

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

dpolit@ucsd.edu

April 2007

Abstract

A new class of large-sample covariance and spectral density matrix estimators is proposed based on the notion of flat-top kernels. The new estimators are shown to be higher-order accurate when higher-order accuracy is possible. A discussion on kernel choice is presented as well as a supporting finite-sample simulation. The problem of spectral estimation under a potential lack of finite fourth moments is also addressed. 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 flat-top estimator can be modified to become positive semi-definite (even strictly positive definite) while maintaining its higher-order accuracy. In addition, an easy (and consistent) procedure for optimal bandwidth choice is given; this procedure estimates the optimal bandwidth associated with each individual element of the target matrix, automatically sensing (and adapting to) the underlying correlation structure.

*Research partially supported by NSF grant SES-04-18136 funded jointly by the Economics and Statistics Divisions of NSF. Many thanks are due to Arthur Berg for numerous helpful interventions, to Peter Robinson for a critical reading and suggestions of some key early references, and to Dimitrios Gatzouras for his help with the proof of Lemma 9.1. The S+ software for the practical computation of the different spectral density estimators was compiled with the invaluable help of Isheeta Nargis and Arif Dowla of www.stochasticlogic.com, and is now publicly available from: www.math.ucsd.edu/~dpolitis/SOFT/SfunctionsFLAT-TOPS.html.

HIGHER-ORDER ACCURATE, POSITIVE SEMI-DEFINITE ESTIMATION OF LARGE-SAMPLE COVARIANCE AND SPECTRAL DENSITY MATRICES

1 Introduction

Many applications of multivariate time series analysis involve the nonparametric estimation of spectral density matrices. For example, the large-sample covariance matrix of the sample mean of a stationary sequence equals 2π times its spectral density matrix evaluated at the origin. Pioneering work in multivariate spectral estimation was conducted by E.J. Hannan, E. Parzen, M. Rosenblatt, D. Brillinger, and other prominent statistical researchers in the 1950s and 60s. See, e.g., the papers by Hannan (1957, 1958), Parzen (1957, 1961), Priestley (1962), Brillinger and Rosenblatt (1967), as well as the book-length treatments in Hannan (1970), Brillinger (1981), Priestley (1981), and Rosenblatt (1985) that contain a number of additional references.

The subject was revived more recently in the time series econometrics literature where typical applications—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 was put forth in influential papers by Newey and West (1987) and Andrews (1991); see also 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 often suboptimal as their rate of convergence is $T^{2/5}$ even in situations when higher-order accuracy is possible, e.g., a rate closer to $T^{1/2}$; see Samarov (1977).
- (b) The problem of optimal bandwidth choice for the HAC estimators has not been conclusively 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 paper we address the above three issues. A new class of HAC covariance matrix and spectral density matrix estimators is proposed based on the notion of a *flat-top* kernel defined in Politis (2001) that is a generalization of the trapezoidal kernels of Politis and Romano (1995). The new estimators are shown to be higher-order accurate when higher-order accuracy is possible; a discussion on kernel choice is presented as well as a supporting finite-sample simulation.

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 flat-top estimator can be modified to become positive semi-definite (even strictly positive 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, automatically sensing (and adapting to) the underlying correlation structure.

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 framework of higher-order accurate, positive semi-definite estimation of spectral density matrices.

The structure of the paper is as follows. In the next section, the flat-top estimators of a spectral density matrix are defined, and a general theorem on their asymptotic accuracy is given. Section 3 addresses the difficult problem of spectral estimation under a potential lack of finite fourth moments; surprisingly, it is shown that the flat-top estimators retain—for the most part—their higher-order accuracy. Section 5 introduces a modification of the flat-top matrix estimators that results into an estimator that is positive semi-definite (even positive definite—if so desired) *without* affecting the estimators' higher-order accuracy.

Section 6 discusses some interesting kernels of the flat-top family, while Section 7 is devoted to the issue of data-dependent bandwidth choice. An empirical rule for choosing the bandwidth of a flat-top kernel is given extending the trapezoidal kernel bandwidth choice of Politis (2003); a general asymptotic theorem shows the bandwidth choice rule works by automatically adapting to the underlying (unknown) correlation structure. Section 7 presents some finite-sample simulations complementing our asymptotic results where the high accuracy and rate of convergence of the flat-top estimators are manifested in practice. Finally,

Appendix A addresses in detail the set-up of large-sample HAC covariance matrix estimation that is of interest in econometric applications; Lemma 8.1, in particular, is important as it allows the large-sample covariance matrix estimation to be cast in the framework of spectral density matrix estimation. All technical proofs are placed in Appendix B.

2 Spectral density matrix estimation

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

$$\Gamma(j) = EV_t V_{t+j}' \text{ for } j \geq 0, \text{ and } \Gamma(j) = \Gamma(-j)' \text{ for } j < 0. \quad (1)$$

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

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

where $i = \sqrt{-1}$. The $d \times d$ 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 is defined as

$$\Omega = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{k=1}^T \sum_{j=1}^T EV_k V_j'. \quad (3)$$

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

$$\lambda_{g,c}(x) = \begin{cases} 1 & \text{if } |x| \leq c \\ g(x) & \text{else;} \end{cases} \quad (4)$$

here $c > 0$ is a parameter, and $g : \mathbb{R} \rightarrow [-1, 1]$ 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 $|x| \geq \text{some } x_0$, 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

$$\lambda_{TR,c}(x) = \begin{cases} 1 & \text{if } |x| \leq c \\ \frac{|x|-1}{c-1} & \text{if } c < |x| \leq 1 \\ 0 & \text{else} \end{cases} \quad (5)$$

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 ‘truncated’ kernel defined as $\kappa_{trunc}(x) = 1$ if $|x| \leq 1$ and $\kappa_{trunc}(x) = 0$ else, and the well-known triangular Bartlett kernel $\kappa_B(x) = (1 - |x|)^+$. As a matter of fact, $\lambda_{TR,c}(x)$ reduces to $\kappa_{trunc}(x)$ and/or $\kappa_B(x)$ by letting c tend to 1 or 0 respectively. Here, and throughout the paper, the notation $(y)^+$ indicates the positive part of y , i.e., $(y)^+ = \max(y, 0)$.

Let S be a $d \times d$ 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:

$$\hat{F}_{jk} = \frac{1}{2\pi} \sum_{m=-T}^T \lambda_{g,c}(m/S_{jk}) \hat{\Gamma}_{jk}(m) e^{-imw} \quad (6)$$

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 as

$$\hat{\Gamma}(j) = \frac{1}{T} \sum_{t=1}^{T-j} V_t V'_{t+j} \quad \text{for } j \geq 0; \quad \hat{\Gamma}(j) = \hat{\Gamma}(-j)' \quad \text{for } j < 0, \quad (7)$$

and $\hat{\Gamma}(j) = 0$ for $|j| \geq T$. 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 2.1 *Assume conditions strong enough to ensure that¹*

$$\text{Var}(\hat{F}_{jk}) = O(S_{jk}/T) \quad \text{for any fixed } j, k; \quad (8)$$

¹There exist different sets of conditions sufficient for eq. (8). 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).

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

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

as usual, the notation $A \sim B$ means $A/B \rightarrow 1$.

(iii) If $\Gamma_{jk}(m) = 0$ for $|m| > \text{some } q$, then letting $S_{jk} = \max(\lceil q/c \rceil, 1)$, yields²

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

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 2.1 are usual conditions of weak dependence. For example, if $\Gamma_{jj}(m) = 0$ for $|m| > \text{some } q$, then the j th 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 2.1 demonstrates the improvement in rate of convergence afforded by the use of flat-top kernels as compared to the $O_P(T^{-2/5})$ error associated with traditional second-order kernels. Most importantly, flat-top kernels are seen to *attain* the lower bounds for the order of magnitude of the error of a quadratic spectral density estimator under the three cases of Theorem 2.1; these lower bounds are due to Samarov (1977).

Note that Theorem 2.1 not only gives the rate of convergence of \hat{F}_{jk} to F_{jk} , but at the same time it suggests the optimal values of the bandwidth parameter S_{jk} ; here optimality is meant with respect to 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

²Taking the maximum of $\lceil q/c \rceil$ and 1 is done to cover the possibility that $q = 0$.

