UC San Diego
Recent Work

Title
Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices

Permalink
https://escholarship.org/uc/item/7qg2m9rz

Author
Politis, Dimitris

Publication Date
2005-03-01

eScholarship.org Powered by the California Digital Library
University of California

Higher-order accurate, positive semi-definite estimation of large-sample covariance and spectral density matrices

Dimitris N. Politis*
Dept. of Mathematics and Economics
University of California, San Diego
La Jolla, CA 92093-0112, USA
dpolitis@ucsd.edu

March 14, 2005

1 Introduction

Many applications of time series econometrics—such as hypothesis tests from generalized method of moments estimation (Hansen (1982)) or general dynamic models (Gallant and White (1988))—require accurate estimation of large-sample covariance matrices that is robust to autocorrelation and heteroskedasticity. A general theory towards heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimation has been put forth in the landmark papers of Newey and West (1987) and Andrews (1991); see also the related work of Gallant (1987), Andrews and Monahan (1992), Hansen (1992), and Newey and West (1994).

Nevertheless, the current state-of-the-art seems to be lacking in three respects:
(a) The accuracy of the HAC covariance estimators is suboptimal; their rate of convergence is $T^{2/5}$ even in situations when higher-order accuracy is possible, e.g., a rate closer to $T^{1/2}$.
(b) The problem of optimal bandwidth choice for the HAC estimators has not been con-

*Research partially supported by NSF grant SES-04-18136 funded jointly by the Economics and Statistics Divisions of NSF. Many thanks are due to Dimitrios Gatzouras for his help with the proof of Lemma 8.1, and to Arthur Berg for a careful proof-reading and bringing the Eaton-Tyler paper to the author's attention.

clusively addressed. For example, the 'plug-in' procedure of Andrews (1991) will not give consistent estimation of the optimal bandwidth unless the parametric model used to estimate the 'plug-in' values holds true. On the other hand, cross-validation methods may give consistent bandwidth estimates but their consistency is typically achieved at a very slow rate; see e.g. Robinson (1991) and the references therein.
(c) The existing literature focuses on obtaining a single optimal bandwidth, common for estimating all elements of the target matrix; this is suboptimal as each element of the target matrix generally comes with its own individual optimal bandwidth.

In this note we attempt to fix the above three issues. A new class of HAC covariance matrix estimators is proposed based on the notion of a flat-top kernel as in Politis and Romano (1995) and Politis (2001). The new estimators are shown to be higher-order accurate when higher-order accuracy is possible, and a discussion on kernel choice is given.

The higher-order accuracy of flat-top kernel estimators typically comes at the sacrifice of the positive semi-definite property. Nevertheless, we show how a modified flat-top estimator is positive semi-definite while maintaining its higher-order accuracy. In addition, it is shown that there is an easy (and consistent) procedure for optimal bandwidth choice for flat-top kernel HAC estimators; this procedure estimates the optimal bandwidth associated with each individual element of the target matrix.

Since estimation of the large-sample covariance matrix of a sample mean or generalized method of moments estimator is tantamount to estimation of a spectral density matrix evaluated at the origin, the paper treats the more general problem of higher-order accurate, positive semi-definite estimation of spectral density matrices. The problem of spectral estimation under a potential lack of finite fourth moments is also addressed.

2 Background

Consider the general framework of Andrews (1991) or Hansen (1992) in which the problem at hand is estimation of the large-sample covariance matrix Ω of the sample mean of a second-order stationary (and weakly dependent) sequence of mean zero random vectors $V_t = V_t(\theta)$, $t = 1, \ldots, T$, where V_t takes values in $\mathbb{R}^d$, i.e.,

$$\Omega = \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{T} \sum_{j=1}^{T} EV_k V_j'. \tag{1}$$

Here θ is an unknown parameter assumed to have a $\sqrt{T}$–consistent estimator $\hat{\theta}$, yielding the estimated sequence $\hat{V}_t = V_t(\hat{\theta})$. We then define the usual autocovariance estimators

$$\hat{\hat{\Gamma}}(j) = \frac{1}{T}\sum_{t=1}^{T-j} \hat{V}_t \hat{V}_{t+j}' \quad \text{for} \quad j \geq 0, \quad \text{and} \quad \hat{\hat{\Gamma}}(j) = \hat{\hat{\Gamma}}(-j)' \quad \text{for} \quad j < 0.$$

The general HAC kernel estimator of Ω has the form

$$\hat{\hat{\Omega}} = \sum_{j=-T}^{T} \kappa(j/s_{_T})\hat{\hat{\Gamma}}(j),$$

where the kernel $\kappa(\cdot)$ and the bandwidth/truncation parameter $s_{_T} \in [1, T]$ satisfy some standard conditions. A typical condition on κ is:

$$\kappa : \mathbb{R} \to [-1, 1], \ \kappa \text{ is symmetric, continuous at 0 and for all but a finite number of points,}$$
$$\text{and satisfying} \ \ \kappa(0) = 1 \ \ \text{and} \quad \int_{\mathbb{R}} \kappa^2(x)dx < \infty, \tag{2}$$

The kernel $\kappa(\cdot)$ is called a 'spectral window generator' by Andrews (1991) as it corresponds to the function $K(w) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} \kappa(j)e^{-ijw}$ that is useful for smoothing the periodogram; here $i = \sqrt{-1}$. With the exception of the 'truncated' window defined as $\kappa_{trunc}(x) = 1$ if $|x| \leq 1$, and $\kappa_{trunc}(x) = 0$ else, the kernels considered by Andrews (1991) and Newey and West (1987) are positive semi-definite, i.e., their respective spectral window $K(w)$ is a nonnegative function. Nevertheless, this is not a useful restriction inasmuch as higher-order accuracy of $\hat{\hat{\Omega}}$ is concerned; more details are found in the next Section.

We now consider the idealized estimator

$$\hat{\Omega} = \sum_{j=-T}^{T} \kappa(j/s_{_T})\hat{\Gamma}(j), \tag{3}$$

that is computed as if the sequence $V_t, t = 1, \ldots, T$ were directly observable; in the above,

$$\hat{\Gamma}(j) = \frac{1}{T}\sum_{t=1}^{T-j} V_t V_{t+j}' \quad \text{for} \quad j \geq 0, \quad \text{and} \quad \hat{\Gamma}(j) = \hat{\Gamma}(-j)' \quad \text{for} \quad j < 0. \tag{4}$$

Interestingly, the estimators $\hat{\Omega}$ and $\hat{\hat{\Omega}}$ are asymptotically equivalent under general conditions such as Assumptions A, B and C of Andrews (1991) or Condition (V2) of Hansen (1992); see e.g. Theorem 1(b) of Andrews (1991). Intuitively, this is due to the slower rate of convergence of both $\hat{\Omega}$ and $\hat{\hat{\Omega}}$ as compared to the $\sqrt{T}$–consistency of $\hat{\theta}$ and $V_t(\hat{\theta})$.

In view of the results of our next Section, we now give a slight generalization of Theorem 1(b) of Andrews (1991) to cover a possible choice of the bandwidth parameter s_T that does not necessarily tend to infinity (or it does at a slow, logarithmic rate); see e.g. Theorem 3.1 (ii) and (iii) in what follows.

Lemma 2.1 *Assume Assumptions A, B and C of Andrews (1991) hold true, and that κ satisfies eq. (2). Further assume that, as $T \to \infty$, we have $s_T/T \to 0$ and that:*
(i) $s_T^{-1} \sum_{j=-T+1}^{T-1} |\kappa(j/s_T)| = O(1)$;
(ii) $Bias(\hat{\Omega}) = O(\sqrt{s_T/T})$; and
(iii) $s_T \to \infty$ or $EV_t \frac{\partial}{\partial\theta} V_{t-j} = 0$ for all j.
Then, $\hat{\hat{\Omega}} = \Omega + O_P(\sqrt{s_T/T})$, $\hat{\Omega} = \Omega + O_P(\sqrt{s_T/T})$, and $\hat{\Omega} - \hat{\hat{\Omega}} = o_P(\sqrt{s_T/T})$.

Note that condition (i) of Lemma 2.1 is immediately satisfied if the kernel κ 'cuts-off', e.g., if $\kappa(x) = 0$ for $|x| >$ some x_0. Condition (ii) of Lemma 2.1 can be viewed as a restriction (a lower bound) on the rate of growth of s_T.

