# Adaptive bandwidth choice*

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

March 18, 2002

**Abstract**

In this paper, we consider the problem of bandwidth choice in the parallel settings of nonparametric kernel smoothed spectral density and probability density estimation. We propose a new class of 'plug-in' type bandwidth estimators, and show their favorable asymptotic properties. The new estimators automatically adapt to the degree of underlying smoothness which is unknown. The main idea behind the new estimators is the use of infinite-order 'flat-top' kernels for estimation of the constants implicit in the formulas giving the asymptotically optimal bandwidth choices. The proposed bandwidth choice rule for 'flat-top' kernels has a direct analogy with the notion of thresholding in wavelets. It is shown that the use of infinite-order kernels in the pilot estimator has a twofold advantage: (a) accurate estimation of the bandwidth constants, and (b) easy determination of the required 'pilot' kernel bandwidth.

**Key words:** Bandwidth Choice, Density Estimation, Kernel Smoothing, Nonparametric function estimation, Spectral Estimation, Time Series.

*Research partially supported by NSF Grant DMS-01-04059.

# 1 Introduction

Let $f : \mathbf{R} \to \mathbf{R}$ be an unknown function to be estimated from the data $X_1, \ldots, X_N$. In the typical nonparametric set-up, nothing is assumed about $f$ except that it possesses a certain degree of smoothness. Usually, a preliminary estimator of $f$ can be easily calculated that, however, lacks the required smoothness; e.g., in the case where $f$ is a probability density and the preliminary estimator is a histogram. Often, the preliminary estimator is even inconsistent; e.g., in the case where $f$ is a spectral density and the preliminary estimator is the periodogram. Rosenblatt (1991) discusses the two cases of spectral and probability density estimation in a unified framework.

In order to obtain an estimator with good properties, for example, large-sample consistency and smoothness, one can smooth the preliminary estimator by convolving it with a function $\Lambda_h : \mathbf{R} \to \mathbf{R}$ called the 'kernel', and satisfying $\int \Lambda_h(x)dx = 1$. Let $\hat{f}$ denote such a kernel-smoothed estimator of $f$; a precise definition will be given in Sections 2 and 3 in the specific contexts of spectral and probability density estimation. It is convenient to also define the Fourier transform of the kernel as $\lambda_h(s) = \int \Lambda_h(x)e^{isx}dx$.

Typically, as the sample size $N$ increases, the kernel $\Lambda_h(\cdot)$ becomes more and more concentrated near the origin. To achieve this behavior, we let $\Lambda_h(\cdot)$ and $\lambda_h(\cdot)$ depend on a real-valued, positive 'bandwidth' parameter $h$, that is, we assume that $\Lambda_h(x) = h^{-1}\Lambda(x/h)$, and $\lambda_h(s) = \lambda(hs)$, where $\Lambda(\cdot)$ and $\lambda(\cdot)$ are some fixed (not depending on $h$) bounded functions, satisfying $\lambda(s) = \int \Lambda(x)e^{isx}dx$; the bandwidth $h$ will be assumed to be a decreasing function of the sample size.

If $\Lambda$ has finite moments up to $q$th order, and moments of order up to $q-1$ equal to zero, then $q$ is called the 'order' of the kernel $\Lambda$. If the unknown function $f$ has $p$ bounded continuous derivatives, it typically follows that

$$Var(\hat{f}(x)) = \frac{C_{f,\Lambda}(x)}{hN} + o(\frac{1}{hN}), \tag{1}$$

and

$$Bias(\hat{f}(x)) = c_{f,\Lambda}(x)h^k + o(h^k), \tag{2}$$

where $k = \min(q, p)$, and $C_{f,\Lambda}(x), c_{f,\Lambda}(x)$ are bounded functions depending on $\Lambda$, on $f$, and on $f$'s derivatives; cf. Rosenblatt (1991, p. 8).

This idea of choosing a kernel of order $q$ bigger (or equal) than $p$ in order to get the $Bias(\hat{f}(x))$ to be $O(h^p)$ dates back to Parzen (1962) and Bartlett (1963); some more recent

references on 'higher-order' kernels include the following: Devroye (1987, 1992), Gasser, Müller, and Mammitzsch (1985), Granovsky and Müller (1991), Jones (1995), Marron (1994), Müller (1988), and Scott (1992).

It is a well-known fact that optimal bandwidth selection is a crucial issue in such nonparametric smoothing problems. The goal typically is minimization of the large-sample Mean Squared Error (MSE) of $\hat{f}(x)$; however, to do this minimization, the practitioner needs to know or estimate the degree of smoothness $p$, as well as the constants $C_{f,\Lambda}(x)$ and $c_{f,\Lambda}(x)$. Traditionally, this problem has been approached either using a pilot estimator plug-in or by a cross-validation idea. Cross-validation has been well-studied in both density and spectral density estimation settings; see Hall (1983), Hurvich (1985), Hurvich and Beltrao (1990), Rachdi (1998), Rachdi and Youndjé (2000).

The paper at hand focuses instead on the plug-in method originated in Woodroofe (1970), and actively developed in the 80s and 90s; see Jones et al. (1996) or Loader (1999a,b) for a literature review. To briefly describe it, the plug-in approach typically amounts to using a *pilot* estimator of $f$ in order to consistently estimate the constants $C_{f,\Lambda}(x)$ and $c_{f,\Lambda}(x)$, assuming some value for $p$. Consequently, the estimated bandwidth is the one minimizing the resulting large-sample MSE expression (with estimated constants). However, two reasons have made many practitioners sceptical of this approach:
(a) The inaccuracies in the pilot estimator; for example, an iterative pilot plug-in method for bandwidth choice recently proposed by Brockmann et al. (1993) and followed up by Bühlmann (1996) yields an estimator $\bar{h}_{opt}$ of the asymptotically MSE-optimal bandwidth $h_{opt}$ that is consistent but at a very slow rate of convergence satisfying

$$\bar{h}_{opt} = h_{opt}(1 + O_P(N^{-2/9})); \tag{3}$$

see also Abramson (1982) where the (non-iterative) pilot plug-in idea is analyzed.
(b) The difficulty in properly choosing the bandwidth of the pilot; see e.g. Loader (1999a,b) or Politis et al. (1999, p. 190) for more discussion.

The driving observation of the paper at hand is that both problems (a) and (b) of the plug-in method intrinsically have to do with the use of finite-order kernels. For example, eq. (3) pertains to using *second-order* kernels for smoothing *and* for the pilot estimator construction. Although, the asymptotic accuracy, i.e., problem (a), can be significantly improved by using a higher-order (e.g., fourth-order) kernel for the pilot—see e.g. Hall et al. (1991), or Bühlmann (1996)—, problem (b) remains: optimally choosing the bandwidth

of the pilot.

Another important issue is the unknown degree of smoothness $p$. Notably, when one is using second-order kernels the issue is not so important: the practitioner just needs to be able to assume that $p$ is at least 2 in order to take advantage of eq. (2). However, if one is willing to use higher-order kernels, then it is important to be able to estimate $p$, explicitly or implicitly, so that a kernel of high enough order is used with a resulting small bias. In effect, the ideal procedure would automatically guarantee that the order of the kernel to be used is equal or bigger than the underlying (unknown) degree of smoothness, i.e., it would be an *adaptive* procedure; the lack of this adaptivity is perhaps the major drawback of methods based on kernels of higher (but finite) order.

