Aggregation of Spectral Density Estimators

Christopher Chang

Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112, USA; chrchang@alumni.caltech.edu

Dimitris Politis

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

Abstract

Given stationary time series data, we study the problem of finding the best linear combination of a set of lag window spectral density estimators with respect to the mean squared risk. We present an aggregation procedure and prove a sharp oracle inequality for its risk. We also provide simulations demonstrating the performance of our aggregation procedure, given Bartlett and other estimators of varying bandwidths as input. This extends work by P. Rigollet and A. Tsybakov on aggregation of density estimators.

Keywords: Estimation, Spectral analysis, Time series analysis

Introduction

Consider stationary time series data $X_1, \ldots, X_n$ having mean zero and spectral density

$$p(\lambda) := \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma(j) e^{-i\lambda j}, \lambda \in [-\pi, \pi) \tag{1}$$

where $\gamma(k)$ is the autocovariance at lag k. For an estimator $\hat{p}(X_1, \ldots, X_n)$ of p, define the L_2-risk

$$R_n(\hat{p}, p) = E\left[\int_{-\pi}^{\pi} (\hat{p}(x) - p(x))^2 \, dx\right]. \tag{2}$$

Let $\hat{p}_1, \ldots, \hat{p}_J$ be a collection of lag window (i.e. kernel) spectral density estimators of p—see eq. (5) for a precise definition. We investigate the construction of a new estimator $\hat{p}_n^L$ which is asymptotically as good, in terms of L_2-risk, as using the best possible linear combination of $\hat{p}_1, \ldots, \hat{p}_J$; more precisely, $\hat{p}_n^L$ satisfies the oracle inequality

$$R_n(\hat{p}_n^L, p) \leq \inf_{\lambda \in \mathbb{R}^J} R_n(\sum_{j=1}^{J} \lambda_j \hat{p}_j, p) + \Delta_{n,J} \tag{3}$$

where $\Delta_{n,J}$ is a small remainder term independent of p.

Such an estimator may find a variety of applications. For instance, one may try to bypass the difficult issue of bandwidth (or model) selection by setting the $\hat{p}$s to cover a wide spread of possibly reasonable bandwidths/models. Alternatively, when a linear combination of kernels outperforms all the individual inputs, e.g. when the $\hat{p}$s are Bartlett windows as in Politis and Romano (1995), our aggregating estimator is capable of discovering it.

Kernel density estimation dates back to Rosenblatt (1956) and Parzen (1962); Priestley (1981) and Brillinger (1981) discuss its application to spectral densities. More recently, Yang (2000) and Rigollet and Tsybakov (2007) analyzed aggregation of probability density estimators, while Wang et al. (2011) studied

the related problem of linear aggregation in nonparametric regression. We extend Rigollet and Tsybakov's (2007) work to spectral estimation.

To perform aggregation, we use a sample splitting scheme. The time series data is divided into a training set, a buffer zone, and a validation set; with an exponential mixing rate, the buffer zone need not be more than logarithmic in the size of the other sets to ensure approximate independence between the training and validation sets.

The estimator, and theoretical results concerning its performance, are presented in section 2. Simulation studies are conducted in section 3, and our conclusions are stated in section 4.

1. Theoretical Results

1.1. Aggregation Procedure

Split the time series $X_1, \ldots, X_n$ into a training set $X_1, \ldots, X_{n_t}$, a buffer zone $X_{n_t+1}, \ldots, X_{n_t+n_b}$, and a validation set $X_{n_t+n_b+1}, \ldots, X_{n_t+n_b+n_v}$, where the first and third sets can be treated as independent. We investigate appropriate choices of n_t, n_b, and n_v at the end of this section.

With the training set, we produce an initial estimate

$$\hat{\gamma_1}(k) := \frac{1}{n_t} \sum_{j=1}^{n_t-k} X_{j+k} X_j \tag{4}$$

of the autocovariance function, after centering the data. (In practice, the data will be centered to the sample mean rather than the true mean, but the resulting discrepancy is asymptotically negligible w.r.t. autocovariance and spectral density estimation. So, for simplicity of presentation, we center at the true mean above.)