In view of Lemma 2.1, in what follows we will focus on theoretically analyzing (our version of) $\hat{\Omega}$, safe in the knowledge that the asymptotic behavior of the corresponding $\hat{\hat{\Omega}}$ will be identical.

3 Spectral density matrix estimation

Here, and throughout the rest of the paper, we consider observations $V_1, \dots, V_T$ from a second-order stationary d-variate time series $\{V_t, t \in \mathbb{Z}\}$ possessing mean zero and autocovariance matrix sequence $\Gamma(j)$ defined as

$$\Gamma(j) = EV_t V_{t+j}' \quad \text{for} \quad j \geq 0, \quad \text{and} \quad \Gamma(j) = \Gamma(-j)' \quad \text{for} \quad j < 0. \tag{5}$$

Under typical weak dependence conditions—see e.g. Hannan (1970), Brillinger (1981), Brockwell and Davis (1991), or Hamilton (1994)—the spectral density matrix evaluated at point w is defined as

$$F(w) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \Gamma(k) e^{-ikw} \tag{6}$$

where $i = \sqrt{-1}$. The dxd matrix $F(w)$ is positive semi-definite and Hermitian for any $w \in [-\pi, \pi]$ but note that its off-diagonal elements are, in general, complex-valued; $F_{jk}(w)$

will denote the (j,k) element of $F(w)$. Nevertheless, $F(0)$ has all its elements real-valued, and it is easy to see that $F(0) = \Omega/(2\pi)$ where Ω was defined in eq. (1). Hence, accurate estimation of $F(0)$ is tantamount to accurate estimation of Ω. In what follows, we will consider the more general problem of estimation of $F(w)$ at an arbitrary (fixed) point $w \in [-\pi, \pi]$; since w will be fixed, the short-hand notation F will be used to denote $F(w)$, and F_{jk} will denote the (j,k) element of F.

To describe our new spectral matrix estimator, we need the notion of a 'flat-top' kernel. The general family of flat-top kernels was introduced in Politis (2001). Its typical member is $\lambda_{g,c}(x)$ where

$$\lambda_{g,c}(x) = \begin{cases} 1 & \text{if } |x| \le c \\ g(x) & \text{else;} \end{cases} \tag{7}$$

here $c > 0$ is a parameter, and $g : \mathbb{R} \to [-1,1]$ is a symmetric function, continuous at all but a finite number of points, and satisfying $g(c) = 1$, and $\int_{\mathbb{R}} g^2(x)dx < \infty$. The kernel $\lambda_{g,c}(x)$ is 'flat', i.e., constant, over the region $[-c, c]$, hence the name flat-top.

If g is such that $g(x) = 0$ for $|x| \ge$ some x_0, then the kernel $\lambda_{g,c}(x)$ has a hard cut-off. The simplest representative of such a flat-top kernel has a trapezoidal shape defined as

$$\lambda_{TR,c}(x) = \begin{cases} 1 & \text{if } |x| \le c \\ \frac{|x|-1}{c-1} & \text{if } c < |x| \le 1 \\ 0 & \text{else} \end{cases} \tag{8}$$

with $c \in (0,1]$, i.e., the function g performs a linear interpolation between the values $g(c) = 1$ and $g(1) = 0$. The trapezoidal kernel's favorable properties were documented in Politis and Romano (1995). The trapezoid may be seen as a cross between the square truncated kernel $\kappa_{trunc}(x)$, and the well-known triangular Bartlett kernel $\kappa_B(x) = (1 - |x|)^+$; the notation $(y)^+$ indicates the positive part of y, i.e., $(y)^+ = \max(y, 0)$.

Let S be a dxd matrix of bandwidth parameters with (j,k) element denoted by S_{jk}. As usual, S is thought of as a function of T although this dependence will not be explicitly denoted. The estimator of F that we will consider is $\hat{F}$ with (j,k) element given by:

$$\hat{F}_{jk} = \frac{1}{2\pi} \sum_{m=-T}^{T} \lambda_{g,c}(m/S_{jk}) \hat{\Gamma}_{jk}(m) e^{-imw} \tag{9}$$

where $\lambda_{g,c}$ is some chosen member of the flat-top family, and $\hat{\Gamma}_{jk}(m)$ is the (j,k) element of the sample autocovariance matrix $\hat{\Gamma}(m)$ defined in eq. (4). Note that the dependence of $\hat{F}_{jk}$ on the chosen $\lambda_{g,c}$ is not explicitly denoted.

The favorable large-sample properties of $\hat{F}$ are manifested in the following theorem.

Theorem 3.1 *Assume conditions strong enough to ensure that**

$$Var(\hat{F}_{jk}) = O(S_{jk}/T) \quad \textit{for any fixed} \;\; j,k; \tag{10}$$

Then, for each combination of j and k, the following are true.
(i) If $\sum_{m=-\infty}^{\infty} |m|^r |\Gamma_{jk}(m)| < \infty$ for some real number $r \geq 1$, then letting S_{jk} proportional to $T^{1/(2r+1)}$ yields

$$\hat{F}_{jk} = F_{jk} + O_P(T^{-r/(2r+1)}).$$

(ii) If $|\Gamma_{jk}(m)| \leq Ce^{-am}$ for some constants $C, a > 0$, then letting $S_{jk} \sim A \log T$, for some appropriate constant A, yields

$$\hat{F}_{jk} = F_{jk} + O_P(\frac{\sqrt{\log T}}{\sqrt{T}});$$

as usual, the notation $A \sim B$ means $A/B \to 1$.
(iii) If $\Gamma_{jk}(m) = 0$ for $|m| >$ some q, then letting $S_{jk} = \max(\lceil q/c \rceil, 1)$, yields†

$$\hat{F}_{jk} = F_{jk} + O_P(\frac{1}{\sqrt{T}});$$

here $\lceil x \rceil$ is the 'ceiling' function, i.e., the smallest integer larger or equal to x.

The conditions of the three parts of Theorem 3.1 are usual conditions of weak dependence. For example, if $\Gamma_{jj}(m) = 0$ for $|m| >$ some q, then the jth coordinate of V_t, say $V_t^{(j)}$, can be thought to follow a Moving Average (MA) model of order q. Similarly, the condition $|\Gamma_{jj}(m)| \leq Ce^{-am}$ is satisfied if $V_t^{(j)}$ follows a stationary ARMA (p,q) model, i.e., AutoRegressive with Moving Average residuals; see e.g. Brockwell and Davis (1991). The polynomial decay in condition (i) is a worst-case scenario; suffices to note that in order to even define the spectral density of $V_t^{(j)}$ the typical condition is $\sum_{m=-\infty}^{\infty} |\Gamma_{jj}(m)| < \infty$, i.e., $r = 0$ in condition (i).

Theorem 3.1 gives the rate of convergence of $\hat{F}_{jk}$ to F_{jk}, at the same time suggesting the optimal values of the bandwidth parameter S_{jk}; here optimality is meant with respect to

*There exist different sets of conditions sufficient for eq. (10). Assumption A of Andrews (1991) is such a condition based on summability of fourth cumulants; different conditions based on moment and mixing assumptions are also available, see e.g. Hannan (1970), Brillinger (1981), or Brockwell and Davis (1991).

†Taking the maximum of $\lceil q/c \rceil$ and 1 is done to cover the possibility that $q = 0$.

optimizing the rate of convergence of $\hat{F}_{jk}$. As is apparent, the optimal S_{jk} crucially depends on the rate of decay of $\Gamma_{jk}(m)$ as m increases. If we had some reason to believe that the rate of decay of $\Gamma_{jk}(m)$ is the *same* for all j, k, then we could let S_{jk} equal some common value s_T, in which case our estimator would take the familiar simple form

$$\hat{F}_{simple} = \frac{1}{2\pi} \sum_{m=-T}^{T} \lambda_{g,c}(m/s_T)\hat{\Gamma}(m)e^{-imw}; \tag{11}$$

letting $w = 0$, it is seen that the above is of the same exact form as the Newey-West (1987) and Andrews (1991) estimator $\hat{\Omega}$ given in eq. (3). Nevertheless, there is typically no reason to believe that the rate of decay of $\Gamma_{jk}(m)$ is common for all j, k. Thus, $\hat{F}$ is generally preferable to $\hat{F}_{simple}$.