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

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

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. (29). 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 2.1 (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 2.1 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 $T^{2/5}$.

Remark 2.1 If a chosen bandwidth happens not to be small as compared to the sample size, then the standard asymptotics—such as eq. (8)—might not provide accurate approximations, and the so-called “fixed- b ” asymptotics of Kiefer and Vogelsang (2002), and Hashimzade and Vogelsang (2004) are a valuable alternative. There is no inherent discrepancy between the notion of flat-top kernels and “fixed- b ” asymptotics.³ Indeed, the latter may very well be used in connection with flat-top kernels but it seems that the improvements will be marginal (if at all). The reason for this is that for small bandwidths,

³The “steep-origin” kernels of Phillips, Sun, and Jin (2004) are competitors to the “fixed- b ” asymptotics but the underlying idea is the same, i.e., better approximations when the bandwidth happens to be large; here the kernel is raised to a power instead of being re-scaled by the bandwidth parameter. Note though that a flat-top kernel raised to a power can never become of “steep-origin” as it remains a flat-top; thus, the implied re-scaling will be unsuccessful, and flat-top kernels can not be used in this connection.

the “fixed- b ” asymptotics coincide with the traditional approximations, and that flat-top kernels are characterized by ultra-small optimal bandwidths; see e.g. the logarithmic and constant optimal bandwidths in Theorem 2.1 (ii) and (iii).

3 Spectral estimation in the absence of finite fourth moments

As mentioned in the last section, eq. (8) is typically satisfied for kernel estimators such as \hat{F} . Nevertheless, it has been conjectured that some financial time series might not possess finite fourth moments; see e.g. Hall and Yao (2003) or Politis (2004) for a discussion. But 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

$$f(w) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \rho(k) e^{-ikw}; \quad (10)$$

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:

$$\hat{f}_{jk} = \frac{1}{2\pi} \sum_{m=-T}^T \lambda_{g,c}(m/S_{jk}) \hat{\rho}_{jk}(m) e^{-imw}. \quad (11)$$

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. (8) is then suggested:

$$Var(\hat{f}_{jk}) = O(S_{jk}/T) \text{ for any fixed } j, k. \quad (12)$$

Eq. (12) 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:

$$\hat{\Gamma}_{jj}(0) = \Gamma_{jj}(0) + O_P(1/T^\alpha), \quad \text{for all } j, \quad \text{and some } \alpha \in (0, 1/2]. \quad (13)$$

For our purposes we will require the slightly stronger condition:

$$E \left| \hat{\Gamma}_{jj}(0) - \Gamma_{jj}(0) \right|^{1+\delta} = O(1/T^{\alpha(1+\delta)}) \quad \text{for all } j, \quad \text{and some } \delta > 0 \quad \text{and } \alpha \in (0, 1/2]. \quad (14)$$

The following theorem is a generalization of Theorem 2.1 to the setting where finite fourth moments are potentially lacking.

Theorem 3.1 *Fix values for j, k , and assume conditions (12), (14), and that⁴*

$$S_{jk}^{-1} \sum_{j=-T+1}^{T-1} |\lambda_{g,c}(j/S_{jk})| = O(1). \quad (15)$$

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*

$$\hat{f}_{jk} = f_{jk} + O_P(T^{-\alpha r/(r+1)}), \quad (16)$$

and

$$\hat{F}_{jk} = F_{jk} + O_P(T^{-\alpha r/(r+1)}). \quad (17)$$

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

$$\hat{f}_{jk} = f_{jk} + O_P\left(\frac{\log T}{T^\alpha}\right) \quad \text{and} \quad \hat{F}_{jk} = F_{jk} + O_P\left(\frac{\log T}{T^\alpha}\right). \quad (18)$$

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

$$\hat{f}_{jk} = f_{jk} + O_P\left(\frac{\log \log T}{T^\alpha}\right) \quad \text{and} \quad \hat{F}_{jk} = F_{jk} + O_P\left(\frac{\log \log T}{T^\alpha}\right) \quad (19)$$

Note that, even under the potential absence of finite fourth moments, \hat{F} maintains its higher-order accuracy. Parts (ii) and (iii) of Theorem 3.1 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 3.1, the optimal rates for the bandwidth S_{jk} are insensitive to whether fourth moments are finite or not.

⁴As in condition (i) of Lemma 8.1, eq. (15) is easily satisfied such as when $\lambda_{g,c}(x)$ has a hard ‘cut-off’, i.e., $\lambda_{g,c}(x) = 0$ for $|x| > \text{some } x_0$.

4 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 2.1 (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.⁵

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:

$$F = U\Lambda U^* \quad \text{and} \quad \hat{F} = \hat{U}\hat{\Lambda}\hat{U}^* \tag{20}$$

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 $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_d)$, $\hat{\Lambda} = \text{diag}(\hat{\lambda}_1, \dots, \hat{\lambda}_d)$ are diagonal matrices containing the respective eigenvalues.

Noting that the entries of Λ are all nonnegative suggests the following fix to the possible

⁵Note, 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 that have been widely reported; see Figure 1 in Politis and Romano (1995) for a comparative graph of the sidelobes. Because of its discontinuity, the truncated kernel is arguably the *worst* representative of the flat-top family; more details on kernel choice are given in Section 5.

negativity of \hat{F} . Let $\hat{\Lambda}^+ = \text{diag}(\hat{\lambda}_1^+, \dots, \hat{\lambda}_d^+)$ 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

$$\hat{F}^+ = \hat{U} \hat{\Lambda}^+ \hat{U}^*. \quad (21)$$

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 4.1 *Let R_T be a sequence such that $R_T \rightarrow \infty$ as $T \rightarrow \infty$. If $\hat{F} = F + O_P(1/R_T)$, then $\hat{F}^+ = F + O_P(1/R_T)$ as well.⁶*

To take it one step further, it may be the case that the estimand F is not only positive semi-definite but strictly positive definite. Alternatively, it can be of interest to consider the inverse of an estimated F as in the case of estimating a large-sample covariance matrix for the purpose of creating a studentized statistic. Notably, the idea of studentizing a mean-like statistic by a nonparametric spectrum estimate can be traced back to Jowett (1954), Hannan (1957), and Brillinger (1979); see Robinson and Velasco (1997) for a review, and Velasco and Robinson (2001) for some recent developments.

For such applications, it may be desirable to have a *strictly* positive definite estimator of F that maintains the high accuracy of the flat-top estimators. For this reason, let $\epsilon_T > 0$ be some chosen sequence, and define $\hat{\Lambda}^\epsilon = \text{diag}(\hat{\lambda}_1^\epsilon, \dots, \hat{\lambda}_d^\epsilon)$ where $\hat{\lambda}_j^\epsilon = \max(\hat{\lambda}_j, \epsilon_T)$. Also define the strictly positive definite estimator

$$\hat{F}^\epsilon = \hat{U} \hat{\Lambda}^\epsilon \hat{U}^*. \quad (22)$$

The following corollary to Theorem 4.1 shows that \hat{F}^ϵ also inherits the higher-order accuracy of \hat{F} if ϵ_T is chosen right. Thus, \hat{F}^ϵ is a higher-order accurate, strictly positive definite estimator.

Corollary 4.1 *Let R_T be a sequence such that $R_T \rightarrow \infty$ as $T \rightarrow \infty$, and let the strictly positive sequence ϵ_T be $o(1/R_T)$. If $\hat{F} = F + O_P(1/R_T)$, then $\hat{F}^\epsilon = F + O_P(1/R_T)$ as well.*

⁶The notation $A = O_P(1/R_T)$ for some matrix A means that each element of A is $O_P(1/R_T)$.

Note that for spectral estimation problems we always have $1/\sqrt{T} = O(1/R_T)$. So any choice of $\epsilon_T > 0$ satisfying $\epsilon_T = o(1/\sqrt{T})$ will satisfy the requirements of Corollary 4.1. However, in order to avoid the introduction of unnecessary finite-sample bias it is recommended to take $\epsilon_T > 0$ quite smaller than $1/\sqrt{T}$. But then again ϵ_T should not be too small in order not to risk the matrix \hat{F}^ϵ being ill-conditioned which leads to computational difficulties; letting $\epsilon_T = 1/T^a$ with $1 \leq a \leq 2$ seems like a reasonable practical compromise.