Interestingly, there seems to have been little discussion in the literature regarding the use of an *infinite-order* kernel for the pilot; infinite-order kernels have been discussed in Devroye (1987, 1992) and Györfi and Devroye (1984). If one could have a practical way to choose the bandwidth of an infinite-order kernel—thus alleviating problem (b) above—, then, the speed of convergence of the higher-order kernel would help alleviate problem (a) as well. In addition, using an infinite-order kernel ensures (by definition) that the order of the kernel is always at least as big as the underlying degree of smoothness.

As a matter of fact, Politis and Romano (1993, 1995, 1996, 1999) have proposed a family of *flat-top* kernels of infinite order, and have shown their favorable asympotic properties (and optimal rate of convergence); perhaps equally importantly, they have given an easy-to-use, empirical *rule-of-thumb* for flat-top bandwidth choice. In what follows, we first expand upon the Politis and Romano (1995) rule-of-thumb for bandwidth choice, and show that it automatically *adapts* to the underlying (unknown) degree of smoothness; thus, function estimation with automatic (and accurate) bandwidth choice is possible when one is willing to use the aforementioned flat-top kernels. Nevertheless, many authors prefer to use second-order kernels; see e.g. Marron and Wand (1992) for a discussion. In this case, we propose using a flat-top kernel as a pilot; the result is that both problems (a) and (b) are alleviated due to the fast rate of convergence associated with infinite-order kernels together with the ease of empirically picking the flat-top kernel's bandwidth.

In the next section, we carefully define and show the favorable performance of our proposed bandwidth estimator in the case of the spectral density using either a flat-top kernel for smoothing, or a second-order kernel for smoothing with a flat-top kernel pilot; the results when a finite-order kernel is used for smoothing with a flat-top kernel pilot are

similar and are omitted. In section 3, we give the analogous results for the probability density case. All technical proofs are placed in the appendix.

## 2 Spectral density function

Suppose $X_1, \ldots, X_N$ are observations from the (strictly) stationary real-valued sequence $\{X_n, n \in \mathbf{Z}\}$ having mean $\mu = EX_t$, and autocovariance sequence $R(s) = E(X_t - \mu)(X_{t+|s|} - \mu)$; here both $\mu$ and $R(\cdot)$ are unknown. We now consider the problem of estimating the spectral density function $f(w) = (2\pi)^{-1} \sum_{s=-\infty}^{\infty} e^{iws} R(s)$, for $w$ $[-\pi, \pi]$, by means of a kernel $\Lambda_h$. The corresponding kernel estimator may be written as:

$$\hat{f}(w) = \Lambda_h * I_N(w) \tag{4}$$

where $I_N(w) = (2\pi)^{-1} \sum_{s=-N+1}^{N-1} e^{iws} \hat{R}(s)$ is the periodogram and $*$ denotes convolution; here $\Lambda_h(x) = h^{-1}\Lambda(x/h)$, and the real number $h$ is the bandwidth. An alternative way to represent $\hat{f}(w)$ is via the Fourier transform of the kernel (otherwise called the 'lag-window') namely $\lambda_h(s) = \lambda(hs)$, giving:

$$\hat{f}(w) = (2\pi)^{-1} \sum_{s=-\infty}^{\infty} e^{iws} \lambda_h(s) \hat{R}(s), \tag{5}$$

where $\hat{R}(k) = N^{-1} \sum_{i=1}^{N-|k|} (X_i - \bar{X}_N)(X_{i+|k|} - \bar{X}_N)$ is the lag-$k$ sample autocovariance for $|k| < N$; note that $\hat{R}(k)$ is defined to be zero for $|k| \geq N$—see Brockwell and Davis (1991) for details.

### 2.1 Bandwidth selection for flat-top kernels

To achieve accurate estimation of an infinite sum of the type $\sum_{k=-\infty}^{\infty} |k|^p e^{ikw} R(k)$, for some fixed integer $p \geq 0$, we propose to use the 'flat-top' lag-windows of Politis and Romano (1995); notably, the case $p = 0$ is the spectral density, the case $p = 2$ is (a multiple of) the second derivative of the spectral density, and so forth. Thus, using the form of eq. (5), we estimate

$$f_p(w) \equiv (2\pi)^{-1} \sum_{k=-\infty}^{\infty} |k|^p e^{ikw} R(k) \quad \text{by} \quad \tilde{f}_p(w) = (2\pi)^{-1} \sum_{k=-M}^{M} \lambda^T(k/M) |k|^p e^{ikw} \hat{R}(k), \tag{6}$$

where the function $\lambda^T$ has a trapezoidal shape symmetric around zero, i.e.,

$$\lambda^T(t) = \begin{cases} 1 & \text{if } |t| \in [0, 1/2] \\ 2(1-|t|) & \text{if } |t| \in [1/2, 1] \\ 0 & \text{else.} \end{cases} \tag{7}$$

Note that $\lambda^T$ is the simplest representative of the family of flat-top kernels; other choices are also available possessing similar properties—see Politis and Romano (1993, 1995, 1999) or Politis (2001). The following theorem quantifies those asymptotic properties; note that the case $p = 0$ of the theorem is covered in Politis and Romano (1995).

**Theorem 2.1** *Assume conditions strong enough to ensure that*[*]

$$Var(\tilde{f}_p(w)) = O(M/N). \tag{8}$$

*(i) Assume that* $\sum_{s=-\infty}^{\infty} |s|^{(r+p)}|R(s)| < \infty$ *for some positive integer* $r$*; then letting* $M$ *proportional to* $N^{1/(2r+1)}$ *yields*

$$\tilde{f}_p(w) = f_p(w) + O_P(N^{-r/(2r+1)}).$$

*(ii) If* $|R(s)| \leq Ce^{-as}$ *for some constants* $a, C > 0$*, then letting* $M \sim A \log N$*, for some appropriate constant* $A$*, yields*

$$\tilde{f}_p(w) = f_p(w) + O_P(\frac{\sqrt{\log N}}{\sqrt{N}});$$

*as usual, the notation* $A \sim B$ *means* $A/B \to 1$.
*(iii) If* $R(s) = 0$ *for* $|s| >$ *some* $q$*, then letting* $M = 2q$*, yields*

$$\tilde{f}_p(w) = f_p(w) + O_P(\frac{1}{\sqrt{N}}).$$

The quantity $1/M$ serves as the bandwidth of the associated kernel $\Lambda^T$ that is the (inverse) Fourier transform of the lag-window $\lambda^T$. Note that a closed-form formula for $\Lambda^T$ is given in Politis and Romano (1993, 1995) by means of a linear combination of two Fejer

[*]There exist different sets of conditions sufficient for eq. (8); for example, we may assume $E|X_t|^{6+\delta} < \infty$, and $\sum_{k=1}^{\infty} k^2(\alpha_X(k))^{\frac{\delta}{6+\delta}} < \infty$ for some $\delta > 0$. The $\alpha$-mixing coefficients are defined in the following way: let $\mathcal{F}_n^m$ be the $\sigma$-algebra generated by $\{X_t, n \leq t \leq m\}$, and define $\alpha_X(k) = \sup_n \sup_{A,B} |P(A \cap B) - P(A)P(B)|$, where $A$ and $B$ vary over the $\sigma$-fields $\mathcal{F}_{-\infty}^n$ and $\mathcal{F}_{n+k}^{\infty}$, respectively; see Rosenblatt (1956).

kernels. Thus, choosing $M$ is tantamount to bandwidth choice for the flat-top kernel/lag-window pair $\Lambda^T$ and $\lambda^T$. Theorem 2.1 gives the optimal (with respect to minimization of the large-sample MSE of $\tilde{f}_p(w)$) values of $M$.