To elaborate, consider the following example. Let $V_t = (V_t^{(1)}, V_t^{(2)}, V_t^{(3)})'$ where $V_t^{(1)}$ follows an MA(q_1) model, $V_t^{(2)}$ follows an MA(q_2) model independent of $V_t^{(1)}$, and $V_t^{(3)} = V_{t-L}^{(2)}$ for all t. Suppose that the trapezoidal kernel $\lambda_{TR,1/2}(x)$ is used, i.e., $c = 1/2$. Then, Theorem 3.1 (iii) suggests the following optimal bandwidth parameters: $S_{11} = 2q_1$, $S_{22} = 2q_2$, $S_{33} = 2q_2$, $S_{12} = S_{21} = 1$, $S_{13} = S_{31} = 1$, and $S_{23} = S_{32} = 2(q_2 + L)$.

Parts (ii), (iii)—as well as part (i) with $r > 2$—of Theorem 3.1 show that the rate of convergence of $\hat{F}$ is superior to the Newey-West (1987) estimator based on Bartlett's kernel, as well as to all second order kernel estimators considered by Andrews (1991); the Newey-West (1987) estimator only achieves a rate of convergence of $T^{1/3}$, while the second order kernels (including the optimal quadratic spectral window) achieve a rate of convergence of $T^{2/5}$.

4 Spectral estimation in the absence of finite fourth moments

As mentioned in the last section, eq. (10) is typically satisfied for kernel estimators such as $\hat{F}$. Nevertheless, if the series $\{V_t\}$ does not possess finite fourth moments, then $Var(\hat{F}_{jk})$ is not well-defined. For this reason, it is convenient to also define the correlation/cross-correlation matrix $\rho(m)$ with (j,k) element given by $\rho_{jk}(m) = \Gamma_{jk}(m)/\sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}$, and estimated by $\hat{\rho}_{jk}(m) = \hat{\Gamma}_{jk}(m)/\sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)}$. We can then define the normalized spectral density matrix evaluated at point w as

$$f(w) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k) e^{-ikw}; \tag{12}$$

the short-hand notation f will again be used to denote $f(w)$, and f_{jk} will denote the (j,k) element of f. The corresponding flat-top kernel estimator of f is $\hat{f}$ with (j,k) element given by:

$$\hat{f}_{jk} = \frac{1}{2\pi} \sum_{m=-T}^{T} \lambda_{g,c}(m/S_{jk}) \hat{\rho}_{jk}(m) e^{-imw}. \tag{13}$$

Because $\hat{\rho}_{jk}(m)$ is bounded (by unity), $Var(\hat{f}_{jk})$ is well-defined even if $\{V_t\}$ does not possess finite fourth moments. The following alternative to eq. (10) is then suggested:

$$Var(\hat{f}_{jk}) = O(S_{jk}/T) \quad \text{for any fixed} \quad j,k. \tag{14}$$

Eq. (14) is now typically satisfied under regularity conditions; see e.g. Robinson (1991) and Hansen (1992) who considered the problem of spectral estimation in the absence of finite fourth moments.

A further consequence of lack of finite fourth moments is that, although $\hat{\rho}(m)$ will still be $\sqrt{T}$—consistent under appropriate weak dependence assumptions, $\hat{\Gamma}(m)$ is consistent but typically at slower rate; see e.g. Brockwell and Davis (1991) or Embrechts et al. (1997). A reasonable assumption adopted by Robinson (1991) is:

$$\hat{\Gamma}_{jj}(0) = \Gamma_{jj}(0) + O_P(1/T^{\alpha}), \quad \text{for all} \quad j, \quad \text{and some} \quad \alpha \in (0, 1/2]. \tag{15}$$

For our purposes we will require the slightly stronger condition:

$$E\left|\hat{\Gamma}_{jj}(0) - \Gamma_{jj}(0)\right|^{1+\delta} = O(1/T^{\alpha(1+\delta)}) \quad \text{for all} \quad j, \quad \text{and some} \quad \delta > 0 \quad \text{and} \quad \alpha \in (0, 1/2]. \tag{16}$$

The following theorem is a generalization of Theorem 3.1 to the setting where finite fourth moments are potentially lacking.

Theorem 4.1 *Fix values for j,k, and assume conditions (14), (16), and that*[‡]

$$S_{jk}^{-1} \sum_{j=-T+1}^{T-1} |\lambda_{g,c}(j/S_{jk})| = O(1). \tag{17}$$

[‡]As in condition (i) of Lemma 2.1, eq. (17) is easily satisfied such as when $\lambda_{g,c}(x)$ has a hard 'cut-off', i.e., $\lambda_{g,c}(x) = 0$ for $|x| >$ some x_0.

Also assume $\Gamma_{jj}(0) > 0$ *for all* j.
(i) If $\sum_{m=-\infty}^{\infty} |m|^r |\Gamma_{jk}(m)| < \infty$ *for some real number* $r \geq 1$, *then letting* S_{jk} *proportional to* $T^{\alpha/(r+1)}$ *yields*

$$\hat{f}_{jk} = f_{jk} + O_P(T^{-\alpha r/(r+1)}), \tag{18}$$

and

$$\hat{F}_{jk} = F_{jk} + O_P(T^{-\alpha r/(r+1)}). \tag{19}$$

(ii) If $|\Gamma_{jk}(m)| \leq Ce^{-am}$ *for some constants* $C, a > 0$, *then letting* $S_{jk} \sim A \log T$, *for some appropriate constant* A, *yields*

$$\hat{f}_{jk} = f_{jk} + O_P(\frac{\log T}{T^\alpha}) \quad \textit{and} \quad \hat{F}_{jk} = F_{jk} + O_P(\frac{\log T}{T^\alpha}). \tag{20}$$

(iii) If $\Gamma_{jk}(m) = 0$ *for* $|m| >$ *some* q, *then letting* $S_{jk} = \max(\lceil q/c \rceil, 1)$, *yields*

$$\hat{f}_{jk} = f_{jk} + O_P(\frac{\log\log T}{T^\alpha}) \quad \textit{and} \quad \hat{F}_{jk} = F_{jk} + O_P(\frac{\log\log T}{T^\alpha}) \tag{21}$$

Note that, even under the potential absence of finite fourth moments, $\hat{F}$ maintains its higher-order accuracy. Parts (ii) and (iii) of Theorem 4.1 show that the rate of convergence of $\hat{F}$ comes very close to T^α which is the rate of convergence of $\hat{\Gamma}(0)$. Interestingly, under the premises of either part (ii) or (iii) of Theorem 4.1, the optimal rates for the bandwidth S_{jk} are insensitive to whether fourth moments are finite or not.

5 Positive semi-definite spectral estimation

Flat-top kernels are infinite-order kernels, and therefore they are capable of achieving higher-order accuracy when that is possible. For example, it is apparent that, under the MA(q)–type condition of Theorem 3.1 (iii), $\sqrt{T}$–consistent estimation of F_{jk} is possible since F_{jk} is a function of only finitely many (q) parameters. The flat-top estimator $\hat{F}_{jk}$ indeed attains $\sqrt{T}$–consistency in that case, and the flatness of the kernel over the interval $[-c, c]$ is crucial for this attainment.

The disadvantage of flat-top kernels, however, is that they are not positive semi-definite, i.e., the matrix $\hat{F}$ is not almost surely positive semi-definite for all w. The fast rate of convergence of $\hat{F}$ to a positive semi-definite matrix indicates that the incidents of a non-positive semi-definite $\hat{F}$ may be rare; this fact was documented in the simulations of Andrews (1991) with respect to the truncated kernel that technically belongs to the flat-top family.[§]

However, the positive semi-definiteness is an important philosophical point especially in the case of $w = 0$ when the object is estimation of a covariance matrix. It is likely for this reason that the focus in the recent literature starting with Newey-West (1987) has been on positive semi-definite estimators. Nonetheless, we now show how the flat-top estimator $\hat{F}$ can be easily modified to render a positive semi-definite estimator.