We then propose the following candidate estimators:

$$p_j(\lambda) := \frac{1}{\sqrt{2\pi}} \sum_{k=-b_j}^{b_j} \hat{\gamma}_1(k) \cdot w_j\left(\frac{k}{b_j}\right) \frac{e^{ik\lambda}}{\sqrt{2\pi}} \tag{5}$$

where the b_js ($j = 1, \ldots, J$) are candidate bandwidths arrived at via some selection procedure, and the w_js ($j = 1, \ldots, J$) are lag windows with $w_j(0) = 1$, $w_j(x) \leq 1$ for $x \in (-1, 1)$, and $w_j(x) = 0$ for $|x| \geq 1$ for all j. The p_js have some linear span $\mathcal{L}$ in L_2 whose dimension is denoted by M where $M \leq J$. Now construct an orthonormal basis $\{\phi_j\}$ ($j = 1, \ldots, M$), and note that the ϕ_js are–by necessity–trigonometric polynomials of degree at most $b := \max_j b_j$, i.e.,

$$\phi_j(\lambda) = \sum_{k=-b}^{b} a_{j,k} \frac{e^{ik\lambda}}{\sqrt{2\pi}} \tag{6}$$

so the coefficient $a_{j,k}$ is the inner product of ϕ_j and $\frac{e^{ik\lambda}}{\sqrt{2\pi}}$ in L_2.

Then, based on our validation set, we produce a different estimate of the autocovariance function, namely

$$\hat{\gamma_2}(k) := \frac{1}{n_v} \sum_{j=1}^{n_v-k} X_{n_t+n_b+j+k} X_{n_t+n_b+j} \tag{7}$$

and compute the coefficients

$$\hat{K_j} := \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \hat{\gamma_2}(k) a_{j,k} \tag{8}$$

Finally, our proposed aggregate estimator of the spectral density is given by

$$\hat{p}(\lambda) := \sum_{j}^{M} \hat{K}_j \phi_j(\lambda). \tag{9}$$

1.2. Performance Bounds

We start with the simplest mixing assumption, m-dependence (i.e. for all positive integers j and k where $k \geq m$, X_j and X_{j+k} are independent).

Theorem 1.1. *If $\frac{b}{n} \to 0$, $EX_t^4 < \infty$, and the time series satisfies m-dependence, the L_2 risk is bounded above as follows:*

$$\begin{aligned} R_n(\hat{p}, p) \leq \min_{c_1,\ldots,c_J} ||\sum_{j=1}^{J} c_j p_j - p||^2 + \frac{bp^2(0)M}{n_v \pi} \\ + o(bM/n_v), \end{aligned} \tag{10}$$

where p is the true spectral density and $||\cdot||$ denotes the L_2 norm $(\int_{-\pi}^{\pi} (\cdot(x))^2\, dx)^{1/2}$.

Proof: Projecting p onto $\mathcal{L}$, we get $p_{\mathcal{L}}^* := \sum_{j=1}^{M} K_j^* \phi_j$, where K_j^* is the scalar product of p and ϕ_j in L_2. Then, by the Pythagorean theorem, we have

$$||\hat{p} - p||^2 = \sum_{j=1}^{M} (\hat{K}_j - K_j^*)^2 + ||p_{\mathcal{L}}^* - p||^2. \tag{11}$$

Next, we have $E[\hat{K}_j] = \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} E[\hat{\gamma}_2(k) a_{j,k}]$. Under m-dependence, the size-n_b buffer zone is sufficient to make all the $\hat{\gamma}_2(k)$s (functions only of the validation set) independent of the $a_{j,k}$s (functions only of the training set), so

$$E[\hat{K}_j] = \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} E[\hat{\gamma}_2(k)] E[a_{j,k}] = \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \left(1 - \frac{|k|}{n_v}\right) \gamma(k) a_{j,k} \tag{12}$$

Now, $p(\lambda) = \frac{1}{\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \gamma(k) \frac{e^{ik\lambda}}{\sqrt{2\pi}}$, so

$$E[K_j^*] = E[\langle p, \phi_j \rangle] = \frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \gamma(k) a_{j,k} \tag{13}$$

Then,

$$\begin{aligned} E[(\hat{K}_j - K_j^*)^2] &= \mathrm{Var}[\hat{K}_j] + (\mathrm{Bias}[\hat{K}_j])^2 = \mathrm{Var}\left[\frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \hat{\gamma}_2(k) a_{j,k}\right] \\ &= \left(\frac{1}{\sqrt{2\pi}} \sum_{k=-b}^{b} \frac{|k|}{n_v} \gamma(k) a_{j,k}\right)^2 = \frac{1}{2\pi} \mathrm{Var}\left[\sum_{k=-b}^{b} \hat{\gamma}_2(k) a_{j,k}\right] + \frac{2}{n_v^2 \pi} \left(\sum_{k=1}^{b} k \gamma(k) a_{j,k}\right)^2 \end{aligned} \tag{14}$$