As mentioned in the Introduction, besides the favorable asymptotic properties and speed of convergence associated with flat-top kernel and demonstrated in Theorem 2.1, another reason for using the flat-top lag-window is that choosing its bandwidth in practice is intuitive and doable by a simple inspection of the correlogram, i.e., the plot of $\hat{R}(k)$ vs. $k$. Motivated by case $p = 0$ of Theorem 2.1 (iii), Politis and Romano (1995) suggested looking for a point, say $\hat{m}$, after which the correlogram appears negligible, i.e., $\hat{R}(k) \simeq 0$ for $k > \hat{m}$ (but $\hat{R}(\hat{m}) \neq 0$). Of course, $\hat{R}(k) \simeq 0$ is taken to mean that $\hat{R}(k)$ is *not* significantly different from zero, i.e., an implied hypothesis test. After identifying $\hat{m}$, the recommendation is to just take $M = 2\hat{m}$. We will now provide some mathematical evidence supporting that the $M = 2\hat{m}$ rule automatically captures the correct order of magnitude for $M$, thus adapting to the three cases of Theorem 2.1. Before doing that, let us carefully define the empirical rule.

**EMPIRICAL RULE OF PICKING $M$:** *Let $\rho(k) = R(k)/R(0)$, and $\hat{\rho}(k) = \hat{R}(k)/\hat{R}(0)$. Let $\hat{m}$ be the smallest positive integer such that $|\hat{\rho}(\hat{m}+k)| < c\sqrt{\log N/N}$, for $k = 1, \ldots, K_N$, where $c > 0$ is a fixed constant, and $K_N$ is a positive, nondecreasing integer-valued function of $N$ such that $K_N = o(\log N)$. Then, let $\hat{M} = 2\hat{m}$.*

Note that, because $\hat{\rho}(k) = 0$ for $|k| \geq N$, the above minimization problem is always well-defined, although a case where $\hat{m}$ and $M$ turn out comparable to $N$ deserves further scrutiny; as is well-known, we need $\hat{m}$ and $M$ to be of smaller order than $N$ in order to have estimators with small variance.

**Remark 2.1** The form of the threshold $c\sqrt{\log N/N}$ is reminiscent of the Donoho-Johnstone thresholding rule for wavelet coefficients; see e.g. Donoho and Johnstone (1994), Donoho et al. (1996) or Härdle et al. (1998). In fact, there is more here than just a passing analogy. The spectral estimators $\tilde{f}_p(w)$ employing the flat-top lag-window $\lambda^T$ can indeed be thought of as wavelet estimators with 'soft' thresholding, and the cosine functions serving as wavelets. Similarly, the so-called 'truncated' periodogram $T(w) = (2\pi)^{-1} \sum_{k=-M}^{M} e^{ikw} \hat{R}(k)$, can be thought of as a wavelet estimator with 'hard' thresholding; the truncated periodogram also

belongs in the general class of flat-top kernels of Politis and Romano (1993, 1995) although it is *not* recommendable in practice because of the slowly-decaying sidelobes of the Dirichlet kernel.

**Remark 2.2** The above empirical rule of picking $M$ is easy to use and quite intuitive especially since practitioners invariably look at the correlogram at the first stage of any time series analysis. Indeed, the $\pm 1.96/\sqrt{N}$ bands around zero are typically produced by most statistical packages indicating approximate 95% confidence intervals/hypothesis tests for each of the autocorrelations (under an implied linearity condition for the time series $\{X_t\}$). However, it should be stressed that this empirical rule simply does *not* work for lag-windows that are not in the flat-top class of Politis and Romano (1995, 1999). As a matter of fact, the empirical rule as stated above is tailor-made for the *particular* flat-top kernel given in eq. (7); however, it is easy to modify it for use with other flat-top kernels.

**Remark 2.3** The constant $c$ and the value of $K_N$ are the practitioner's choice; indeed, *any* values for $c > 0$ and $1 \leq K_N \leq N$ would work for our asymptotic results albeit leading to very different finite-sample performances. Nevertheless, we can get some guidance on practically useful choices for $c$ and $K_N$ by the comparison with the construction of confidence intervals/hypothesis tests for the autocorrelations. For concreteness, let us now specify that log will denote logarithm with base 10. If the sample size under consideration is in the [100,1000] range—as it is quite typical—, the factor $\sqrt{\log N}$ ranges between 1.41 and 1.73 so its influence is rather small. Thus, the simple choice $K_N = 1$ and $c = 1$ roughly corresponds to an implied approximate 90% confidence interval, or equivalently, a level 0.10 hypothesis test, for the autocorrelation $\rho(\hat{m}+1)$. Nevertheless, it is advisable to use a bigger $K_N$—see Remark 2.4; we therefore recommend taking $K_N$ to be about 5, while at the same time increasing $c$ to a value around 2 so that our empirical rule would roughly correspond to 95% *simultaneous* intervals for $\rho(\hat{m}+k)$ with $k = 1, \ldots, K_N$ by Bonferroni's inequality.

The performance of our empirical rule of picking $M$ is quantified in the following theorem.

**Theorem 2.2** *Assume conditions strong enough to ensure that*[†] *for all finite* $n$,

$$\max_{k=1,\ldots,n} |\hat{\rho}(s+k) - \rho(s+k)| = O_P(1/\sqrt{N}) \tag{9}$$

[†]There exist different sets of conditions sufficient for eq. (9); see Brockwell and Davis (1991) or Romano

*uniformly in s, and*

$$\max_{k=0,1,\ldots,N-1} |\hat{\rho}(k) - \rho(k)| = O_P(\sqrt{\frac{\log N}{N}}). \tag{10}$$

*Also assume* $|R(k)| > 0$ *for all* $k \leq$ *some* $k_0$.
*(i) Assume that* $R(k) = Ck^{-d}$ *for* $k > k_0$, *and for some* $C > 0$, *and a positive integer* $d$. *Then,*

$$\hat{M} \overset{P}{\sim} \frac{A_1 N^{1/2d}}{(\log N)^{1/2d}}$$

*for some positive constant* $A_1$; *note that the notation* $A \overset{P}{\sim} B$ *means* $A/B \to 1$ *in probability.*
*(ii) Assume* $R(k) = C\xi^k$ *for* $k > k_0$, *where* $C > 0$, *and* $|\xi| < 1$ *are some constants. Then,*

$$\hat{M} \overset{P}{\sim} A_2 \log N$$

*where* $A_2 = -1/\log|\xi|$.
*(iii) If* $R(k) = 0$ *for all* $k > q \equiv k_0$, *but* $R(q) \neq 0$, *then*

$$\hat{M} = 2q + o_P(1).$$

Comparing the empirical rule $\hat{M}$ to the theoretically optimal values of $M$ given in Theorem 2.1 we see that $\hat{M}$ manages to capture exactly the theoretically optimal rate in cases (ii) and (iii) of Theorem 2.2. In case (i) of Theorem 2.2, $\hat{M}$ increases essentially as a power of $N$ since the $2d$-th root of the logarithm changes in an ultra-slow way with $N$; note that the empirically found exponent $1/2d$ is slightly smaller than the theoretically optimal given in Theorem 2.1 (i) but the difference is small, and becomes even smaller for large $d$.