Recall that a Hermitian matrix has all real eigenvalues, and can be diagonalized by a unitary transformation. Thus, consider the unitary decompositions of the Hermitian matrices F and $\hat{F}$, namely:

$$F = U\Lambda U^* \quad \text{and} \quad \hat{F} = \hat{U}\hat{\Lambda}\hat{U}^* \tag{22}$$

where $U, \hat{U}$ are unitary (complex-valued) matrices, i.e., they satisfy $U^{-1} = U^*$ and $\hat{U}^{-1} = \hat{U}^*$ where * denotes the conjugate transpose; the columns of U and $\hat{U}$ are the orthonormal eigenvectors of F and $\hat{F}$ respectively, and $\Lambda = diag(\lambda_1, \ldots, \lambda_d)$, $\hat{\Lambda} = diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$ are diagonal matrices containing the respective eigenvalues.

Noting that the entries of Λ are all nonnegative suggests the following fix to the possible negativity of $\hat{F}$. Let $\hat{\Lambda}^+ = diag(\hat{\lambda}_1^+, \ldots, \hat{\lambda}_d^+)$ where $\hat{\lambda}_j^+ = \max(\hat{\lambda}_j^+, 0)$, i.e., the entries of $\hat{\Lambda}^+$ are given by the positive part of the entries of $\hat{\Lambda}$, and define the positive semi-definite estimator

$$\hat{F}^+ = \hat{U}\hat{\Lambda}^+\hat{U}^*. \tag{23}$$

The following theorem shows that, in addition to being positive semi-definite, $\hat{F}^+$ inherits the higher-order accuracy of $\hat{F}$; $\hat{F}^+$ is therefore our proposed higher-order accurate, positive semi-definite estimator.

Theorem 5.1 *Let R_T be a sequence such that $R_T \to \infty$ as $T \to \infty$. If $\hat{F} = F + O_P(1/R_T)$, then $\hat{F}^+ = F + O_P(1/R_T)$ as well.*[¶]

[§]Note, however, that the discontinuity of the truncated kernel gives its corresponding spectral window very pronounced 'sidelobes', and hence high variance (because of large l_2–norm) and unfavorable finite-sample behavior; see e.g. Politis and Romano (1995). More details on kernel choice are given in Section 6.

[¶]The notation $A = O_P(1/R_T)$ for some matrix A means that each element of A is $O_P(1/R_T)$.

6 Flat-top kernel choice

The favorable asymptotic rates of Theorems 3.1 and 4.1 are achievable by any member of the flat-top family. Nevertheless, finite-sample properties will be dependent upon kernel choice. For example, as mentioned in the previous section, the truncated kernel $\kappa_{trunc}(x)$ is one of the worse representatives of the flat-top family because of the pronounced 'sidelobes' of the Dirichlet kernel which is its corresponding spectral window—see e.g. Figure 2 of Politis and Romano (1995). Since half of those sidelobes are on the negative side, they unnecessarily inflate the L_2–norm of the spectral window under the constraint that its L_1–norm is unity; as is well-known, a large L_2–norm implies a large variance.[‖]

In order to reduce the size of a spectral window's sidelobes, the flat-top kernel must be chosen as smooth as possible. The poor finite-sample performance of the truncated kernel is due to the discontinuity of the function $\kappa_{trunc}(x)$ at points ± 1. The trapezoidal kernel $\lambda_{TR,c}(x)$ is continuous everywhere, and is thus much better performing than the truncated. Even better finite-sample behavior is expected if the 'corners' of the trapezoid $\lambda_{TR,c}(x)$ are smoothed out. For example, McMurry and Politis (2004) constructed a member of the flat-top family that is infinitely differentiable; it is defined as

$$\lambda_{ID,b,c}(x) = \begin{cases} 1 & \text{if} \quad |x| \leq c \\ \exp\left(-b\exp(-b/(|x|-c)^2)/(|x|-1)^2\right) & \text{if} \quad c < |x| < 1 \\ 0 & \text{if} \quad |x| \geq 1 \end{cases} \tag{24}$$

where $c \in (0, 1]$, and $b > 0$ is a shape parameter, making the transition from $\lambda_{ID,b,c}(c) = 1$ to $\lambda_{ID,b,c}(1) = 0$ more or less abrupt.

Nevertheless, the already good performance of the trapezoidal kernel indicates that one might not have to use an infinitely differentiable kernel to gather appreciable finite-sample benefits. For example, we can create a flat-top kernel by adding a piecewise cubic tail, similar to that of Parzen's (1961) kernel, to the $[-c, c]$ flat-top region. The resulting flat-

[‖]The variance is still of order $O(S_{jk}/T)$ as eq. (10) demands, but the proportionality constant in the term $O(S_{jk}/T)$ is large for the Dirichlet kernel.

top kernel would be defined as:

$$\lambda_{PR,c}(x) = \begin{cases} 1 & \text{if} \quad 0 \le x \le c \\ 1 - 6(x-c)^2 + 6|x-c|^3 & \text{if} \quad c \le x \le c + 1/2 \\ 2(1 - |x-c|)^3 & \text{if} \quad c + 1/2 < x < c + 1 \\ 0 & \text{if} \quad x \ge c + 1 \\ \lambda_{PR,c}(-x) & \text{if} \quad x < 0. \end{cases} \tag{25}$$

Similarly, we can create a flat-top kernel by a modification of Priestley's (1962) 'quadratic spectral kernel':

$$\kappa_{QS}(x) = \frac{3}{x^2}\left(\frac{\sin x}{x} - \cos x\right)$$

that has been found optimal** among positive semi-definite second order kernels; see e.g. Priestley (1962) or Epanechnikov (1969). The modification would amount to defining:

$$\lambda_{QS,b,c}(x) = \begin{cases} 1 & \text{if} \quad 0 \le x \le c \\ \frac{3}{b^2(x-c)^2}\left(\frac{\sin(b(x-c))}{b(x-c)} - \cos(b(x-c))\right) & \text{if} \quad x > c \\ \lambda_{QS,b,c}(-x) & \text{if} \quad x < 0, \end{cases} \tag{26}$$

so that $\lambda_{QS,b,c}(x)$ has the required $[-c, c]$ flat-top region, but inherits the tails of $\kappa_{QS}(x)$. Note that $\kappa_{QS}(x)$ tends to zero for large x but does not vanish after a cut-off point. The parameter $b > 0$ in $\lambda_{QS,b,c}(x)$ is again a shape parameter scaling the magnitude of the tail. Since c 'scales' together with b, we can let $c = 1$ in connection with $\lambda_{QS,b,c}(x)$, so that b is the only remaining shape parameter.

Having chosen the shape of the function g, the remaining parameters c and/or b have to be chosen as well. For the trapezoidal kernel $\lambda_{TR,c}(x)$, the recommendation of Politis and Romano (1995) is to take c in the neighborhood of 1/2; the rationale is that the extreme values $c \to 0$ and $c \to 1$ are both to be avoided, corresponding to the aforementioned poorly performing kernels, the Bartlett and truncated kernel respectively.

For the infinitely differentiable kernel $\lambda_{ID,b,c}(x)$ there is an interplay between the two parameters b and c; for example, even with c close to 0, there is a range of values of b that will make $\lambda_{ID,b,c}(x)$ look very much like the trapezoidal $\lambda_{TR,1/2}(x)$ with ultra-smoothed

**Priestley's kernel $\kappa_{QS}(x)$ leads to the so-called Epanechnikov spectral window of quadratic form, i.e., $K_{QS}(w) = (1 - w^2)^+$ that satisfies a number of optimality criteria among positive semi-definite second order kernels; see Andrews (1991).