Note that $\hat{K}_j$ can be seen as a lag window spectral density estimator at $\lambda = 0$, except the kernel function is allowed to be negative and doesn't necessarily evaluate to 1 at zero. Parzen's (1957) formula for the variance of such an estimator does not require nonnegativity of the kernel function, but does require that

it be normalized to $K(0) = 1$; we can fix the latter by replacing $a_{j,k}$ with $\frac{a_{j,k}}{a_{j,0}}$ and then multiplying the resulting formulaic variance by $a_{j,0}^2$. (This just cancels out.) As an asymptotic result, it also requires that the kernel function be continuous rather than discrete, so we interpolate $a_{j,k+x} = (1-x)a_{j,k} + xa_{j,k+1}$ for $0 < x < 1$. Then, applying Parzen's (1957) formula,

$$\operatorname{Var}\left[\sum_{k=-b}^{b} \hat{\gamma_2}(k)a_{j,k}\right] = \left[\frac{2a_{j,0}^2 b}{n_v}p^2(0)\int_{-\infty}^{\infty}\frac{a_{j,k}^2}{a_{j,0}^2}\,dk\right] + o(b/n_v) \tag{15}$$

and plugging this into (14),

$$E[(\hat{K_j} - K_j^*)^2] = \frac{b}{n_v\pi}p^2(0)\int_{-\infty}^{\infty}a_{j,k}^2\,dk + \frac{2}{n_v^2\pi}\left(\sum_{k=1}^{b}k\gamma(k)\right)^2 + o(b/n_v). \tag{16}$$

$\sum_{k=-b}^{b} a_{j,k}^2 = 1$, so, by convexity of x^2, the integral is bounded above by 1. The square of the bias can be absorbed into the $o(b/n_v)$ term. We conclude that

$$E[||\hat{p} - p||^2] \leq \min_{\hat{K_1},\ldots,\hat{K_M}}\left|\left|\sum_{j=1}^{M}\hat{K_j}p_j - p\right|\right|^2 + \frac{bp^2(0)M}{n_v\pi} + o(bM/n_v). \quad \square \tag{17}$$

Next, we consider the exponential mixing. Defining $\alpha(\cdot)$ as in Definition A.0.1 in Politis et al. (1999), we have:

Theorem 1.2. *If $\frac{b}{n} \to 0$, $EX_t^4 < \infty$, the time series satisfies the α-mixing assumption $\alpha(k) \leq c^k$ for some constant $c > 1$ and all $k \geq n_b$, and n_b is chosen such that $n_b \geq (2+\epsilon)\log_c n$ for some $\epsilon > 0$, the L_2 risk of our estimator has the same upper bound as in Theorem 1.1.*

Proof: We wish for the dependence between the $\hat{\gamma_2}$'s and the $a_{j,k}$'s to have an impact of order $o(b/n)$ on $||\hat{p} - p||^2 - \min\left|\left|\sum_{j=1}^{M}\hat{K_j}p_j - p\right|\right|^2$.

By Lemma A.0.1 in Politis et al. (1999), with $\xi = \hat{\gamma_2}(k), \zeta = a_{j,k}, p = 2$, and $q = \infty$, we have

$$|\operatorname{Cov}(\hat{\gamma_2}(k), a_{j,k})| \leq 8(E|\hat{\gamma_2}|^2)^{1/2} \cdot 1 \cdot \sqrt{\alpha(n_b)} \tag{18}$$

since $|a_{j,k}| \leq 1$ (because, by construction of the orthonormal basis, $\sum_j a_{j,k}^2 = 1$);

$$\leq 8\sqrt{\frac{(n_v-k)^2}{n_v^2}\gamma^2(k) + \operatorname{Var}\hat{\gamma}_2(k)}\sqrt{\alpha(n_b)} = \Theta(8\gamma(k)\sqrt{\alpha(n_b)}) = \Theta(8\gamma(k)c^{-n_b/2}) \tag{19}$$

where Θ is Bachmann-Landau notation denoting an asymptotic lower and upper bound.