Remark 4.1 Some concluding remarks are in order here. Since Theorem 2.1 shows that the flat-top \hat{F} is consistent at a very fast rate, we expect it to be close to its target value F and share its properties (positive definiteness, etc.). This is indeed true, and supported by the finite-sample simulations of Section 7.

To elaborate, if the eigenvalues of the estimand F are relatively large, i.e., not close to zero, then with high probability the eigenvalues of \hat{F} will be positive as well and there is no need for \hat{F}^+ or even \hat{F}^ϵ . On the other hand, if an eigenvalue of F is zero (or close to zero), then the small bias of \hat{F} *demands* that the corresponding eigenvalue of \hat{F} has a distribution that is centered right around zero which therefore generates many negative values (as many as 50%); see Figure 3 (b) for an illustration. However, this is *not* to be seen as a hindrance; rather, it is very informative, giving strong evidence that the target eigenvalue is close to zero, and that consequently taking \hat{F}^+ or \hat{F}^ϵ is most appropriate.

Consider for example the one-dimensional case ($d = 1$), and note that the usual asymptotic normality of kernel estimators clashes with the desire of unbiasedness; this is especially apparent either in small/medium-size samples, or in large samples with a target value near zero. In other words, restricting our attention to just non-negative estimators is tantamount to limiting ourselves to working with severely biased estimators; see e.g. Figure 3 (a).

The thesis of this paper is to not impose the non-negativity restriction at the outset; rather, to work with the most accurate (and less biased) estimators, and fix the possible non-positivities at the end. Because of the high accuracy of the proposed estimators, non-positivities will be observed in practice effectively only when the target value is zero (or close to zero) in which case an estimated value of zero (or positive but close to zero) is right on target.

5 Flat-top kernel choice

The favorable asymptotic rates of Theorems 2.1 and 3.1 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.⁷

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

$$\lambda_{ID,b,c}(x) = \begin{cases} 1 & \text{if } |x| \leq c \\ \exp(-b \exp(-b/(|x| - c)^2)/(|x| - 1)^2) & \text{if } c < |x| < 1 \\ 0 & \text{if } |x| \geq 1 \end{cases} \quad (23)$$

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-

⁷The variance is still of order $O(S_{jk}/T)$ as eq. (8) demands, but the proportionality constant in the term $O(S_{jk}/T)$ is large for the Dirichlet kernel.

top kernel would be defined as:

$$\lambda_{PR,c}(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq c \\ 1 - 6(x-c)^2 + 6|x-c|^3 & \text{if } c \leq x \leq c + 1/2 \\ 2(1 - |x-c|)^3 & \text{if } c + 1/2 < x < c + 1 \\ 0 & \text{if } x \geq c + 1 \\ \lambda_{PR,c}(-x) & \text{if } x < 0. \end{cases} \quad (24)$$

The original Parzen kernel $\kappa_{PR}(x)$ is seen to be equal to $\lambda_{PR,0}(x)$.

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

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

that has been found optimal⁸ among positive semi-definite second order kernels; see e.g. Priestley (1962) or Epanechnikov (1969). The modification would amount to defining:

$$\lambda_{QS,b,c}(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq c \\ \frac{3}{b^2(x-c)^2} \left(\frac{\sin(b(x-c))}{b(x-c)} - \cos(b(x-c)) \right) & \text{if } x > c \\ \lambda_{QS,b,c}(-x) & \text{if } x < 0, \end{cases} \quad (25)$$

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 $c \rightarrow 0$ and $c \rightarrow 1$ 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

⁸Priestley'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).

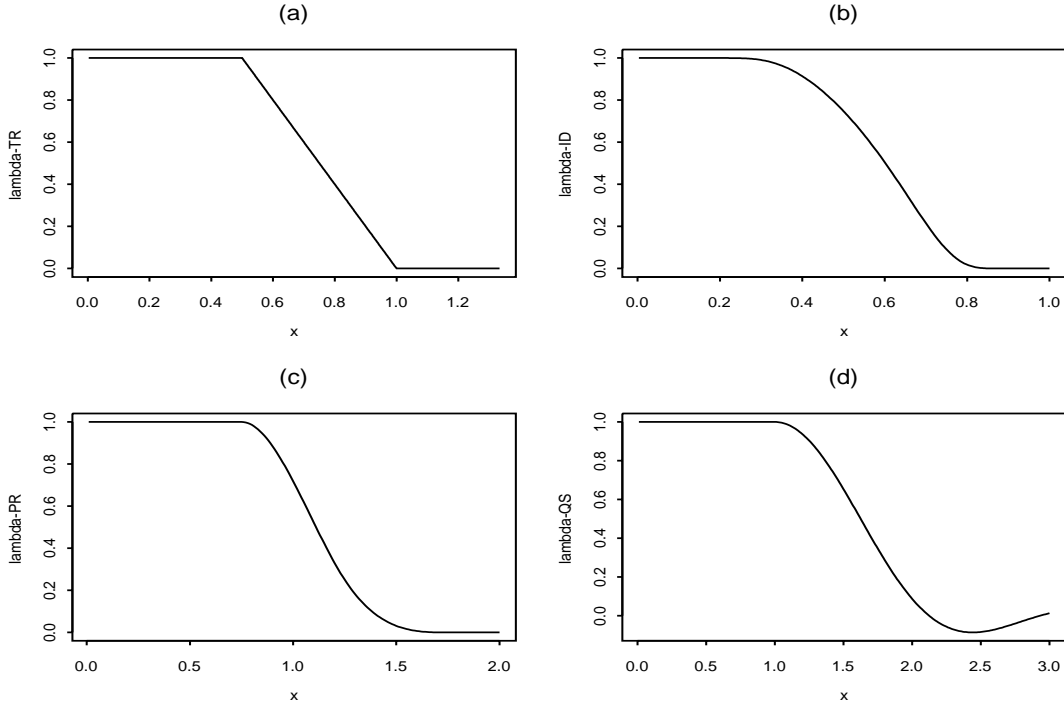


Figure 1: (a) Plot of $\lambda_{TR,1/2}(x)$ vs. $x > 0$; (b) Plot of $\lambda_{ID,0.25,0.05}(x)$ vs. $x > 0$; (c) Plot of $\lambda_{PR,0.75}(x)$ vs. $x > 0$; (d) Plot of $\lambda_{QS,4,1}(x)$ vs. $x > 0$.

will make $\lambda_{ID,b,c}(x)$ look very much like the trapezoidal $\lambda_{TR,1/2}(x)$ with ultra-smoothed 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.

6 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 2.1 and 3.1, 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 2.1 and 3.1 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 2.1 and 3.1 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 $m = 0, 1, \dots, K_T$, 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 $m = 0, 1, \dots, K_T$, 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 $m = 0, 1, \dots, K_T$. 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 univariate case (i.e., $d = 1$ or $j = k$ in the above), the bandwidth choice rule was empirically suggested by Politis and Romano (1995) for the trapezoidal kernel; it was then rigorously studied in Politis (2003) but still only for the trapezoidal kernel. 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 $m = 1, \dots, K_T$ by Bonferroni's inequality.

These approximate confidence intervals are based on a null hypothesis that the series is i.i.d. in which case the large-sample variance of $\rho_{jk}(m)$ is $1/T$ for any m . Nevertheless, in nonlinear/non-normal time series, uncorrelatedness is a weaker assumption than independence. So, a more conservative approach would be to use a resampling and/or subsampling approach—cf. Lahiri (2003) or Politis, Romano and Wolf (1999)—in order to estimate these variances and adjust C_0 appropriately. For example, if the standard deviation of $\sqrt{T}\rho_{jk}(m)$ for some m among the lags under consideration is estimated by ν , then let $C_0 \simeq 2\nu$ and $K_T = \max(5, \sqrt{\log_{10} T})$ as before.