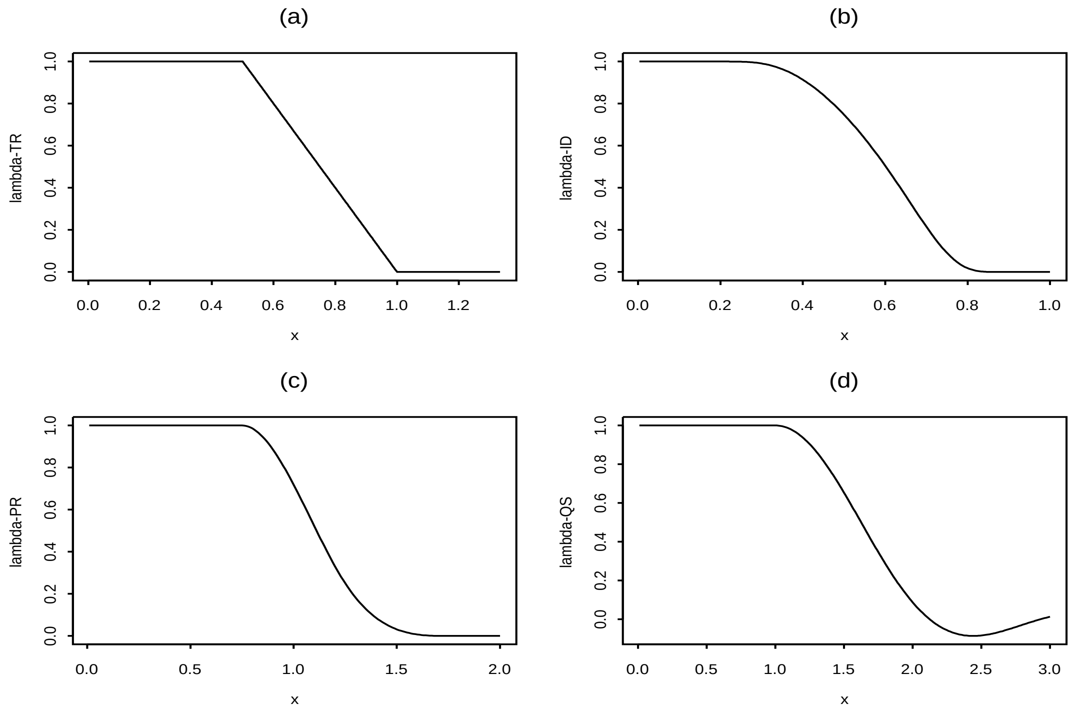

Figure 1: (a) Plot of $\lambda_{TR,1/2}(x)$ vs. $x > 0$; (b) Plot of $\lambda_{ID,0.25,0.05}(x)$ vs. $x > 0$; (c) Plot of $\lambda_{PR,0.75}(x)$ vs. $x > 0$; (d) Plot of $\lambda_{QS,4,1}(x)$ vs. $x > 0$.

corners. Similarly, to implement the kernels $\lambda_{PR,c}(x)$ and/or $\lambda_{QS,b,1}(x)$, the parameters c and b must be chosen respectively.

The problem of identifying the optimal shape of a flat-top kernel is still open, and more work is needed in that respect. In the meantime, motivated by the good performance of the trapezoidal kernel $\lambda_{TR,1/2}(x)$, the following rule-of-thumb may be suggested: choose the parameter(s) of a flat-top kernel such that the resulting shape is similar to $\lambda_{TR,1/2}(x)$ with smoothed corners. For example, letting $c = 0.05$ and $b = 1/4$ has this desired effect in connection with $\lambda_{ID,b,c}(x)$, i.e., $\lambda_{ID,0.25,0.05}(x)$ 'looks' like a smoothed version of $\lambda_{TR,1/2}(x)$. To get $\lambda_{PR,c}(x)$ and $\lambda_{QS,b,1}(x)$ to yield a similar balance between the flat-top region and the tail, the values $c = 0.75$ and $b = 4$ may be used respectively. Plots of the flat-top kernels $\lambda_{TR,1/2}(x)$, $\lambda_{ID,0.25,0.05}(x)$, $\lambda_{PR,0.75}(x)$ and $\lambda_{QS,4,1}(x)$ are shown in Figure 1.

7 Data-dependent bandwidth choice

In this section, assume that a member of the flat-top family, say $\lambda_{g,c}$, has been identified to be used for $\hat{F}^+$ and $\hat{F}$. Besides the favorable asymptotic properties and speed of convergence associated with flat-top kernels as demonstrated in Theorems 3.1 and 4.1, a further reason for using a flat-top lag-window is that choosing its bandwidth in practice is intuitive and doable by a simple inspection of the correlogram/cross-correlogram, i.e., a plot of $\hat{\rho}_{jk}(m)$ vs. m where $\hat{\rho}_{jk}(m) = \hat{\Gamma}_{jk}(m)/\sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)}$ for all j, k.

The proposed bandwidth choice rule is motivated by case (iii) of Theorems 3.1 and 4.1 and boils down to looking for a point, say $\hat{q}$, after which the correlogram appears negligible, i.e., $\hat{\rho}_{jk}(m) \simeq 0$ for $|m| > \hat{q}$ (but $\hat{\rho}_{jk}(\hat{q}) \neq 0$). Of course, $\hat{\rho}_{jk}(m) \simeq 0$ is taken to mean that $\hat{\rho}_{jk}(m)$ is not significantly different from zero, i.e., an implied hypothesis test. After identifying $\hat{q}$, the recommendation is to just take $\hat{S}_{jk} = \max(\lceil \hat{q}/c \rceil, 1)$ as part (iii) of Theorems 3.1 and 4.1 suggests. Although it may be overoptimistic to expect that our data will follow a finite-order MA(q) model, the validity of this simple rule in general situations is due to the fact that an MA(q) model—with high enough q—can always serve as an approximation at least as far as the spectral density is concerned; see e.g. Brockwell and Davis (1991).

The intuitive interpretation of the above bandwidth choice rule is an effort to extend the 'flat-top' region of $\lambda_{g,c}$ over the whole of the region where $\hat{\rho}_{jk}(m)$ is thought to be significant so as not to downweigh it and introduce bias. Nevertheless, the 'flat-top' region of $\lambda_{g,c}$ can be greater than $[-c, c]$ depending on the choice of function g. Even if $g(x)$ is strictly decreasing for $x > c$, its rate of decrease near c may be slow enough so that $\lambda_{g,c}(x) \simeq 1$ for x in an interval much greater than $[-c, c]$; see, for example, Figure 1 (b) regarding the infinitely differentiable $\lambda_{IS,b,c}(s)$ with $b = 1/4$ and $c = 0.05$. Thus, we are led to define the 'effective' flat-top region of $\lambda_{g,c}$ as the interval $[-c_{ef}, c_{ef}]$ where c_{ef} is the largest number such that $\lambda_{g,c}(x) \geq 1 - \epsilon$ for all x in $[-c_{ef}, c_{ef}]$; here ϵ is some small chosen number, e.g. $\epsilon = 0.01$.

Now we can rigorously define the empirical bandwidth choice rule. Note that in the case $j \neq k$, $\rho_{jk}(m)$ is the cross-correlation sequence which is not symmetric in m; rather than looking at both positive and negative m, we choose to look at both $\rho_{jk}(m)$ and $\rho_{kj}(m)$ for only positive m which is equivalent.

EMPIRICAL RULE OF CHOOSING S_{jk} FOR FLAT-TOP KERNEL $\lambda_{g,c}$.
Case *$j = k$: Let $\hat{q}$ be the smallest nonnegative integer such that $|\hat{\rho}_{jk}(\hat{q}+m)| < C_0\sqrt{\log_{10} T/T}$, for $m = 0, 1, \ldots, K_T$, where $C_0 > 0$ is a fixed constant, and K_T is a positive, nondecreasing integer-valued function of T such that $K_T = o(\log T)$. Then, let $\hat{S}_{jk} = \max(\lceil \hat{q}/c_{ef} \rceil, 1)$.*
Case *$j \neq k$: Let $\hat{q}_{jk}$ be the smallest nonnegative integer such that $|\hat{\rho}_{jk}(\hat{q}_{jk} + m)| < C_0\sqrt{\log_{10} T/T}$, for $m = 0, 1, \ldots, K_T$, where $C_0 > 0$ is a fixed constant, and K_T is a positive, nondecreasing integer-valued function of T such that $K_T = o(\log T)$. Similarly, let $\hat{q}_{kj}$ be the smallest nonnegative integer such that $|\hat{\rho}_{kj}(\hat{q}_{kj} + m)| < C_0\sqrt{\log_{10} T/T}$, for $m = 0, 1, \ldots, K_T$. Then, let $\hat{q} = \max(\hat{q}_{jk}, \hat{q}_{kj})$, and $\hat{S}_{jk} = \hat{S}_{kj} = \max(\lceil \hat{q}/c_{ef} \rceil, 1)$.*

In the case $j = k$, the above bandwidth choice rule was empirically suggested by Politis and Romano (1995) for the trapezoidal kernel; it was then rigorously studied in Politis (2003). Note that the constant C_0 and the form of K_T are the practitioner's choice. Politis (2003) makes the concrete recommendations $C_0 \simeq 2$ and $K_T = \max(5, \sqrt{\log_{10} T})$ that have the interpretation of yielding (approximate) 95% simultaneous confidence intervals for $\rho_{jk}(\hat{q} + m)$ with $m = 1, \ldots, K_T$ by Bonferroni's inequality. Nevertheless, the practitioner should always be vigilant in a case where altering the value of C_0 slightly leads to radically different values of $\hat{q}$. In such a case, the rule-of-thumb is to use the smaller of the two potential estimates $\hat{q}$ in the sense that flat-top kernels work best with small bandwidth parameters; see Politis and White (2004) for an example of this phenomenon.