Plugging this back into $E[\hat{K_j}]$, we get an additional term with absolute value bounded by $\Theta\left(\frac{1}{\sqrt{2\pi}}\sum_{k=-b}^{b}8\gamma(k)c^{-n_b/2}\right)$. Since we chose $n_b \geq (2+\epsilon)\log_c n$, $c^{-n_b/2} \leq n^{-1-(\epsilon/2)}$ so the term's impact on $E[\hat{K_j}]$ is $o(b/n)$. Thus, its impact on $E[(\hat{K_j} - K_j^*)^2]$ is also $o(b/n)$ as desired. $\square$

Theorem 1.3. *If $\frac{b}{n} \to 0$, $EX_t^4 < \infty$, the time series satisfies the α-mixing assumption $\alpha(k) = O(k^{-c})$ for all $k \geq n_b$ and some $c > 2$, and n_b is chosen such that $n_b \geq n^{\frac{2}{c}+\epsilon}$ for some $\epsilon > 0$, the L_2 risk of our estimator has the same upper bound as in Theorem 1.1.*

Proof: The proof is identical to that of Theorem 1.2 up to (19). Plugging (19) into $E[\hat{K}_j]$, we get an additional term with absolute value bounded by $O\left(\frac{1}{\sqrt{2\pi}}\sum_{k=-b}^{b} 8\gamma(k)n_b^{-c/2}\right)$. Since we chose $n_b \geq n^{\frac{2}{c}+\epsilon}$, the term's impact on $E[\hat{K}_j]$ is $o(b/n)$, and the result follows. □

Remark. If $\gamma(k)$ decays at only a polynomial rate, Theorem 3.1 from Politis (2011) is only able to bound $\min_{c_1,\ldots,c_M} \left|\left|\sum_{j=1}^{M} c_j p_j - p\right|\right|^2$ by a term of order $n_t^{\frac{1}{2r+1}-1}$ where $r \geq 1$ satisfies $\sum_{k=1}^{\infty} k^r\gamma(k) < \infty$. In this case, when the bandwidth candidates are of smaller order than $n_v^{\frac{1}{2r+1}}$, n_v should be larger than n_t.

However, if $\gamma(k)$ decays at least exponentially, the same theorem offers a bound of $O\left(\frac{\log n_t}{n_t}\right)$. In this case, if the bandwidth candidates increase more than logarithmically in n_v, we'll want to choose $n_v > n_t$.

2. Simulation Results

The Bartlett kernel is defined by

$$w(x) = \begin{cases} 1-|x| & \text{for } |x|<1; \\ 0 & \text{elsewhere} \end{cases} \tag{20}$$

In the following simulations, we aggregate the estimators

$$p_j(\lambda) = \frac{1}{\sqrt{2\pi}} \sum_{k=-b_j}^{b_j} \hat{\gamma}_1(k) w(\frac{k}{b_j}) \frac{e^{ik\lambda}}{\sqrt{2\pi}}, \tag{21}$$

for various collections of b_js.

Let $\{Z_t\} \sim IID(0,\sigma^2)$. The MA(1) model $X_t = Z_t + \theta Z_{t-1}$ then has autocovariances $\gamma(0) = (1+\theta^2)\sigma^2$, $\gamma(1) = \theta\sigma^2$, and $\gamma(k) = 0$ for $k > 1$. From Politis (2003), the large-sample optimal block size is $(6n)^{1/3}\left|\frac{\sum_{k=1}^{\infty} k\gamma(k)}{\sum_{k=-\infty}^{\infty}\gamma(k)}\right|^{2/3}$, which evaluates to $(6n)^{1/3}\left|\frac{\theta}{(1+\theta)^2}\right|^{2/3}$ in the MA(1) case. Most of our simulations use $\theta = 0.5$, for which this reduces to $\frac{2n^{1/3}}{3}$.

In the tables below, "length" denotes the length of the time series, the b_js in the aggregate are listed under "bandwidth", "avg. $\hat{K}$" denotes the average weight assigned by the aggregate to the bandwidth, and "MISE" is the empirical mean integrated square error (MISE) of the kernel spectral density estimate. All values are averages over 200 trials, except for the length 125k time series (for which only 50 trials were averaged); simulation standard errors are in parentheses.