Thus, $\hat{M}$ automatically adapts to the underlying rate of decay of the autocorrelation function, switching between the polynomial, logarithmic, and constant rates that are optimal respectively in the three cases of Theorem 2.1. In addition, focusing on the interesting case (ii), a further adaptivity of $\hat{M}$ is apparent as the correct constant $A_2 = -1/\log|\xi|$ is also captured; note that, as expected, the constant $A_2$ is large in the case of a large $|\xi|$ which corresponds to stronger dependence (thus requiring a bigger value of $M$).

and Thombs (1996). As a matter of fact, under further regularity conditions, the process $\sqrt{N}(\hat{\rho}(\cdot) - \rho(\cdot))$ is asymptotically Gaussian with autocovariance tending to zero; consequently, eq. (10) would follow from the theory of extremes of dependent sequences—see e.g. Leadbetter et al. (1983).

**Remark 2.4** Note that, under either one of the conditions of Theorem 2.2, we could take $K_N = 1$ in the empirical rule of selecting $M$ with good results. However, this was at the expense of slightly strong assumptions; for example, the assumption $|R(k)| > 0$ for all $k \leq$ some $k_0$ is quite strong. Also observe that condition (ii) of Theorem 2.2 is a subcase of condition (ii) of Theorem 2.1, and condition (i) is only a subcase of condition (i) of Theorem 2.1. Nevertheless, condition (ii) of Theorem 2.2 has some interest as it corresponds to the case of an AR(1) model for $\{X_t\}$. We generalize conditions (i) and (ii) of Theorem 2.2 in the theorem below in which the benefit of using a $K_N$ bigger than one is apparent.

**Theorem 2.3** *Assume conditions strong enough to ensure (9) and (10) hold. Also assume that the sequence $R(k)$ does not have more than $K_N - 1$ zeros in its first $k_0$ lags (i.e., for $k = 1, \ldots, k_0$).*
*(i) Assume that $R(k) = Ck^{-d}\cos(ak + \theta)$ for $k > k_0$, and for some $C > 0$ and a positive integer $d$, and some constants $a \geq \frac{\pi}{K_N}, \theta \in [0, 2\pi]$. Then,*

$$\hat{M} \overset{P}{\sim} \frac{A_1 N^{1/2d}}{(\log N)^{1/2d}}$$

*for some positive constant $A_1$.*
*(ii) Assume $R(k) = C\xi^k \cos(ak + \theta)$ for $k > k_0$, where $C > 0, |\xi| < 1, a \geq \frac{\pi}{K_N}, \theta \in [0, 2\pi]$ are some constants. Then,*

$$\hat{M} \overset{P}{\sim} A_2 \log N$$

*where $A_2 = -1/\log|\xi|$.*
*(iii) If $R(k) = 0$ for $|k| >$ some integer $q$ (with $q < k_0 + K_N$), but $R(q) \neq 0$, then*

$$\hat{M} = 2q + o_P(1).$$

**Remark 2.5** Observe that the assumptions of Theorem 2.3 explicitly involve $K_N$; the interpretation of those conditions is that the empirical rule of picking $M$ will work as expected if the $K_N$ is chosen large enough so that the assumptions are satisfied. For example, the condition $a \geq \pi/K_N$ should be interpreted: the empirical rule of picking $M$ will work well if $K_N$ is (or eventually becomes with increasing $N$) bigger than $\pi/a$.

**Remark 2.6** Note that condition (iii) of Theorem 2.3 is identical to condition (iii) of Theorem 2.1, and corresponds to the case where $\{X_t\}$ is an MA($q$) model; see Brockwell

and Davis (1991). Condition (ii) of Theorem 2.3 attempts to capture the interesting case where $\{X_t\}$ satisfies a stationary ARMA model. To see this, recall that the autocovariance $R(k)$ of a stationary ARMA model satisfies

$$R(k) \sim C\xi^k \cos(ak + \theta), \quad \text{for large } k, \tag{11}$$

where $\xi$ is essentially the modulus of the characteristic polynomial root (or conjugate double roots) closest to the unit circle; cf. Brockwell and Davis (1991). Our Theorem 2.3 (ii), condition: $R(k) = C\xi^k \cos(ak + \theta)$ for $k > k_0$, is a simple approximation to eq. (11). In our final result of this subsection, we focus on eq. (11) as given, that is, without resorting to an approximation.

**Theorem 2.4** *Assume conditions strong enough to ensure the validity of eq. (9) and (10). Assume that either* $R(k) = C\xi^k + \epsilon_k$ *or* $R(k) = C\xi^k \cos(ak + \theta) + \epsilon_k$, *where* $C > 0, |\xi| < 1, a \geq \frac{\pi}{K_N}, \theta \in [0, 2\pi]$ *are some constants, and* $\epsilon_k = o(\xi^k)$ *as* $k$ *increases. Also assume that the sequence* $R(k)$ *does not have more than* $K_N - 1$ *consecutive zeros in the first* $[A_0 \log N]$ *places, i.e., in the index set* $\{1, 2, \ldots, [A_0 \log N]\}$, *where* $A_0$ *is some constant satisfying* $A_0 \geq -1/(2 \log |\xi|)$.
*Then,*

$$\hat{M} \overset{P}{\sim} A_2 \log N$$

*where* $A_2 = -1/\log |\xi|$.

## 2.2 Bandwidth selection for second order kernels using flat-top pilots

We now focus on the spectral estimator $\hat{f}(w) = \Lambda_h * I_N(w)$ of (4) with $\Lambda_h$ being a *second-*order nonnegative kernel, i.e., we assume that $\int \Lambda_h(x)dx = 1$, $\int x\Lambda_h(x)dx = 0$, and that $\Lambda_h(x) \geq 0$ for all $x$; see e.g. Priestley (1981). From eq. (2) it then follows that

$$Bias(\hat{f}(w)) = c_{f,\Lambda}(w)h^2 + o(h^2), \quad with \quad c_{f,\Lambda}(w) = -\frac{f''(w)}{2}\int x^2\Lambda(x)dx; \tag{12}$$

cf. Rosenblatt (1991).

In addition, under standard regularity conditions (e.g., the ones appearing in the footnote to Theorem 2.1), eq. (1) holds true, and in particular:

$$Var(\hat{f}(w)) = \frac{C_{f,\Lambda}(w)}{hN} + o(\frac{1}{hN}), \quad with \quad C_{f,\Lambda}(w) = f^2(w)\eta(w)\int \lambda^2(t)dt \tag{13}$$

where $\eta(w) = 2$ if $w$ is an integer multiple of $\pi$, and $\eta(w) = 1$ otherwise.