The performance of our empirical bandwidth choice rule is quantified in the following theorem; the case $j = k$ of the theorem was given in Politis (2003) for the trapezoidal flat-top kernel.

Theorem 7.1 *Fix j, k, and assume conditions strong enough to ensure that*[††] *for all finite N,*

$$\max_{m=1,\ldots,N} |\hat{\rho}_{jk}(n + m) - \rho_{jk}(n + m)| = O_P(1/\sqrt{T}) \tag{27}$$

[††]There exist different sets of conditions sufficient for eq. (27); see Brockwell and Davis (1991) or Romano and Thombs (1996). As a matter of fact, under further regularity conditions, the process $\sqrt{T}(\hat{\rho}_{jk}(\cdot) - \rho_{jk}(\cdot))$ is asymptotically Gaussian with autocovariance tending to zero; consequently, eq. (28) would follow from the theory of extremes of dependent sequences—see e.g. Leadbetter et al. (1983).

uniformly in n, and

$$\max_{m=0,1,\ldots,T-1} |\hat{\rho}_{jk}(m) - \rho_{jk}(m)| = O_P(\sqrt{\frac{\log T}{T}}). \quad (28)$$

Also assume that the sequence $\rho_{jk}(m)$ does not have more than $K_T - 1$ consecutive zeros[‡‡] *in its first m_0 lags (i.e., for $m = 0, 1, \ldots, m_0$).*

(i) Assume that for $m > m_0$ we have $\rho_{jk}(m) = C_1 m^{-p_1}$ or $\rho_{jk}(m) = C_1 m^{-p_1} \cos(a_1 m + \theta_1)$, and $\rho_{kj}(m) = C_2 m^{-p_2}$ or $\rho_{kj}(m) = C_2 m^{-p_2} \cos(a_2 m + \theta_2)$, for some positive integers p_1, p_2, and some constants satisfying $C_v > 0$, $a_v \geq \frac{\pi}{K_T}$, and $\theta_v \in [0, 2\pi]$ for $v = 1, 2$. Then,

$$\hat{S}_{jk} \overset{P}{\sim} \frac{A_1 T^{1/(2p)}}{(\log T)^{1/(2p)}} \quad \textit{where} \quad p = \max(p_1, p_2)$$

for some positive constant A_1; the notation $A \overset{P}{\sim} B$ means $A/B \xrightarrow{P} 1$.

(ii) Assume that for $m > m_0$ we have $\rho_{jk}(m) = C_1 \xi_1^m$ or $\rho_{jk}(m) = C_1 \xi_1^m \cos(a_1 m + \theta_1)$, and $\rho_{kj}(m) = C_2 \xi_2^m$ or $\rho_{kj}(m) = C_2 \xi_2^m \cos(a_2 m + \theta_2)$, where the constants satisfy $C_v > 0$, $|\xi_v| < 1$, $a_v \geq \frac{\pi}{K_T}$, and $\theta_v \in [0, 2\pi]$ for $v = 1, 2$. Then,

$$\hat{S}_{jk} \overset{P}{\sim} A_2 \log T$$

where $A_2 = -1/\max(\log|\xi_1|, \log|\xi_2|)$.

(iii) If $|\rho_{jk}(m)| + |\rho_{kj}(m)| = 0$ for $m >$ some nonnegative integer q (with $q < m_0 + K_T$), but $|\rho_{jk}(q)| + |\rho_{kj}(q)| \neq 0$, then

$$\hat{S}_{jk} = \max(\lceil q/c_{ef} \rceil, 1) + o_P(1).$$

Comparing the empirical rule $\hat{S}_{jk}$ to the theoretically optimal values of S_{jk} given in Theorem 3.1 we see that $\hat{S}_{jk}$ manages to capture exactly the theoretically optimal rate in cases (ii) and (iii) of Theorem 7.1. In case (i) of Theorem 7.1, $\hat{S}_{jk}$ increases essentially as a power of T since the $2p$-th root of the logarithm changes in an ultra-slow way with T; note that the empirically found exponent $1/(2p)$ is slightly smaller than the theoretically optimal bandwidth given in part (i) of Theorem 3.1 but the difference is small, and becomes even smaller for large p. Thus, $\hat{S}_{jk}$ is seen to automatically adapt to the underlying rate of decay of the correlation/cross-correlation function, switching between the polynomial, logarithmic, and constant rates that are optimal respectively in the three cases of Theorems 3.1 and 4.1.

[‡‡]Because of this assumption, it is advisable to take K_T be an increasing function of T, albeit at the very slow rate suggested by the recommendation $K_T = \max(5, \sqrt{\log_{10} T})$.

8 Appendix: Technical proofs

Proof of Lemma 2.1. The case $s_T \to \infty$ is covered in Theorem 1 of Andrews (1991); thus, we now assume $EV_t \frac{\partial}{\partial\theta} V_{t-j} = 0$ for all j.

A careful reading of the proof of Theorem 1(b) of Andrews (1991) indicates that the proof first hinges on showing that $(Ts_T)^{-1/2} \sum_{j=-T+1}^{T-1} \kappa(|j|/s_T) \to 0$; but this follows immediately from our condition (i).

Now noting that $T^{-1}\sum_{t=j+1}^{T} V_t \xrightarrow{P} 0$ from a Weak Law of Large Numbers under Assumption A, we further need to show that $T^{-1}\sum_{t=j+1}^{T} V_t \frac{\partial}{\partial\theta} V_{t-j} \xrightarrow{P} 0$. But this follows from a Weak Law of Large Numbers for the cross-correlation of the series V_t to the series $\frac{\partial}{\partial\theta} V_{t-j}$ under Assumption C and our assumption $EV_t \frac{\partial}{\partial\theta} V_{t-j} = 0$.□

Proof of Theorem 3.1. In view of eq. (10), the proof amounts to bounding the bias of $\hat{F}_{jk}$ under the different weak dependence conditions. Note that $E\hat{\Gamma}_{jk}(m) = (1 - \frac{|m|}{T})\Gamma_{jk}(m)$. Thus, we have

$$Bias(\hat{F}_{jk}) \equiv E\hat{F}_{jk} - F_{jk} = A_1 + A_2 + A_3$$

where

$$A_1 = \frac{1}{2\pi} \sum_{m=-T+1}^{T-1} \left(\lambda_{g,c}(\frac{m}{S_{jk}}) - 1 \right) \Gamma_{jk}(m) e^{-imw}$$

$$A_2 = -\frac{1}{2\pi T} \sum_{m=-T+1}^{T-1} |m| \lambda_{g,c}(\frac{m}{S_{jk}}) \Gamma_{jk}(m) e^{-imw}$$

$$A_3 = -\frac{1}{2\pi} \sum_{|m| \geq T} \Gamma_{jk}(m) e^{-imw}.$$

But $|A_3| \leq \frac{1}{2\pi} \sum_{|m|\geq T} |\Gamma_{jk}(m)| \leq \frac{1}{2\pi T} \sum_{|m|\geq T} |m||\Gamma_{jk}(m)| = o(1/T)$, since under any of the three conditions (i), (ii) or (iii) we have $\sum_m |m||\Gamma_{jk}(m)| < \infty$.

Similarly, $|A_2| = O(1/T)$, using the fact that $|\lambda_{g,c}(\frac{m}{S_{jk}})| \leq 1$.