In any case, 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 6.1 Fix j, k , and assume conditions strong enough to ensure that⁹ for all finite N ,

$$\max_{m=1, \dots, N} |\hat{\rho}_{jk}(n+m) - \rho_{jk}(n+m)| = O_P(1/\sqrt{T}) \quad (26)$$

uniformly in n , and

$$\max_{m=0, 1, \dots, T-1} |\hat{\rho}_{jk}(m) - \rho_{jk}(m)| = O_P\left(\sqrt{\frac{\log T}{T}}\right). \quad (27)$$

Also assume that the sequence $\rho_{jk}(m)$ does not have more than $K_T - 1$ consecutive zeros¹⁰ in its first m_0 lags (i.e., for $m = 0, 1, \dots, m_0$).

(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,

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

for some positive constant A_1 ; the notation $A \stackrel{P}{\sim} B$ 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,

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

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

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

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

⁹There exist different sets of conditions sufficient for eq. (26); 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. (27) would follow from the theory of extremes of dependent sequences—see e.g. Leadbetter et al. (1983).

¹⁰Because 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 $K_T = \max(5, \sqrt{\log_{10} T})$.

Comparing the empirical rule \hat{S}_{jk} to the theoretically optimal values of S_{jk} given in Theorem 2.1 we see that \hat{S}_{jk} manages to capture exactly the theoretically optimal rate in cases (ii) and (iii) of Theorem 6.1. In case (i) of Theorem 6.1, \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 2.1 but the difference is small, and becomes even smaller for large p .

Thus, \hat{S}_{jk} is seen to *adapt* to the underlying rate of decay of the correlation and cross-correlation functions, automatically switching between the polynomial, logarithmic, and constant rates that are optimal respectively in the three cases of Theorems 2.1 and 3.1.

7 Some finite-sample simulations

We now present some finite-sample simulations to complement our asymptotic results. The simulations are not meant to be exhaustive; rather, their goal is to illustrate the main issues discussed in the paper. We will focus on estimating $F(w)$ with $w = 0$ for bivariate series ($d = 2$) generated by two simple ARMA models; we will consider the usual ‘traditional’ estimators and compare them to the proposed flat-top kernels.

Throughout this section, the bandwidths of the ‘traditional’ kernels κ_B (Bartlett), κ_{PR} (Parzen), and κ_{QS} (optimal 2nd order kernel) were estimated using equations (6.2) and (6.4) of Andrews (1991), i.e., the notion of estimating the bandwidth constants by fitting an AR(1) model. By contrast, the bandwidths of all flat-top kernels were estimated using our empirical rule of Section 6. For the truncated kernel κ_{trunc} both bandwidth choices, i.e., the Andrews bandwidth—see footnote 5 in Andrews (1991, p. 834)—and our empirical rule, were used and are denoted by Truncated-A and Truncated-E respectively.

For the simulation, $B = 999$ bivariate time series stretches, each of length T , were generated using the two models below.

MODEL I: $V_t^{(1)} = 0.75V_{t-1}^{(1)} + Z_t^{(1)}$, and $V_t^{(2)} = 2(Z_t^{(2)} + Z_{t-1}^{(2)})$ where $V_t^{(k)}$ denotes the k th coordinate series of the bivariate series $\{V_t\}$.

MODEL II: $V_t^{(1)} = Z_t^{(1)} - Z_{t-1}^{(1)}$, and $V_t^{(2)} = W_t + V_{t+7}^{(1)}$ where $W_t = -0.75W_{t-1} + Z_t^{(2)}$.

In all the above, the error series $\{Z_t^{(1)}\}$ and $\{Z_t^{(2)}\}$ are i.i.d. standard normal and independent to each other.

Model I involves two coordinates independent to each other, an AR(1) and a MA(1), both exhibiting positive dependence. The independence of the two coordinates implies that $F_{12}(w) = 0$ for all w which in turn implies that the optimal value of S_{12} is as small as possible, i.e, one; the other target values are $F_{11}(0) = 8/\pi = F_{22}(0)$.

Table 1a shows the empirically found Mean Squared Errors (MSE) of different estimators relative to (i.e., divided by) the MSE of the optimal second order estimator with kernel κ_{QS} ; the data followed Model I with $T = 100$. It is apparent that in the case of F_{11} all traditional kernels (Bartlett, Parzen and the optimal κ_{QS}) do quite well and outperform the recommended flat-top kernels of Figure 1. However, this seems to be due to the fact that we are using an AR(1) formula for the bandwidths of traditional kernels and an AR(1) model happens to be correct in this case. That the bandwidth is the most prominent issue here is manifested by comparing the truncated kernel with AR(1) bandwidth (Truncated-A) to the one with bandwidth estimated by our empirical rule (Truncated-E). In fact, Truncated-A seems to be the overall *best* estimator of the AR(1) spectrum F_{11} with strong positive dependence present; see also Table II of Andrews (1991).

The situation is reversed in the estimation of F_{12} and F_{22} . Here, the problematic use of the same bandwidth for all coordinates of the target matrix F is apparent as Truncated-E, having coordinate-specific estimated bandwidth, outperforms Truncated-A. In fact, the best estimator of F_{12} and F_{22} appears to be the positive semi-definite estimator \hat{F}^+ corresponding to the truncated kernel with bandwidth matrix estimated by our empirical rule (Truncated-E).

This may not seem surprising since *had we known* that an MA(1) model holds for F_{22} , we would estimate F_{22} by a model-based estimator that would be tantamount to a truncated estimator in this case. However, note that the MA(1) information is *not* used here; rather, our empirical rule is able to sense and automatically adapt to this MA(1) structure, and this is a major success with a sample size as small as 100.

Figure 2 (a) shows a histogram of our empirical rule \hat{S}_{11} for use with the trapezoidal kernel $\lambda_{TR,1/2}$ as computed over the 999 Monte Carlo iterations. The mean of the histogram is about 9 which is right about what we would use had we known that the underlying model is an AR(1). It is the variability in this histogram that inflates the variances of our flat-top estimators with estimated bandwidths.

A histogram of the corresponding \hat{S}_{22} is not very informative as the overwhelming majority (93%) of the computed \hat{S}_{22} were found to equal 2 which corresponds to an MA(1) structure. Figure 2 (b) shows a plot of \hat{S}_{22} as computed over the Monte Carlo iterations that more clearly shows the bandwidth estimation procedure in action.

Note that the entries of Table 1a corresponding to the matrix $\hat{F}+$ are nearly identical to those of \hat{F} indicating a very *low* proportion of \hat{F} matrices that were *not* positive semi-definite even for $T = 100$. The reason for this is twofold: (i) the target values (of those eigenvalues) are relatively large, i.e., not close to zero, and (ii) the bandwidths chosen were appropriate resulting in accurate estimators.

As expected, taking the positive part yields an overall improvement; note though that the improvement is a global one, and not necessarily uniform over all coordinates of \hat{F} ; for example, in most flat-top kernels of Table 1a it seems that \hat{F}_{12}^+ is quite improved as compared to \hat{F}_{11} at the expense of having \hat{F}_{11}^+ just slightly inferior to \hat{F}_{11} .

	\hat{F}_{11}	\hat{F}_{12}	\hat{F}_{22}	\hat{F}_{11}^+	\hat{F}_{12}^+	\hat{F}_{22}^+
κ_B (Bartlett)	1.02	0.64	0.69	n/a	n/a	n/a
κ_{PR} (Parzen)	1.04	1.12	1.09	n/a	n/a	n/a
κ_{trunc} (Truncated-A)	0.97	0.82	0.94	n/a	n/a	n/a
κ_{trunc} (Truncated-E)	1.55	0.70	0.46	1.56	0.64	0.47
$\lambda_{TR,1/2}$ (Trapezoid)	1.74	0.93	0.58	1.75	0.84	0.58
$\lambda_{PR,3/4}$ (Flat-top Parzen)	1.76	1.05	0.72	1.77	0.97	0.72
$\lambda_{QS,4,1}$ (Flat-top Quadratic)	1.78	0.99	0.62	1.80	0.91	0.62
$\lambda_{ID,1/4,0.05}$ (Flat-top Inf. Diff.)	1.82	1.20	0.78	1.83	1.12	0.78

Table 1a. Entries represent the empirical MSEs of different estimators relative to (i.e., divided by) the MSE of the optimal second order estimator with kernel κ_{QS} ; Model I with $T = 100$. [Minimum MSE is indicated by boldface.]