The asymptotically MSE–optimal bandwidth choice is then given by:

$$h_{opt} = \left( \frac{C_{f,\Lambda}(w)}{4c^2_{f,\Lambda}(w)\ N} \right)^{1/5}. \tag{14}$$

**Remark 2.7** Note that $h_{opt}$ is a *local* asymptotically optimal bandwidth as its goal is to minimize the large-sample MSE of $\hat{f}(w)$ at a *particular* $w$-point; thus, a more accurate notation for $h_{opt}$ might be $h_{opt}(w)$ but the shorter $h_{opt}$ should create no confusion. A *global* bandwidth may also be constructed that would have the minimization of the large-sample limit of $\int_{-\pi}^{\pi} MSE(\hat{f}(w))dw$ as its target; the arguments are similar.

Let $\tilde{c}_{f,\Lambda}(w), \tilde{C}_{f,\Lambda}(w)$ denote the estimates of $c_{f,\Lambda}(w), C_{f,\Lambda}(w)$ using the estimators $\tilde{f}(w)$ and $\tilde{f}_2(w)$ in place of the unknowns $f(w)$ and $f_2(w)$. Recall that $\tilde{f}(w)$ and $\tilde{f}_2(w)$ were defined in the previous subsection using the flat-top kernel $\Lambda^T$; note that $f''(w) = -f_2(w)$, so our estimator of $f''(w)$ is $-\tilde{f}_2(w)$.

Our data-dependent approximation to $h_{opt}$ is given by:

$$\hat{h}_{opt} = \left( \frac{\tilde{C}_{f,\Lambda}(w)}{4\tilde{c}^2_{f,\Lambda}(w)\ N} \right)^{1/5}. \tag{15}$$

We first give a result concerning the performance of $\hat{h}_{opt}$ when the flat-top kernel's bandwidth $1/M$ is chosen in a non data-dependent way.

**Theorem 2.5** *(i) If* $\sum_{k=-\infty}^{\infty} |k|^{r+2}|R(k)| < \infty$ *for some* $r \geq 2$, *then letting* $M \sim AN^{1/(2r+1)}$ *for some* $A > 0$, *we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(N^{-r/(2r+1)})).$$

*(i′) If* $\sum_{k=-\infty}^{\infty} k^4|R(k)| < \infty$, *then letting* $M \sim AN^{1/5}$ *for some* $A > 0$, *we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(N^{-2/5})).$$

*(ii) If* $|R(k)| \leq Ce^{-ak}$ *for some* $a, C > 0$, *then letting* $M \sim A \log N$, *for some appropriate constant* $A$, *we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(\frac{\sqrt{\log N}}{\sqrt{N}})).$$

*(iii) If* $R(k) = 0$ *for* $|k| >$ *some* $q$, *then letting* $M = 2q$, *we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(\frac{1}{\sqrt{N}})).$$

A simplistic way of applying Theorem 2.5 is to focus on the worst-case scenario autocovariance rate of decay, i.e., case $(i')$ where $\sum_{k=-\infty}^{\infty} k^4|R(k)| < \infty$, and choose $M \sim AN^{1/5}$; then, you are guaranteed that $\hat{h}_{opt} = h_{opt}(1 + O_P(N^{-2/5}))$ which is still a big improvement over eq. (3). Nevertheless, this conservative approach is suboptimal if the autocovariance decays much faster. In addition, the practitioner is faced with the problem of choosing the constant $A$ which—for any given finite-sample set-up—is tantamount to choosing $M$ itself! A way to by-pass both of these problems is given by our automatic, data-dependent empirical rule of picking $M$ (presented in the previous subsection) in which the rate of decay of the autocovariance is assessed by looking at the correlogram.

**Theorem 2.6** *Using a flat-top pilot with bandwidth equal to $1/\hat{M}$ as given by our empirical rule of picking $M$, we obtain an estimated bandwidth $\hat{h}_{opt}$ with the following properties.*
*(i) Assume $d > 6$, and either the conditions of Theorem 2.2 (i), or the conditions of Theorem 2.3 (i). Then,*

$$\hat{h}_{opt} = h_{opt}(1 + O_P((\frac{\log N}{N})^{(d-5)/(2d)})) \quad \textit{if } d \textit{ is an integer,}$$

*and*

$$\hat{h}_{opt} = h_{opt}(1 + O_P((\frac{\log N}{N})^{[d-4]/(2d)})) \quad \textit{if } d \textit{ is not an integer,}$$

*where $[\cdot]$ denotes the integer part.*
*(ii) Under the conditions of Theorem 2.2 (ii), or the conditions of Theorem 2.3 (ii), or the conditions of Theorem 2.4, we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(\frac{\sqrt{\log N}}{\sqrt{N}})).$$

*(iii) Under the conditions of Theorem 2.2 (iii), or the conditions of Theorem 2.3 (iii), we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(\frac{1}{\sqrt{N}})).$$

Theorem 2.6 demonstrates the increased accuracy of our estimated bandwidth $\hat{h}_{opt}$ as compared to rates such as the one given in eq. (3). In addition, the estimated bandwidth $\hat{h}_{opt}$ inherits the automatic adaptivity to the (unknown) underlying degree of smoothness by way of our empirical rule of picking $M$ in the bandwidth choice for the flat-top pilot.

# 3 Probability density function

We now briefly address the problem of probability density estimation. To do this we assume that the first marginal distribution of the (strictly) stationary series $\{X_n, n \in \mathbf{Z}\}$ possesses the (unknown) probability density $f(x)$. The analogy between the spectral density and the probability density is well known; see e.g. Rosenblatt (1991).

The kernel smoothed estimator of $f(x)$, for some fixed $x \in \mathbf{R}$, based on a kernel $\Lambda$ is given by

$$\hat{f}(x) = \frac{1}{N} \sum_{k=1}^{N} \Lambda_h (x - X_k). \tag{16}$$

For $s \in \mathbf{R}$, define the characteristic function $R(s) = \int_{-\infty}^{\infty} e^{-isx} f(x)dx$, and let $\hat{R}(s)$ be the sample characteristic function defined by $\hat{R}(s) = \frac{1}{N} \sum_{k=1}^{N} e^{-isX_k}$. Recall that a Fourier inversion gives:

$$f(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{isx} R(s)ds \quad \text{and} \quad \hat{f}(x) = (2\pi)^{-1} \int_{-\infty}^{\infty} \lambda_h(s) e^{isx} \hat{R}(s)ds.$$

With the aforementioned definitions, the analogy between spectral density and probability density is complete, the main difference being that, while the spectral density must be an even function, the probability density does not necessarily have symmetry; consequently, $R(s)$ is now a complex valued function of the *real* argument $s$.

## 3.1 Bandwidth selection for flat-top kernels

By analogy to eq. (6), we estimate

$$f_p(x) \equiv (2\pi)^{-1} \int_{-\infty}^{\infty} |s|^p e^{isx} R(s)ds \quad \text{by} \quad \tilde{f}_p(x) = (2\pi)^{-1} \int_{-M}^{M} \lambda^T(s/M)|s|^p e^{isx} \hat{R}(s)ds; \tag{17}$$

thus, $\tilde{f}_p(x)$ is our flat-top kernel estimator of the probability density ($p = 0$) and its derivatives ($p > 0$). The following theorem is the analog of Theorem 2.1 in the probability density case; the case $p = 0$ was previously covered in Politis and Romano (1993, 1999).