Now note that $A_1 = a_1 + a_2$, where

$$a_1 = \frac{1}{2\pi} \sum_{|m| \leq cS_{jk}} \left(\lambda_{g,c}(\frac{m}{S_{jk}}) - 1 \right) \Gamma_{jk}(m) e^{-imw}$$

$$a_2 = \frac{1}{2\pi} \sum_{cS_{jk} < |m| \leq T} \left(\lambda_{g,c}(\frac{m}{S_{jk}}) - 1 \right) \Gamma_{jk}(m) e^{-imw}$$

First observe that $a_1 = 0$, because $\lambda_{g,c}(\frac{m}{S_{jk}}) = 1$ for $|m| \leq cS_{jk}$. Now

$$|a_2| \leq \frac{1}{\pi} \sum_{cS_{jk}<m\leq T} |\lambda_{g,c}(\frac{m}{S_{jk}}) - 1||\Gamma_{jk}(m)| \leq \frac{1}{\pi} \sum_{cS_{jk}<m\leq T} 2|\Gamma_{jk}(m)| \tag{29}$$

But under the condition of part (i), we have:

$$|a_2| \leq \frac{1}{\pi} \sum_{cS_{jk}<m\leq T} 2\frac{m^r}{c^r S_{jk}^r}|\Gamma_{jk}(m)| \quad \text{i.e.} \quad Bias(\hat{F}_{jk}) = O(1/S_{jk}^r) + O(1/T) = O(1/S_{jk}^r).$$

Under the condition of part (ii), eq. (29) gives

$$|a_2| \leq \frac{2C}{\pi} \sum_{cS_{jk}<m\leq T} e^{-am},$$

i.e., $Bias(\hat{F}_{jk}) = O(e^{-acS_{jk}}) + O(1/T) = O(1/T)$.

Finally, under the condition of part (iii), we have $a_2 = 0$, i.e., $Bias(\hat{F}_{jk}) = O(1/T)$, and the theorem is proven. $\square$

For the proof of Theorem 4.1, we will need the following auxiliary lemma.

Lemma 8.1 *Eq. (16), together with the assumption $\Gamma_{jj}(0) > 0$ for all j, implies that*

$$E\left|\sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\right|^{1+\delta} = O(1/T^{\alpha(1+\delta)}) \quad \textit{for all} \ \ j,k. \tag{30}$$

Proof of Lemma 8.1. Let $\Delta = 1 + \delta$, and note that:

$$E\left|\sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\right|^{\Delta} =$$

$$= E\left|\sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{jj}(0)\hat{\Gamma}_{kk}(0)} + \sqrt{\Gamma_{jj}(0)\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\right|^{\Delta}$$

$$= E\left|\sqrt{\hat{\Gamma}_{kk}(0)}(\sqrt{\hat{\Gamma}_{jj}(0)} - \sqrt{\Gamma_{jj}(0)}) + \sqrt{\Gamma_{jj}(0)}(\sqrt{\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{kk}(0)})\right|^{\Delta} \leq c_1 A_1 + c_2 A_2$$

where c_1, c_2 are some positive constants. In the above, the simple inequality $(a+b)^{\Delta} \leq 2^{\Delta}\max(a,b)^{\Delta} \leq 2^{\Delta}(a^{\Delta} + b^{\Delta})$ for $a, b \geq 0$ is used, and

$$A_1 = E\sqrt{\hat{\Gamma}_{kk}(0)^{\Delta}} \ |\sqrt{\hat{\Gamma}_{jj}(0)} - \sqrt{\Gamma_{jj}(0)}|^{\Delta} \quad \text{and} \quad A_2 = \sqrt{\Gamma_{jj}(0)^{\Delta}} \ E|\sqrt{\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{kk}(0)}|^{\Delta}.$$

But $\left(\sqrt{\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{kk}(0)}\right)^{\Delta}\left(\sqrt{\hat{\Gamma}_{kk}(0)} + \sqrt{\Gamma_{kk}(0)}\right)^{\Delta} = \left(\hat{\Gamma}_{kk}(0) - \Gamma_{kk}(0)\right)^{\Delta}$, hence

$$E|\sqrt{\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{kk}(0)}|^{\Delta} = E\frac{|\hat{\Gamma}_{kk}(0) - \Gamma_{kk}(0)|^{\Delta}}{\left(\sqrt{\hat{\Gamma}_{kk}(0)} + \sqrt{\Gamma_{kk}(0)}\right)^{\Delta}} \leq E\frac{|\hat{\Gamma}_{kk}(0) - \Gamma_{kk}(0)|^{\Delta}}{\sqrt{\Gamma_{kk}(0)^{\Delta}}} = O(1/T^{\alpha\Delta}) \tag{31}$$

by eq. (16). Therefore, $A_2 = O(1/T^{\alpha\Delta})$.

Note that inequality (31) holds for all k; hence, it follows that

$$A_1 = E|\sqrt{\hat{\Gamma}_{jj}(0)} - \sqrt{\Gamma_{jj}(0)}|^{\Delta}|\sqrt{\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{kk}(0)}|^{\Delta} + O(1/T^{\alpha\Delta}).$$

Finally, observe that the function $h(x) = \sqrt{1-x} - (1-\sqrt{x})$ is nonnegative for all $x \in [0,1]$. Therefore, for any $a \geq b > 0$, we have: $\sqrt{a} - \sqrt{b} = |\sqrt{a} - \sqrt{b}| \leq \sqrt{a-b} = \sqrt{|a-b|}$.

Using the above, it follows that

$$E|\sqrt{\hat{\Gamma}_{jj}(0)} - \sqrt{\Gamma_{jj}(0)}|^{\Delta}|\sqrt{\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{kk}(0)}|^{\Delta} \leq E\sqrt{|\hat{\Gamma}_{jj}(0) - \Gamma_{jj}(0)|^{\Delta}}\sqrt{|\hat{\Gamma}_{kk}(0) - \Gamma_{kk}(0)|^{\Delta}}$$

$$\leq \sqrt{E|\hat{\Gamma}_{jj}(0) - \Gamma_{jj}(0)|^{\Delta}E|\hat{\Gamma}_{kk}(0) - \Gamma_{kk}(0)|^{\Delta}} = O(1/T^{\alpha\Delta}),$$

the second inequality being the Cauchy-Schwarz, and the last claim due to eq. (16). Hence, $A_1 = O(1/T^{\alpha\Delta})$ as well, and the lemma is proven.□.

Proof of Theorem 4.1. Note that (15) follows by eq. (16) using Jensen's and Markov's inequality. Now by (15) we have:

$$\hat{F}_{jk} = \sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)}\hat{f}_{jk} = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\hat{f}_{jk} + O_P(1/T^{\alpha}). \tag{32}$$

Let

$$W_T = \hat{F}_{jk} - \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\hat{f}_{jk} = \left(\sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\right)\hat{f}_{jk} = O_P(1/T^{\alpha}).$$

Focusing on integrability of W_T, note that

$$E|W_T|^{\Delta} \leq \max|\hat{f}_{jk}|^{\Delta}E\left|\sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\right|^{\Delta}.$$

But

$$|\hat{f}_{jk}| \leq \frac{1}{2\pi}\sum_{m=-T}^{T}|\lambda_{g,c}(m/S_{jk})||\hat{\rho}_{jk}(m)||e^{-imw}| \leq \frac{1}{2\pi}\sum_{m=-T}^{T}|\lambda_{g,c}(m/S_{jk})| = O(S_{jk})$$

by assumption (17). Hence, $\max|\hat{f}_{jk}|^\Delta = O(S_{jk}^\Delta)$. Therefore, by eq. (30) we have:

$$E|W_T|^\Delta = O(S_{jk}^\Delta/T^{\alpha\Delta}). \tag{33}$$

Proof of (i) and (ii). Recall that $T^\alpha W_T = O_P(1)$ by eq. (32). Since $S_{jk} \to \infty$, it follows that $\frac{T^\alpha}{S_{jk}}W_T = o_P(1)$. But then eq. (33) implies that the sequence $\frac{T^\alpha}{S_{jk}}W_T$ is uniformly integrable; hence

$$E\frac{T^\alpha}{S_{jk}}W_T = o(1) \quad \text{i.e.,} \quad EW_T = o(S_{jk}/T^\alpha),$$

and therefore

$$E\hat{F}_{jk} = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}E\hat{f}_{jk} + o(S_{jk}/T^\alpha).$$

However, $F_{jk} = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}f_{jk}$; hence,

$$Bias(\hat{F}_{jk}) = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\ Bias(\hat{f}_{jk}) + o(S_{jk}/T^\alpha). \tag{34}$$

But from part (i) of Theorem 3.1 we have: $Bias(\hat{F}_{jk}) = O(1/S_{jk}^r)$; it follows that

$$Bias(\hat{f}_{jk}) = O(1/S_{jk}^r) + o(S_{jk}/T^\alpha). \tag{35}$$

Recall that $Var(\hat{f}_{jk}) = O(S_{jk}/T)$ by eq. (14). Note that the second term in $Bias(\hat{f}_{jk})$ is of bigger order than the standard deviation of $\hat{f}_{jk}$ since $\alpha \leq 1/2 \leq (r+1)/(2r+1)$.