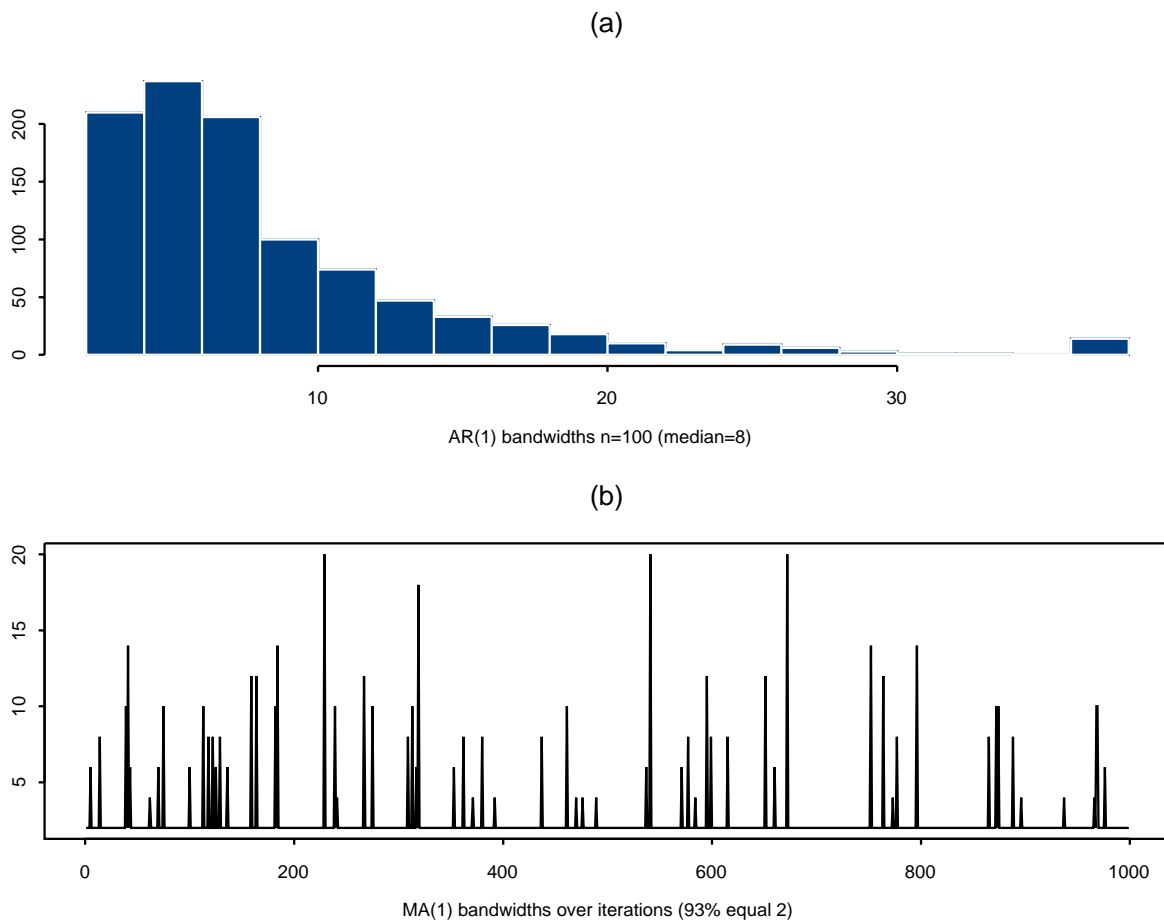


Figure 2: (a) Histogram of \hat{S}_{11} ; (b) Plot of \hat{S}_{22} over the Monte Carlo iterations; Model I with $T=100$.

	\hat{F}_{11} or \hat{F}_{11}^+	\hat{F}_{12} or \hat{F}_{12}^+	\hat{F}_{22} or \hat{F}_{22}^+
κ_B (Bartlett)	1.20	0.79	0.83
κ_{PR} (Parzen)	1.02	1.10	1.10
κ_{trunc} (Truncated-A)	0.95	0.85	0.96
κ_{trunc} (Truncated-E)	1.41	0.27	0.27
$\lambda_{TR,1/2}$ (Trapezoid)	1.62	0.35	0.32
$\lambda_{PR,3/4}$ (Flat-top Parzen)	1.63	0.42	0.43
$\lambda_{QS,4,1}$ (Flat-top Quadratic)	1.68	0.38	0.34
$\lambda_{ID,1/4,0.05}$ (Flat-top Inf. Diff.)	1.77	0.48	0.46

Table 1b. Entries represent the empirical MSEs of different estimators relative to the MSE of the optimal second order estimator with kernel κ_{QS} ; Model I with $T = 500$.

Table 1b is the same as Table 1a with the sample size increased to 500, and the results are qualitatively similar. Most notable here is the approximate halving of the flat-top MSEs of F_{12} and F_{22} going from Table 1a to Table 1b; this lends support to the \sqrt{T} convergence claimed in Theorem 2.1 (iii) in conjunction with Theorem 6.1 (iii).

Interestingly, in Table 1b the MSEs of the flat-top \hat{F}^+ were found *identical* (to 8 decimal points) to those of \hat{F} indicating that there were absolutely *no* occurrences of estimators with negative eigenvalues with the increased sample size; this empirical finding gives credence to the discussion at the end of Section 4. Still, practitioners are urged to use the positive semi-definite variation \hat{F}^+ as a safe-guard.

In order to really see the effect/improvement of \hat{F}^+ vs. \hat{F} , we need to consider a model where the target eigenvalues happen to be close to zero. Model II is characterized by negative (i.e., alternating) dependence which has as its consequence small values for the spectral density at the origin. As a matter of fact, $F_{11}(0)$ is identically zero whereas $F_{22}(0)$ equals 0.052. Coordinate $V_t^{(1)}$ follows an MA(1) model, and $V_t^{(2)}$ follows an ARMA(1,2) model that is—by construction—dependent to coordinate $V_t^{(1)}$ as their cross-correlation is significant at lags around 7. For this reason $F_{12}(w)$ is not identically zero, and the optimal values for S_{12} are not trivial as in Model I; interestingly, however, $F_{12}(0)$ just happens to be zero as well.

Figure 3 (a) shows histograms of the distribution of the Bartlett estimator of F_{11} in the case of Model II with $T=250$; Figure 3 (b) is the same but concerning the trapezoidal $\lambda_{TR,1/2}$

estimator. As expected, the positivity of the Bartlett estimator results into significant bias when the target value is zero. By contrast, the trapezoidal shows minimal bias albeit somewhat larger variance. But even the variance discrepancy is corrected after the positive-part of the trapezoidal estimator is taken strongly suggesting that the flat-top \hat{F}^+ is a superior estimator. A similar phenomenon occurs with a target value near zero as in the estimation of F_{22} that equals 0.052; see Figure 4.

Table 2a shows the empirically found Mean Squared Errors (MSE) of different estimators relative to the MSE of the kernel κ_{QS} with data from Model II with $T = 100$. The first striking feature of Table 2a is that, despite its optimality among second order kernels, kernel κ_{QS} is vastly outperformed by the traditional positive kernels: Bartlett and Parzen. Those in turn are outperformed by any of our four recommended flat-top kernels in their positive semi-definite variation \hat{F}^+ . The (non-recommended) truncated kernel performs rather erratically regardless of bandwidth choice.

As mentioned above, Model II presents a bit of a challenge in estimating S_{12} by our empirical rule and this difficulty is manifested in the results of Table 2a. The reason for this is that whereas $F_{11}(w)$ equals a constant plus a cosine of period 2π , $F_{12}(w)$ involves a cosine of period $2\pi/7$, i.e., it is very ‘wiggly’. Still, the best flat-top performers, the flat-top Parzen and the infinitely differentiable, manage to achieve a MSE that is about a half of that of the reference kernel κ_{QS} . The situation is dramatically improved if the sample size is increased to 500 as Table 2b shows.