**Theorem 3.1** *Assume conditions strong enough to ensure*[‡] *that*

$$Var(\tilde{f}_p(x)) = O(M/N). \tag{18}$$

[‡]There exist different sets of conditions sufficient for eq. (18); for example, Hallin and Tran (1996) show eq. (18) under an assumption of linearity for the time series $\{X_n\}$. Interestingly, as long as the linearity coefficients have a fast (polynomial) decay, the (weak) dependence present in the time series $\{X_n\}$ does

*(i) Assume that* $\int_{-\infty}^{\infty} |s|^{(r+p)}|R(s)|ds < \infty$ *for some positive integer* $r$*; then letting* $M$ *proportional to* $N^{1/(2r+1)}$ *yields*

$$\tilde{f}_p(x) = f_p(x) + O_P(N^{-r/(2r+1)}).$$

*(ii) If* $|R(s)| \leq Ce^{-as}$ *for some constants* $a, C > 0$*, then letting* $M \sim A \log N$*, for some appropriate constant* $A$*, yields*

$$\tilde{f}_p(x) = f_p(x) + O_P(\frac{\sqrt{\log N}}{\sqrt{N}}).$$

*(iii) If* $R(s) = 0$ *for* $|s| >$ *some* $q$*, then letting* $M = 2q$*, yields*

$$\tilde{f}_p(x) = f_p(x) + O_P(\frac{1}{\sqrt{N}}).$$

We now need to define an analog of our empirical rule of picking the flat-top kernel's inverse-bandwidth $M$ in the probability density case. The analog of the correlogram is the plot of the sample characteristic function which, however, is a complex-valued function of the continuous argument $s$; therefore, in what follows, $|R(s)|$ denotes the modulus of a complex number.

**EMPIRICAL RULE OF PICKING** $M$ **(continuous version):** *Let* $\rho(s) = R(s)/R(0)$*, and* $\hat{\rho}(s) = \hat{R}(s)/\hat{R}(0)$*. Let* $\hat{m}$ *be the smallest positive real number such that* $|\hat{\rho}(\hat{m} + s)| < c\sqrt{\log N/N}$*, for all* $s \in (0, K_N)$*, where* $c > 0$ *is a fixed constant, and* $K_N$ *is a positive, nondecreasing real-valued function of* $N$ *such that* $K_N = o(\log N)$*. Then, let* $\hat{M} = 2\hat{m}$*.*

All our Theorems from Section 2.1 have analogs in the probability dansity case. For brevity, we focus on just one of them, Theorem 2.3, and give its extension below.

**Theorem 3.2** *Assume conditions strong enough to ensure (9) and (10) hold. Let* $s_0 > 0$*, and assume that the function* $R(\cdot)$ *satisfies the following: there is no interval* $I$ *of length* $K_N$*, with* $I \subset (0, s_0)$*, such that* $R(s) = 0$ *for all* $s \in I$*.*

not seem to influence the large-sample variance of either $\hat{f}$ or $\tilde{f}_p$; these variances can actually be calculated as if the sequence $\{X_n\}$ were independent. On the other hand, such conditions on weak dependence are necessary since from recent work of Adams and Nobel (1998) it is known that ergodicity alone is not sufficient to guarantee consistent density estimation.

*(i) Assume that* $|R(s)| = |Cs^{-d}\cos(as+\theta)|$ *for* $s > s_0$*, and for some* $C > 0$ *and a positive integer* $d$*, and some constants* $a \geq \frac{\pi}{K_N}, \theta \in [0, 2\pi]$*. Then,*

$$\hat{M} \overset{P}{\sim} \frac{A_1 N^{1/2d}}{(\log N)^{1/2d}}$$

*for some positive constant* $A_1$*.*
*(ii) Assume* $|R(s)| = |C\xi^s \cos(as+\theta)|$ *for* $s > s_0$*, where* $C > 0, |\xi| < 1, a \geq \frac{\pi}{K_N}, \theta \in [0, 2\pi]$ *are some constants. Then,*

$$\hat{M} \overset{P}{\sim} A_2 \log N$$

*where* $A_2 = -1/\log|\xi|$*.*
*(iii) If* $R(s) = 0$ *for* $|s| >$ *some* $q$ *(with* $q < s_0 + K_N$*), but* $R(q) \neq 0$*, then*

$$\hat{M} = 2q + o_P(1).$$

### 3.2 Bandwidth selection for second order kernels using flat-top pilots

It is well known that, under regularity conditions on moments and the degree of dependence, equations (1) and (2) hold true; see Györfi et al. (1989) or Bosq (1996). For example, it is a well known fact that if the sequence $\{X_n\}$ is i.i.d., then

$$C_{f,\Lambda}(x) = f(x) \int \Lambda^2(x)dx;$$

cf. Rosenblatt (1991, p. 7). As mentioned before, the variance constant $C_{f,\Lambda}(x)$ remains the *same* in the stationary case subject to some regularity conditions, e.g., a linear time series with weak dependence. Although there is a slight restriction on the allowable bandwidth rates in the dependent case, fortunately this restriction does not affect the optimal rate for the bandwidth which is still $N^{-1/5}$ as in the i.i.d. case; see Assumption 3 of Hallin and Tran (1996).

For the bias term, the arguments (and formulas) are identical as in the spectral density case; for example, for a second order (nonnegative) kernel $\Lambda$ we have:

$$Bias(\hat{f}(x)) = c_{f,\Lambda}(x)h^2 + o(h^2), \quad with \quad c_{f,\Lambda}(x) = -\frac{f''(x)}{2}\int y^2\Lambda(y)dy; \tag{19}$$

Thus, as before, the asymptotically MSE–optimal bandwidth choice (local at the point $x$) for a second order kernel $\Lambda$ is given by:

$$h_{opt} = \left(\frac{C_{f,\Lambda}(x)}{4c^2_{f,\Lambda}(x)\ N}\right)^{1/5}, \tag{20}$$

and our data-dependent approximation to $h_{opt}$ is given by:

$$\hat{h}_{opt} = \left( \frac{\tilde{C}_{f,\Lambda}(x)}{4\tilde{c}^2_{f,\Lambda}(x)\ N} \right)^{1/5}. \tag{21}$$

By complete analogy to the spectral density case, $\tilde{C}_{f,\Lambda}(x)$ and $\tilde{c}_{f,\Lambda}(x)$ are estimators of $C_{f,\Lambda}(x)$ and $c_{f,\Lambda}(x)$ based on flat-top kernel estimators $\tilde{f}_p(x)$ (for $p = 0$ and $p = 2$) in conjunction with our (continuous version) of the empirical rule of picking $M$. Both Theorems from Section 2.2 have extensions in the probability density case in terms of estimating the bandwidth of the second order kernel $\Lambda$. We focus on Theorem 2.6, and give its extension below.