Hence, minimization of the order of magnitude of the Mean Squared Error of $\hat{f}_{jk}$ gives the stated optimal choice for the bandwidth S_{jk} in part (i) of Theorem 4.1, and the resulting rate of convergence of $\hat{f}_{jk}$ as given in eq. (18). Finally, note that the $O_P(1/T^\alpha)$ term in eq. (32) is negligible compared to the accuracy of $\hat{f}_{jk}$ as given in (18). Thus, eq. (32) together with (18) implies (19), and part (i) is proven.

To prove part (ii), recall that from part (ii) of Theorem 3.1 we have $Bias(\hat{F}_{jk}) = O(1/T)$. Plugging the optimal bandwidth $S_{jk} = A\log T$ in eq. (34) we obtain:

$$Bias(\hat{f}_{jk}) = O(1/T) + o(\log T/T^\alpha) = O(\log T/T^\alpha). \tag{36}$$

Recall that $Var(\hat{f}_{jk}) = O(\log T/T)$ by eq. (14). Hence, minimization of the order of magnitude of the Mean Squared Error of $\hat{f}_{jk}$ gives the stated rate of convergence of $\hat{f}_{jk}$. By eq. (32), $\hat{F}_{jk}$ has the same rate of convergence as $\hat{f}_{jk}$, and part (ii) is proven.

Proof of (iii). Note that $\frac{T^\alpha}{\log\log T} W_T = o_P(1)$. Also note that S_{jk} is constant under the premises of part (iii). Thus, eq. (33) implies $E|T^\alpha W_T|^\Delta = O(1)$, and thus the sequence $\frac{T^\alpha}{\log\log T} W_T$ is uniformly integrable; hence

$$E\frac{T^\alpha}{\log\log T} W_T = o(1) \quad \text{i.e.,} \quad EW_T = o(\log\log T/T^\alpha),$$

and therefore

$$E\hat{F}_{jk} = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}E\hat{f}_{jk} + o(\log\log T/T^\alpha).$$

However, $F_{jk} = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)} f_{jk}$; hence,

$$Bias(\hat{F}_{jk}) = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\ Bias(\hat{f}_{jk}) + o(\log\log T/T^\alpha).$$

But from part (iii) of Theorem 3.1 we have: $Bias(\hat{F}_{jk}) = O(1/T)$; it follows that

$$Bias(\hat{f}_{jk}) = O(1/T) + o(\log\log T/T^\alpha) = O(\log\log T/T^\alpha). \tag{37}$$

Recalling that $Var(\hat{f}_{jk}) = O(1/T)$ by eq. (14), gives the stated rate of convergence for $\hat{f}_{jk}$ which—by eq. (32)—is the same as that of $\hat{F}_{jk}$, and part (iii) of the theorem is proven.□

Proof of Theorem 5.1. The condition $\hat{F} = F + O_P(1/R_T)$ implies

$$\hat{\Lambda} = \Lambda + O_P(1/R_T), \quad \text{and hence} \quad \hat{\lambda}_j = \lambda_j + O_P(1/R_T) \quad \text{for all} \quad j; \tag{38}$$

see e.g. Theorems 3.2 and 4.2 (and the discussion afterwards) of Eaton and Tyler (1991). But, viewed as an estimator of the nonnegative λ_j, $\hat{\lambda}_j^+$ is a better (or, at least, not worse) estimator than $\hat{\lambda}_j$ in the sense that $|\hat{\lambda}_j^+ - \lambda_j| \le |\hat{\lambda}_j - \lambda_j|$ always. Hence, it follows that

$$\hat{\lambda}_j^+ = \lambda_j + O_P(1/R_T) \quad \text{for all} \quad j, \quad \text{and hence} \quad \hat{\Lambda}^+ = \Lambda + O_P(1/R_T). \tag{39}$$

Using eq. (38) and (39) we have the following:

$$F + O_P(1/R_T) = \hat{F} = \hat{U}\hat{\Lambda}\hat{U}^* = \hat{U}\left(\Lambda + O_P(1/R_T)\right)\hat{U}^*$$
$$= \hat{U}\left(\Lambda^+ + O_P(1/R_T)\right)\hat{U}^* = \hat{F}^+ + O_P(1/R_T),$$

the latter since $\hat{U} = U + o_P(1) = O_P(1)$; solving for $\hat{F}^+$ in the above, the theorem is proven.□

Proof of Theorem 7.1. The proof is analogous to the proof of Theorem 2.3 of Politis (2003) and is omitted.□

References

[1] Andrews, D. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation, *Econometrica*, **59**, 817-858.

[2] Andrews, D. and Monahan, J. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator, *Econometrica*, **60**, 953-966.

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

[4] Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods, 2nd ed.*, Springer, New York.

[5] Eaton, M.E. and Tyler, D.E. (1991). On Weilandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix, *Annals Statist.*, **19**, No. 1, 260-271.

[6] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events.* Springer, Berlin.

[7] Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density, *Theory of Prob. and its Applications*, vol. 14, 153-158.

[8] Gallant, A.R. (1987). *Nonlinear Statistical Models*, John Wiley, New York.

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

[10] Hamilton, J.D. (1994). *Time Series Analysis.* Princeton University Press, Princeton.

[11] Hannan, E.J. (1970), *Multiple Time Series*, John Wiley, New York.

[12] Hansen, B.E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes, *Econometrica*, **60**, 967-972.

[13] Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica*, **50**, 1029-1054

[14] Leadbetter, M.R., Lindgren, G., and Rootzen, H. (1983), *Extremes and related properties of random sequences and processes*, Springer-Verlag, New York.

[15] McMurry, T. and Politis, D.N. (2004). Nonparametric regression with infinite order flat-top kernels, *J. Nonparam. Statist.*, vol. 16, no. 3-4, 549–562.

[16] Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55**, 703-708.

[17] Newey, W. and West, K. (1994). Automatic lag selection in covariance matrix estimation, *Rev. Econ. Studies*, **61**, 631-653.

[18] Parzen, E. (1961), Mathematical Considerations in the Estimation of Spectra, *Technometrics*, vol. 3, 167-190.

[19] Politis, D.N. (2001). On nonparametric function estimation with infinite-order flat-top kernels, in *Probability and Statistical Models with applications,* Ch. Charalambides et al. (Eds.), Chapman and Hall/CRC: Boca Raton, pp. 469-483.

[20] Politis, D.N. (2003). Adaptive bandwidth choice, *J. Nonparam. Statist.,* vol. 15, no. 4-5, 517-533.

[21] Politis, D.N., and Romano, J.P. (1995), Bias-corrected nonparametric spectral estimation. *J. Time Ser. Anal.*, **16**, 67–103.

[22] Politis, D.N., and White, H. (2004). Automatic block-length selection for the dependent bootstrap, *Econometric Reviews,* vol. 23, no. 1, pp. 53-70.

[23] Priestley, M.B. (1962), Basic considerations in the estimation of spectra, *Technometrics*, vol. 4, 551-564.

[24] Robinson, P. (1991). Automatic frequency domain inference on semiparametric and nonparametric models, *Econometrica*, vol. 59, 1329-1363.

[25] Romano, J.P. and Thombs, L. (1996). Inference for autocorrelations under weak assumption, *J. Amer. Statist. Assoc.*, 91, 590-600.