Looking at our four flat-top kernels, $\lambda_{TR,1/2}$, $\lambda_{PR,3/4}$, $\lambda_{QS,4,1}$, and $\lambda_{ID,1/4,0.05}$, the improvement offered by the increased sample size of Table 2b is very apparent, and this is in good part due to the bandwidths being chosen by our empirical rule which adapts to the underlying correlation structure. Of course, to realize/maximize those gains, one has to employ the \hat{F}^+ estimators. As conjectured in Section 5, the infinitely differentiable flat-top kernel $\lambda_{ID,1/4,0.05}$ is best overall but with the flat-top Parzen coming in as a (very) close second. Both have impressively low MSEs of the order of 10% as compared to the optimal second order kernel κ_{QS} .

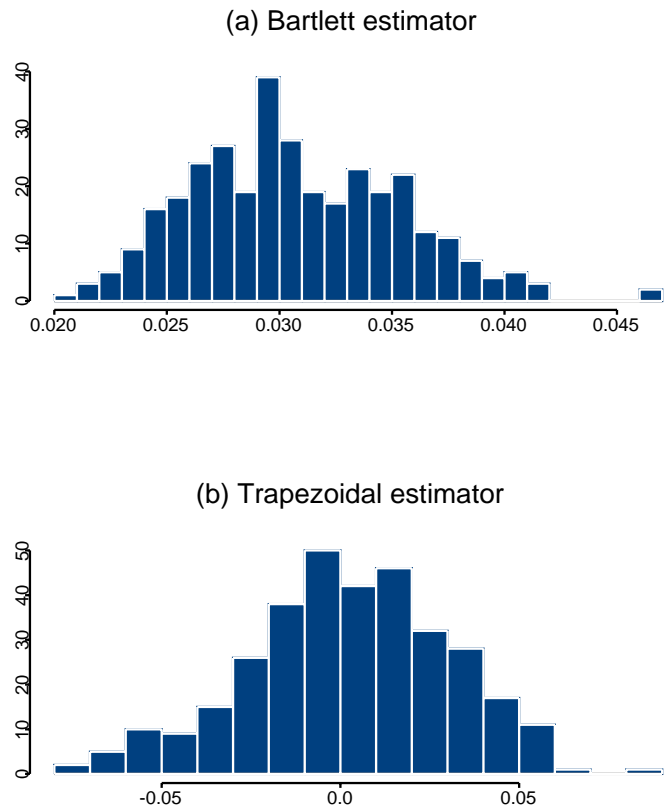


Figure 3: (a) Bartlett estimator of F_{11} ; (b) Trapezoidal estimator of F_{11} ; Model II with $T=250$.

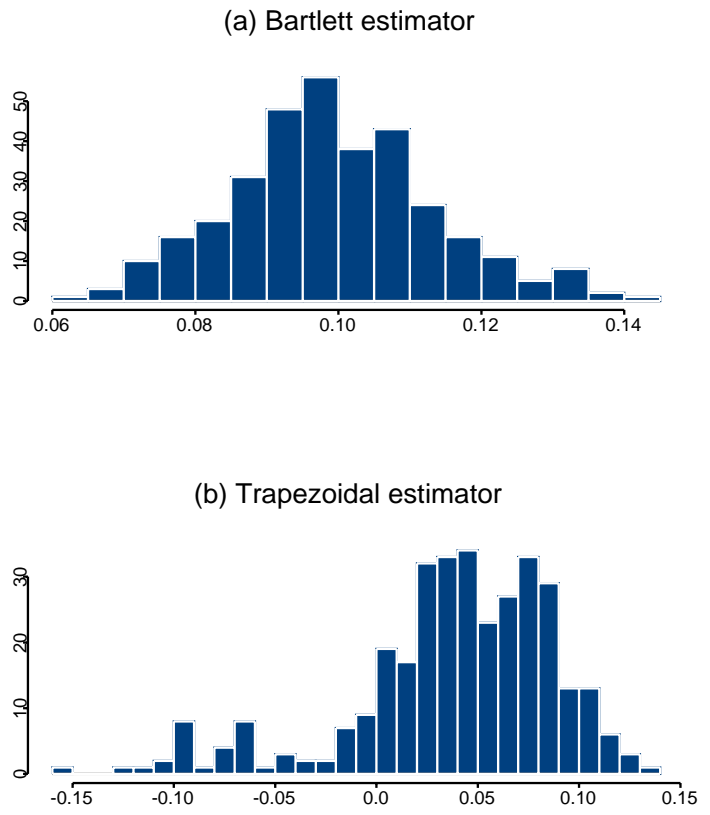


Figure 4: (a) Bartlett estimator of F_{22} ; (b) Trapezoidal estimator of F_{22} ; Model II with $T=250$.

	\hat{F}_{11}	\hat{F}_{12}	\hat{F}_{22}	\hat{F}_{11}^+	\hat{F}_{12}^+	\hat{F}_{22}^+
κ_B (Bartlett)	0.35	0.62	0.68	n/a	n/a	n/a
κ_{PR} (Parzen)	0.81	0.66	0.87	n/a	n/a	n/a
κ_{trunc} (Truncated-A)	0.34	10.9	9.45	n/a	n/a	n/a
κ_{trunc} (Truncated-E)	0.36	19.5	5.12	0.33	6.45	2.38
$\lambda_{TR,1/2}$ (Trapezoid)	0.30	2.97	1.90	0.19	1.09	0.33
$\lambda_{PR,3/4}$ (Flat-top Parzen)	0.11	0.81	0.26	0.07	0.56	0.19
$\lambda_{QS,4,1}$ (Flat-top Quadratic)	0.23	2.25	1.45	0.15	0.90	0.25
$\lambda_{ID,1/4,0.05}$ (Flat-top Inf. Diff.)	0.09	0.89	0.27	0.06	0.62	0.18

Table 2a. Entries represent the empirical MSEs of different estimators relative to the MSE of the optimal second order estimator with kernel κ_{QS} ; Model II with $T = 100$.

	\hat{F}_{11}	\hat{F}_{12}	\hat{F}_{22}	\hat{F}_{11}^+	\hat{F}_{12}^+	\hat{F}_{22}^+
κ_B (Bartlett)	0.37	0.11	0.76	n/a	n/a	n/a
κ_{PR} (Parzen)	0.86	0.30	0.95	n/a	n/a	n/a
κ_{trunc} (Truncated-A)	0.28	11.4	29.0	n/a	n/a	n/a
κ_{trunc} (Truncated-E)	0.30	10.73	5.29	0.19	5.38	3.40
$\lambda_{TR,1/2}$ (Trapezoid)	0.26	0.22	0.33	0.13	0.17	0.29
$\lambda_{PR,3/4}$ (Flat-top Parzen)	0.08	0.16	0.12	0.05	0.14	0.12
$\lambda_{QS,4,1}$ (Flat-top Quadratic)	0.20	0.15	0.16	0.10	0.12	0.14
$\lambda_{ID,1/4,0.05}$ (Flat-top Inf. Diff.)	0.07	0.11	0.11	0.04	0.09	0.11

Table 2b. Entries represent the empirical MSEs of different estimators relative to the MSE of the optimal second order estimator with kernel κ_{QS} ; Model II with $T = 500$.

As a conclusion, note that the notorious truncated kernel gives poor results in Table 2b (Model II) even with the adaptive bandwidth choice, i.e., the Truncated-E version,¹¹ whereas it was the best performer in Table 1 (Model I). It is mixed/incoherent results such as these that turned practitioners away from the truncated kernel early on and made them

¹¹The poor performance of the truncated kernel in an MA(1) case with negative dependence was pointed out by West (1997) who instead proposed a model-based covariance estimator; note that this poor performance is clearly not shared by our recommended flat-top kernels as evidenced by Tables 2a and 2b.

apprehensive regarding infinite-order kernels in general. However, it is the thesis of this paper that those poor results are not associated with the infinite order but rather with the unsmoothness of the truncated kernel.

By contrast, all four of our recommended flat-top kernels of Figure 1 beat the traditional kernels in almost all instances of spectral and cross-spectral estimators considered; the single exception is the AR(1) case F_{11} in Model I, the reason being that in that case the traditional estimators enjoy the benefit of an ultra-accurate, model-based, optimal bandwidth choice from a model that happens to be correct. Given the same benefit, flat-top kernels would do similarly well as the example of Truncated-A in Tables 1a and 1b clearly shows.