**Theorem 3.3** *Using a flat-top pilot with bandwidth equal to $1/\hat{M}$ as given by our empirical rule of picking $M$ (continuous version), we obtain an estimated bandwidth $\hat{h}_{opt}$ with the following properties.*

*(i) Assume $d > 6$, and the conditions of Theorem 3.2 (i). Then,*

$$\hat{h}_{opt} = h_{opt}(1 + O_P((\frac{\log N}{N})^{(d-5)/(2d)})) \quad \textit{if } d \textit{ is an integer,}$$

*and*

$$\hat{h}_{opt} = h_{opt}(1 + O_P((\frac{\log N}{N})^{[d-4]/(2d)})) \quad \textit{if } d \textit{ is not an integer,}$$

*where $[\cdot]$ denotes the integer part.*
*(ii) Under the conditions of Theorem 3.2 (ii), we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(\frac{\sqrt{\log N}}{\sqrt{N}})).$$

*(iii) Under the conditions of Theorem 3.2 (iii), we have*

$$\hat{h}_{opt} = h_{opt}(1 + O_P(\frac{1}{\sqrt{N}})).$$

**Acknowledgement.** Many thanks are due to Professor Mary Ellen Bock of Purdue University for her suggestion in the early 90s that flat-top kernels behave like wavelets.

# 4 Appendix: Technical proofs

PROOF OF THEOREM 2.1 We give the proof of part (ii); the other parts are proven in the same manner. Observe that under the assumed conditions of part (ii) we have that

$$\sum_{k=-M}^{M} \lambda^T(k/M) e^{iwk} \hat{R}(k) = \sum_{k=-\infty}^{\infty} e^{iwk} R(k) + O_P(\sqrt{\log N}/\sqrt{N});$$

see Politis and Romano (1995). By a calculation similar to the one given in Politis and Romano (1995), we can also show that the bias of $\sum_{k=-M}^{M} \lambda^T(k/M)|k|^p e^{iwk} \hat{R}(k)$ is of order $O(e^{-c_1 M})$ for some $c_1 > 0$. Using eq. (8) and $M \sim A \log N$, it follows that

$$\sum_{k=-M}^{M} \lambda^T(k/M)|k|^p e^{iwk} \hat{R}(k) = \sum_{k=-\infty}^{\infty} |k|^p e^{iwk} R(k) + O_P(\sqrt{\log N}/\sqrt{N}),$$

and part (ii) is proven. □

PROOF OF THEOREM 2.2 We first prove part (iii); we will actually prove that $P(\hat{m} = q) \to 1$. Note that $\hat{m} > q$ only if

$$\max_{k=1,\ldots,K_N} |\hat{\rho}(q+k)| \geq c\sqrt{\log N/N} \tag{22}$$

But by eq. (10) and the assumption of $R(k) = 0$ for $k > q$ we have:

$$\max_{k=1,\ldots,K_N} |\hat{\rho}(q+k)| = O_P(\sqrt{\log K_N/N}) = o_P(\sqrt{\log\log N/N}) \tag{23}$$

since $K_N = o(\log N)$. The probability of (22) and (23) happening simultaneously tends to zero, hence $P(\hat{m} > q) \to 0$.

We will now show that $P(\hat{m} < q) \to 0$, or equivalently that $P(\hat{m} = i) \to 0$ for $i = 1, \ldots, q-1$. Note that $\hat{m} = i (< q)$ only if $|\hat{\rho}(i)| < c\sqrt{\log N/N}$. But $\hat{\rho}(i) = \rho(i) + O_P(\sqrt{1/N})$ where $|\rho(i)| > 0$ by assumption. Combining the above, it follows that $P(\hat{m} = i) \to 0$.

Next we give the proof of part (ii). First note that since $|R(k)| > 0$ for all $k \leq k_0$, it follows that $|R(k)| > \epsilon$ for all $k \leq k_0$ and some $\epsilon > 0$; hence, for large enough $N$, $\max_{k=1,\ldots,k_0} |\hat{R}(k) - R(k)| < \epsilon$ with high probability. Therefore, by eq. (9) it follows that $\hat{m} > k_0$ with high probability, and we can just focus on the part of the correlogram to the right of $k_0$, i.e., look at $\hat{R}(k)$ for $k > k_0$ only.

From part (ii) assumptions together with eq. (10) it follows that

$$|\hat{\rho}(k)| = |C\xi^k| + O_P(\sqrt{\log N/N}) \quad \text{uniformly in } k$$

and consequently, that

$$|\hat{\rho}(\hat{m})| = |C\xi^{\hat{m}}| + O_P(\sqrt{\log N/N}). \tag{24}$$

Recall that our empirical rule of selecting $M$ implies $|\hat{\rho}(\hat{m})| \geq c\sqrt{\log N/N}$ but $|\hat{\rho}(\hat{m}+1)| < c\sqrt{\log N/N}$. Combining the above two statements with eq. (24), and using the fact that $\log\log N/\log N \to 0$, it follows that $\hat{m}$ is bounded above and below by $A_2 \log N$ with probability tending to one, and hence

$$\hat{m} \overset{P}{\sim} A_2 \log N.$$

The proof of part (i) is similar to that of part (ii). □

Proof of Theorem 2.3 Part (iii) is proven exactly like part (iii) of Theorem 2.2. We give the proof of part (ii) below. First note that from the condition that the sequence $R(k)$ does not have more than $K_N - 1$ zeros in its first $k_0$ lags, together with eq. (9), it follows that $\hat{m} > k_0$ with high probability, so again we can focus on the part of the correlogram to the right of $k_0$, i.e., look at $\hat{R}(k)$ for $k > k_0$ only.

Now condition $a \geq \pi/K_N$ implies that $K_N$ is bigger than a half-periof of the cosine $\cos(ak+\theta)$. Hence, by eq. (10) we have:

$$\max_{k=1,\ldots,K_N} |\hat{\rho}(\hat{m}+k)| \geq C|\xi|^{\hat{m}+K_N} + O_P(\sqrt{\log N/N}). \tag{25}$$

But our empirical rule for picking $M$ implies that $|\hat{\rho}(\hat{m})| \geq c\sqrt{\log N/N}$, whereas

$$\max_{k=1,\ldots,K_N} |\hat{\rho}(\hat{m}+k)| < c\sqrt{\log N/N}.$$

The above two statements, combined with eq. (24) and (25), and the arguments given in the proof of part (ii) of Theorem 2.2, imply that

$$A_2 \log N - K_N \leq \hat{m} \leq A_2 \log N$$

with high probability. Since, by assumption, $K_N = o(\log N)$, the proof of part (ii) is completed. The proof of part (i) is similar. □

Proof of Theorem 2.4 First note that from the condition that the sequence $R(k)$ does not have more than $K_N - 1$ zeros in its first $[A_0 \log N]$ lags, together with eq. (10),

it follows that $\hat{m} > [A_0 \log N]$ with high probability, so we can focus on the part of the correlogram to the right of $[A_0 \log N]$, i.e., look at $\hat{R}(k)$ for $k > [A_0 \log N]$ only.

The rest of the proof follows exactly like the proof of Theorem 2.3 (ii), together with the observation that, under the assumed conditions,

$$\epsilon_k = o(\xi^k) = o(\xi^{A_0 \log N}) = o(1/\sqrt{N}),$$

for all $k > A_0 \log N$. □

PROOF OF THEOREM 2.5 Theorem 2.5 follows from Theorems 2.1, 2.2, 2.3, and the $\delta$–method. □

PROOF OF THEOREM 2.6 First note that all results of Theorem 2.5 remain valid if, in place of $M$, we use the random (but asymptotically equivalent) quantity $M(1 + o_P(1))$; this can be proven along the same lines as the proof of Lemma 2 in Bühlmann (1996). Therefore, Theorem 2.6 immediately follows given the asymptotic rates achieved by $\tilde{M}$ under the different dependence conditions.

