K.D. West, 1994, Automatic lag selection in covariance matrix estimation, *Review of Economic Studies* 61, pp. 631-653.

Ogaki, M., 1992, An introduction to the generalized method of moments, University of Rochester working paper No. 370.

Parzen, E., 1957, On consistent estimates of the spectrum of a stationary time series, *Annals of Mathematical Statistics* 28, pp. 329-348.

Priestley, M.B., 1982, *Spectral analysis and time series*, Academic Press.

Robinson, P.M., 1991, Automatic frequency domain inference on semiparametric and nonparametric models, *Econometrica* 59, pp. 1329-1364.

Robinson, P.M., 1995, Inference-without-smoothing in the presence of nonparametric autocorrelation, manuscript, London School of Economics.

Shibata, R., 1976, Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika* 63(1), pp. 117-126.

Shibata, R., 1980, Asymptotically efficient selection of the order of the model for estimating parameters of a linear process, *Annals of Statistics* 8, pp. 147-164.

Shibata, R., 1981, An optimal autoregressive spectral estimate, *Annals of Statistics* 9, pp. 300-306.

Schwarz, G., 1978, Estimating the dimension of a model, *Annals of Statistics* 6, pp. 461-464.

West, K.D., 1994, Another heteroskedasticity and autocorrelation consistent covariance matrix estimator, manuscript, University of Wisconsin.

White, H., 1984, *Asymptotic Theory for Econometricians*, Academic Press.