In the first part of the paper, the optimal performance of flat-top kernels was substantiated with asymptotic theorems. It is of particular importance that the optimality of flat-top kernels (after the proposed positive semi-definite transformation) seems to kick in even in sample sizes as small as $T = 100$ making them a valuable tool for practical use.

8 Appendix A: Large-sample covariance matrix estimation

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 $V_t = V_t(\theta)$, $t = 1, \dots, T$, where V_t takes values in \mathbb{R}^d , i.e., Ω as defined in eq. (3).

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

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

As usual, we set $\hat{\Gamma}(j) = 0$ for $|j| \geq T$.

The typical heteroskedasticity and autocorrelation consistent (HAC) kernel estimator of Ω has the form

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

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

$\{\kappa : \mathbb{R} \rightarrow [-1, 1], \kappa$ is symmetric, continuous at 0 and for all but a finite number of points,

$$\text{and satisfying } \kappa(0) = 1 \text{ and } \int_{\mathbb{R}} \kappa^2(x)dx < \infty\}. \quad (28)$$

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 $i = \sqrt{-1}$. In statistics, $\kappa(\cdot)$ is typically called a ‘lag-window’. With the exception of the ‘truncated’ window $\kappa_{trunc}(x)$, 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.

We now consider the idealized estimator

$$\hat{\Omega} = \sum_{j=-T}^T \kappa(j/s_T) \hat{\Gamma}(j), \quad (29)$$

that is computed as if the sequence $V_t, t = 1, \dots, T$ were directly observable; the definition of $\hat{\Gamma}(m)$ for the above is found in eq. (7).

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 order to be able to use results such as Theorem 2.1 (ii) and (iii) in the setting of large-sample HAC covariance matrix estimation, 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).

Lemma 8.1 *Assume Assumptions A, B and C of Andrews (1991) hold true, and that κ satisfies eq. (28). Further assume that, as $T \rightarrow \infty$, we have $s_T/T \rightarrow 0$ and that:*

(i) $s_T^{-1} \sum_{j=-T+1}^{T-1} |\kappa(j/s_T)| = O(1)$;

(ii) $\text{Bias}(\hat{\Omega}) = O(\sqrt{s_T/T})$; and

(iii) $s_T \rightarrow \infty$ 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 $\hat{\Omega} - \hat{\hat{\Omega}} = o_P(\sqrt{s_T/T})$.

Condition (i) of Lemma 8.1 is immediately satisfied if the kernel κ ‘cuts-off’, e.g., if $\kappa(x) = 0$ for $|x| > \text{some } x_0$. Condition (ii) of Lemma 8.1 can be viewed as a restriction (a lower bound) on the rate of growth of s_T .

Note that the flat-top family of kernels (4) satisfies eq. (28). So, let $\hat{\Omega}_\lambda$ be the estimator $\hat{\Omega}$ that uses a flat-top kernel λ instead of κ , and let $\hat{\hat{\Omega}}_\lambda$ denote its corresponding HAC

estimator. Thus, in view of Lemma 8.1, the large-sample properties of $\hat{\Omega}_\lambda$ that are derived in the main body of the paper carry over *verbatim* to the HAC flat-top estimator $\hat{\hat{\Omega}}_\lambda$.

9 Appendix B: Technical proofs

PROOF OF THEOREM 2.1. In view of eq. (8), 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

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

where

$$\begin{aligned} A_1 &= \frac{1}{2\pi} \sum_{m=-T+1}^{T-1} \left(\lambda_{g,c}\left(\frac{m}{S_{jk}}\right) - 1 \right) \Gamma_{jk}(m) e^{-imw} \\ A_2 &= -\frac{1}{2\pi T} \sum_{m=-T+1}^{T-1} |m| \lambda_{g,c}\left(\frac{m}{S_{jk}}\right) \Gamma_{jk}(m) e^{-imw} \\ A_3 &= -\frac{1}{2\pi} \sum_{|m| \geq T} \Gamma_{jk}(m) e^{-imw}. \end{aligned}$$

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

$$\begin{aligned} a_1 &= \frac{1}{2\pi} \sum_{|m| \leq cS_{jk}} \left(\lambda_{g,c}\left(\frac{m}{S_{jk}}\right) - 1 \right) \Gamma_{jk}(m) e^{-imw} \\ a_2 &= \frac{1}{2\pi} \sum_{cS_{jk} < |m| \leq T} \left(\lambda_{g,c}\left(\frac{m}{S_{jk}}\right) - 1 \right) \Gamma_{jk}(m) e^{-imw} \end{aligned}$$

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

$$|a_2| \leq \frac{1}{\pi} \sum_{cS_{jk} < m \leq T} \left| \lambda_{g,c}\left(\frac{m}{S_{jk}}\right) - 1 \right| |\Gamma_{jk}(m)| \leq \frac{1}{\pi} \sum_{cS_{jk} < m \leq T} 2|\Gamma_{jk}(m)| \quad (30)$$

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

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

Under the condition of part (ii), eq. (30) 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 3.1, we will need the following auxiliary lemma.

Lemma 9.1 *Eq. (14), together with the assumption $\Gamma_{jj}(0) > 0$ for all j , implies that*

$$E \left| \sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)} \right|^{1+\delta} = O(1/T^{\alpha(1+\delta)}) \quad \text{for all } j, k. \quad (31)$$

PROOF OF LEMMA 9.1. Let $\Delta = 1 + \delta$, and note that:

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

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

$$E |\sqrt{\hat{\Gamma}_{kk}(0)} - \sqrt{\Gamma_{kk}(0)}|^\Delta = E \frac{|\hat{\Gamma}_{kk}(0) - \Gamma_{kk}(0)|^\Delta}{\left(\sqrt{\hat{\Gamma}_{kk}(0)} + \sqrt{\Gamma_{kk}(0)} \right)^\Delta} \leq E \frac{|\hat{\Gamma}_{kk}(0) - \Gamma_{kk}(0)|^\Delta}{\sqrt{\Gamma_{kk}(0)^\Delta}} = O(1/T^{\alpha\Delta}) \quad (32)$$

by eq. (14). Therefore, $A_2 = O(1/T^{\alpha\Delta})$.

Note that inequality (32) 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

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

the second inequality being the Cauchy-Schwarz, and the last claim due to eq. (14). Hence, $A_1 = O(1/T^{\alpha\Delta})$ as well, and the lemma is proven. \square .

PROOF OF THEOREM 3.1. Note that (13) follows by eq. (14) using Jensen's and Markov's inequality. Now by (13) we have:

$$\hat{F}_{jk} = \sqrt{\hat{\Gamma}_{jj}(0)\hat{\Gamma}_{kk}(0)}\hat{f}_{jk} = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)}\hat{f}_{jk} + O_P(1/T^\alpha). \quad (33)$$

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 (15). Hence, $\max |\hat{f}_{jk}|^\Delta = O(S_{jk}^\Delta)$. Therefore, by eq. (31) we have:

$$E|W_T|^\Delta = O(S_{jk}^\Delta/T^{\alpha\Delta}). \quad (34)$$

Proof of (i) and (ii). Recall that $T^\alpha W_T = O_P(1)$ by eq. (33). Since $S_{jk} \rightarrow \infty$, it follows that $\frac{T^\alpha}{S_{jk}} W_T = o_P(1)$. But then eq. (34) implies that the sequence $\frac{T^\alpha}{S_{jk}} W_T$ is uniformly integrable; hence

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

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,

$$Bias(\hat{F}_{jk}) = \sqrt{\Gamma_{jj}(0)\Gamma_{kk}(0)} Bias(\hat{f}_{jk}) + o(S_{jk}/T^\alpha). \quad (35)$$

But from part (i) of Theorem 2.1 we have: $Bias(\hat{F}_{jk}) = O(1/S_{jk}^r)$; it follows that

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

Recall that $Var(\hat{f}_{jk}) = O(S_{jk}/T)$ by eq. (12). 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 3.1, and the resulting rate of convergence of \hat{f}_{jk} as given in eq. (16). Finally, note that the $O_P(1/T^\alpha)$ term in eq. (33) is negligible compared to the accuracy of \hat{f}_{jk} as given in (16). Thus, eq. (33) together with (16) implies (17), and part (i) is proven.