Regarding case (i), we should elaborate a bit: assuming $R(k) = O(1/k^d)$ it follows that $f_2(w)$ has at least $J$ continuous derivatives where $J = [d - 4]$ if $d$ is not an integer, and $J = d - 5$ if $d$ is an integer. Thus, by an expansion of the type of eq. (2), the bias of $\tilde{f}_2(w)$ will be of order $O(1/M^J)$. In addition, $f(w)$ will have at least $J + 2$ continuous derivatives so the bias of $\tilde{f}(w) = O(1/M^{J+2}) = O(1/M^J)$. Recalling that $Var(\tilde{f}_i(w)) = O(M/N)$ for $i = 0$ or 2, and substituting the rate $N^{1/(2d)}/(\log N)^{1/(2d)}$ achieved by $\hat{M}$ on a set whose probability tends to one, completes the calculation of part (i). □

PROOF OF THEOREM 3.1 The proof of Theorem 3.1 is similar to the proof of Theorem 2.1. □

## References


[1] Abramson, I. (1982), Arbitrariness of the pilot estimator in adaptive kernel methods, *J. Multiv. Anal.*, 12, 562-567.

[2] Adams, T.M. and Nobel, A.B. (1998). On density estimation from an ergodic process, *Annals of Probability*, 26, 794-804.

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

[4] Brockmann, M., Gasser, T., and Herrmann, E. (1993), Locally adaptive bandwidth choice for kernel regression estimators, *J. Amer. Statist. Assoc.*, 88, 1302-1309.

[5] Bühlmann, P. (1996), Locally adaptive lag-window spectral estimation, *J. Time Ser. Anal.*, 17, 247-270.

[6] Bosq, D. (1996), *Nonparametric statistics for stochastic processes*, Lecture Notes in Statistics No. 110, Springer, New York.

[7] Devroye, L. (1987), *A course in density estimation*, Birkhäuser, Boston.

[8] Devroye, L. (1992), A note on the usefulness of superkernels in density estimation, *Ann. Statist.,* vol. 20, no. 4, pp. 2037-2056.

[9] Devroye, L. and Györfi, L. (1984). *Nonparametric Density Estimation: The L1 View*, John Wiley, New York.

[10] Donoho, D.L. and Johnstone, I.M. (1994), Ideal spatial adaptation via wavelet shrinkage, *Biometrika*, 81, 425-455.

[11] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996), Density estimation by wavelet thresholding, *Ann. Statist.*, 24, 508-539.

[12] Gasser, T., Müller, H.G. and Mammitzsch, V. (1985), Kernels for nonparametric curve estimation, *J. Roy. Statist. Soc. B*, vol. 47, pp. 238-252.

[13] Granovsky, B.L. and Müller, H.G.(1991), Optimal kernel methods: A unifying variational principle, *Internat. Statist. Review*, vol. 59, no. 3, pp. 373-388.

[14] Hall, P. (1983), Large sample optimality of least squares cross-validation in density estimation, *Ann. Statist.*, 11, 1156-1174.

[15] Hall, P., Sheather, S.J., Jones, M.C., and Marron, J.S. (1991), On optimal data-based bandwidth selection in kernel density estimation, *Biometrika*, 78, 263-270.

[16] Hallin, M. and Tran, L.T. (1996), Kernel density estimation for linear processes: asymptotic normality and bandwidth selection, *Annals of the Institute of Statistical Mathematics*, 48, 429-449.

[17] Härdle, W., Kerkyacharian, G, Picard, D. and Tsybakov, A. (1998), *Wavelets, approximation, and statistical applications*, Lecture Notes in Statistics No. 129, Springer: New York.

[18] Hurvich, C. (1985), Data-driven choice of a spectrum estimate: extending the applicability of cross-validation methods, *J. Amer. Statist. Assoc.*, 80, 933-940.

[19] Hurvich, C. and Beltrao, K.I. (1990), Cross-validatory choice of a spectrum estimate and its connection with aic, *J. Time Ser. Anal.*, 11, 121-137.

[20] Györfi, L., Härdle, W., Sarda, P., and Vieu, P. (1989). *Nonparametric Curve Estimation from Time Series*, Lecture Notes in Statistics **60**, Springer-Verlag, New York.

[21] Jones, M.C. (1995), On higher order kernels, *J. Nonparametr. Statist.*, vol. 5 , 215-221.

[22] Jones, M.C., Marron, J.S., and Sheather, S.J. (1996), A brief survey of bandwidth selection for density estimation, *J. Amer. Statist. Assoc.*, vol. 91 , 401-407.

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

[24] Loader, C. (1999a). Bandwidth selection: classical or plug-in?, *Annals of Statistics*, **27**, 415–438.

[25] Loader, C. (1999b). *Local Regression and Likelihood*, Springer, New York.

[26] Marron, J.S. (1994), Visual understanding of higher order kernels, *J. Comput. Graphical Statist.*, vol.3, 447-458.

[27] Marron, J.S. and Wand, M.P. (1992), Exact mean integrated squared error, *Ann. Statist.,* vol. 20, 712-736.

[28] Müller, H.G. (1988), *Nonparametric regression analysis of longitudinal data*, Springer-Verlag, Berlin.

[29] 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.

[30] Politis, D.N., and Romano, J.P. (1993), On a Family of Smoothing Kernels of Infinite Order, in *Computing Science and Statistics, Proceedings of the 25th Symposium on the Interface*, San Diego, California, April 14-17, 1993, (M. Tarter and M. Lock, eds.), The Interface Foundation of North America, pp. 141-145.

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

[32] Politis, D.N., and Romano, J.P. (1996), On Flat-top Kernel Spectral Density Estimators for Homogeneous Random Fields, *J. Statist. Plan. Infer.*, vol. 51, 1996, pp. 41-53.

[33] Politis, D.N. and Romano, J.P. (1999), Multivariate density estimation with general flat-top kernels of infinite order, *J. Multivar. Anal.*, vol. 68, 1-25.

[34] Politis, D.N., Romano, J.P. and Wolf, M. (1999), *Subsampling*, Springer: New York.

[35] Priestley, M.B. (1981), *Spectral Analysis and Time Series*, Academic Press.

[36] Rachdi, M. (1998), Choix optimal de la fênetre spectrale pour un processus stationnaire à temps discret $\alpha$-mélangeant, *C. R. Acad. Sci. Paris*, t. 327, Série I, p. 405-408.

[37] Rachdi, M. and Youndjé, E. (2000), Asymptotic optimal spectral bandwidth choice for a discrete stationary process, Document 2000-5, Université de Rouen UFR des sciences mathematiques.

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

[39] Rosenblatt, M. (1956), A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, **42**, 43–47.

[40] Rosenblatt, M. (1991), *Stochastic Curve Estimation*, NSF-CBMS Regional Conference Series vol. 3, Institute of Mathematical Statistics, Hayward.

[41] Scott, D. W. (1992), *Multivariate density estimation: theory, practice, and visualization*, Wiley, New York.

[42] Woodroofe, M. (1970), On choosing a delta sequence, *Ann. Math. Statist.*, 41, 1665-1671.