To prove part (ii), recall that from part (ii) of Theorem 2.1 we have $Bias(\hat{F}_{jk}) = O(1/T)$. Plugging the optimal bandwidth $S_{jk} = A \log T$ in eq. (35) we obtain:

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

Recall that $Var(\hat{f}_{jk}) = O(\log T/T)$ by eq. (12). 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. (33), \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. (34) 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

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

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 2.1 we have: $Bias(\hat{F}_{jk}) = O(1/T)$; it follows that

$$\text{Bias}(\hat{f}_{jk}) = O(1/T) + o(\log \log T/T^\alpha) = O(\log \log T/T^\alpha). \quad (38)$$

Recalling that $\text{Var}(\hat{f}_{jk}) = O(1/T)$ by eq. (12), gives the stated rate of convergence for \hat{f}_{jk} which—by eq. (33)—is the same as that of \hat{F}_{jk} , and part (iii) of the theorem is proven. \square

PROOF OF THEOREM 4.1. The condition $\hat{F} = F + O_P(1/R_T)$ implies

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

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 $|\hat{\lambda}_j^+ - \lambda_j| \leq |\hat{\lambda}_j - \lambda_j|$ always. Hence, it follows that

$$\hat{\lambda}_j^+ = \lambda_j + O_P(1/R_T) \quad \text{for all } j, \quad \text{and hence } \hat{\Lambda}^+ = \Lambda + O_P(1/R_T). \quad (40)$$

Using eq. (39) and (40) we have the following:

$$\begin{aligned} F + O_P(1/R_T) = \hat{F} &= \hat{U} \hat{\Lambda} \hat{U}^* = \hat{U} (\Lambda + O_P(1/R_T)) \hat{U}^* \\ &= \hat{U} (\Lambda^+ + O_P(1/R_T)) \hat{U}^* = \hat{F}^+ + O_P(1/R_T), \end{aligned}$$

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

PROOF OF THEOREM 6.1. The proof is analogous to the proof of Theorem 2.3 of Politis (2003) and is omitted. \square

PROOF OF LEMMA 8.1. The case $s_T \rightarrow \infty$ 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 $(Ts_T)^{-1/2} \sum_{j=-T+1}^{T-1} \kappa(|j|/s_T) \rightarrow 0$; 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 $EV_t \frac{\partial}{\partial \theta} V_{t-j} = 0$. \square

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. (1979). Confidence intervals for the crosscovariance function. *Selecta Statistica Canadiana*, 5, pp. 3-16.
- [4] Brillinger, D.R. (1981), *Time Series: Data Analysis and Theory*, Holden-Day, New York.
- [5] Brillinger, D.R. and Rosenblatt, M. (1967). Asymptotic theory of k th order spectra. In *Spectral Analysis of Time Series*, (B. Harris, Ed.), Wiley, New York, pp. 153-188.
- [6] Brockwell, P. J. and Davis, R. A. (1991), *Time Series: Theory and Methods, 2nd ed.*, Springer, New York.
- [7] 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.
- [8] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events*. Springer, Berlin.
- [9] Epanechnikov, V.A. (1969). Non-parametric estimation of a multivariate probability density, *Theory of Prob. and its Applications*, vol. 14, 153-158.
- [10] Gallant, A.R. (1987). *Nonlinear Statistical Models*, John Wiley, New York.
- [11] Gallant, A.R. and White, H. (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Basil Blackwell, New York.
- [12] Hall, P. and Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors, *Econometrica*, 71, 285-317.
- [13] Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press, Princeton.

- [14] Hannan, E.J. (1957). The variance of the mean of a stationary process, *J. Roy. Statist. Soc., Ser. B*, 19, 282-285
- [15] Hannan, E.J. (1958). The estimation of the spectral density after trend removal. *J. Roy. Statist. Soc., Ser. B*, 20, 323-333.
- [16] Hannan, E.J. (1970), *Multiple Time Series*, John Wiley, New York.
- [17] Hansen, B.E. (1992). Consistent covariance matrix estimation for dependent heterogeneous processes, *Econometrica*, **60**, 967-972.
- [18] Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica*, **50**, 1029-1054
- [19] Hashimzade, N. and Vogelsang, T.J. (2004). Fixed-b asymptotic approximations of the sampling behavior of nonparametric spectral density estimators. Working paper, Dept. of Economics, Cornell University.
- [20] Jowett, G.H. (1954), The comparison of means of sets of observations from sections of independent stochastic series, *J. Roy. Statist. Soc., Ser. B*, 17, 208-227.
- [21] Kiefer, N.M. and Vogelsang, T.J. (2002). Heteroscedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation, *Econometrica*, 70, 1350-1366.
- [22] Lahiri, S.N. (2003), *Resampling Methods for Dependent Data*, Springer Verlag, New York.
- [23] Leadbetter, M.R., Lindgren, G., and Rootzen, H. (1983), *Extremes and related properties of random sequences and processes*, Springer-Verlag, New York.
- [24] 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.
- [25] Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, **55**, 703-708.
- [26] Newey, W. and West, K. (1994). Automatic lag selection in covariance matrix estimation, *Rev. Econ. Studies*, **61**, 631-653.

- [27] Parzen, E. (1957). On Consistent Estimates of the Spectrum of a Stationary Time Series. *Ann. Math. Statist.*, Vol. 28, No. 2., pp. 329-348.
- [28] Parzen, E. (1961), Mathematical Considerations in the Estimation of Spectra, *Technometrics*, vol. 3, 167-190.
- [29] Phillips, P.C.B., Sun, Y. and Jin, S. (2004). Spectral density estimation and robust hypothesis testing using steep origin kernels without truncation. UCSD Dept. of Economics Discussion Paper 2004-15 (to appear in *Intern. Econ. Review*).
- [30] 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.
- [31] Politis, D.N. (2003). Adaptive bandwidth choice, *J. Nonparam. Statist.*, vol. 15, no. 4-5, 517-533.
- [32] Politis, D.N. (2004). A heavy-tailed distribution for ARCH residuals with application to volatility prediction, *Annals of Economics and Finance*, vol. 5, pp. 283-298.
- [33] Politis, D.N., and Romano, J.P. (1995), Bias-corrected nonparametric spectral estimation. *J. Time Ser. Anal.*, **16**, 67-103.
- [34] Politis, D.N., Romano, J.P. and Wolf, M. (1999), *Subsampling*, Springer Verlag, New York.
- [35] Politis, D.N., and White, H. (2004). Automatic block-length selection for the dependent bootstrap, *Econometric Reviews*, vol. 23, no. 1, pp. 53-70.
- [36] Priestley, M.B. (1962), Basic considerations in the estimation of spectra, *Technometrics*, vol. 4, 551-564.
- [37] Priestley, M.B. (1981). *Spectral Analysis and Time Series*, Adademic Press, New York.
- [38] Robinson, P. (1991). Automatic frequency domain inference on semiparametric and nonparametric models, *Econometrica*, vol. 59, 1329-1363.
- [39] Robinson, P.M. and Velasco, C. (1997). Autocorrelation-robust inference. In *Handbook of Statistics*, (G.S. Maddala and C.R. Rao, Eds.), vol. 15, pp. 267-298, Amsterdam: North Holland.

- [40] Romano, J.P. and Thombs, L. (1996). Inference for autocorrelations under weak assumption, *J. Amer. Statist. Assoc.*, 91, 590-600.
- [41] Rosenblatt, M. (1985), *Stationary Sequences and Random Fields*, Birkhäuser, Boston.
- [42] Samarov, A. (1977). Lower bound for the risk of spectral density estimates, *Probl. Inform. Transm.*, 13, pp. 67-72.
- [43] Velasco, C. and Robinson, P.M. (2001). Edgeworth Expansions for Spectral Density Estimates and Studentized Sample Mean, *Econometric Theory*, Vol. 17, pp. 497-539
- [44] West, K.D. (1997), Another heteroskedasticity— and autocorrelation—consistent covariance matrix estimator, *J. Econometrics*, 76, pp. 171-191.