Length	Bandwidth	Avg. $\hat{K}$	MISE
100	3	.8391	.015955 (.000975)
	12	.2250	.028312 (.001675)
	agg.		.022932 (.001518)
500	5	.9608	.004253 (.000275)
	20	.0895	.008967 (.000360)
	agg.		.006289 (.000368)
1000	7	.9936	.002756 (.000171)
	28	.0437	.006518 (.000271)
	agg.		.003739 (.000177)
27000	20	.9869	.000274 (.000009)
	80	.0266	.000654 (.000014)
	agg.		.000293 (.000009)
125000	33	.9778	.000099 (.000005)
	133	.0280	.000237 (.000008)
	agg.		.000102 (.000004)
1000	6	-.5530	.002717 (.000155)
	7	1.5914	.002622 (.000153)
	agg.		.003234 (.000163)
1000	7	.9195	.002787 (.000138)
	14	.1186	.003483 (.000149)
	agg.		.003642 (.000173)
1000	7	1.0387	.002985 (.000149)
	50	-.0028	.011457 (.000341)
	agg.		.003721 (.000200)

Table 1: MA(1) $\theta = 0.5$ Bartlett aggregation results, optimal bandwidth with single alternative.

We first tried aggregations of two bandwidths, with one roughly optimal and the other much larger. Theoretically, we expect the optimal linear combination to basically ignore the second bandwidth, and this is what our aggregates tended towards doing. However, for smaller sample sizes, the lone inefficient alternative raised the MISE by close to 50%. This penalty was reduced to 5-10% once the sample size reached the tens of thousands; see blocks 1–5 of Table 1.

We then tried varying the alternative bandwidth; see blocks 6–8. There was no noticeable difference between the 2x optimal and 7x optimal alternatives. However, if the second bandwidth was instead a near-duplicate of the first, the MISE penalty was found lower. Of course, there would be little potential gain from aggregation in that case.

Length	Bandwidth	Avg. $\hat{K}$	MISE
100	2	-.7411	.019814 (.001053)
	3	.9637	.016115 (.001117)
	5	.8571	.017252 (.001185)
	agg.		.026471 (.001758)
100	2	-1.3568	.021141 (.000866)
	3	3.0841	.017913 (.000931)
	5	-1.1890	.019320 (.001043)
	8	.5268	.024443 (.001247)
	agg.		.031413 (.001734)
1000	3	-1.4652	.005982 (.000199)
	5	3.0790	.003430 (.000175)
	10	-.6013	.003141 (.000169)
	agg.		.003696 (.000206)
1000	4	-.6085	.004167 (.000179)
	7	1.7816	.002973 (.000157)
	14	-.1436	.003627 (.000171)
	agg.		.003801 (.000180)
1000	5	.2156	.003218 (.000192)
	10	.9896	.003148 (.000173)
	20	-.1754	.005082 (.000185)
	agg.		.004350 (.000186)
1000	7	.7596	.002926 (.000174)
	14	.5107	.003740 (.000177)
	28	-.2355	.006496 (.000223)
	agg.		.004669 (.000188)
1000	3	.1063	.005605 (.000199)
	12	1.0519	.003285 (.000144)
	48	-.1280	.010856 (.000245)
	agg.		.004144 (.000211)
1000	5	.1666	.003345 (.000141)
	10	.9547	.003253 (.000129)
	20	-.1622	.005148 (.000161)
	40	.0678	.009356 (.000244)
	agg.		.005392 (.000201)
27000	10	-.3527	.000480 (.000014)
	20	1.5431	.000255 (.000010)
	40	-.1843	.000340 (.000010)
	agg.		.000289 (.000010)
27000	10	-.3709	.000510 (.000016)
	20	1.6316	.000289 (.000012)
	40	-.2834	.000377 (.000012)
	80	.3200	.000683 (.000013)
	agg.		.000338 (.000012)

Table 2: MA(1) $\theta = 0.5$ Bartlett aggregation results, geometric bandwidth spreads.

We then tried increasing the number of aggregate components, with geometric spreads of bandwidths. As expected, the MISE penalty was roughly linear in the number of components, and was more acceptable with larger sample sizes; see Table 2.

It did not really matter whether the aggregate included a near-optimal component; the (3, 5, 10) aggregate outperformed the (4, 7, 14) aggregate and the (3, 12, 48) aggregate noticeably outperformed the (7, 14, 28) aggregate for length 1k time series, despite the fact that the optimal bandwidth was about 7.

Length	Bandwidth	Avg. $\hat{K}$	MISE
100	1	-.6542	.046679 (.000697)
	3	1.7287	.016935 (.000977)
	agg.		.018653 (.001105)
500	1	-.2461	.041030 (.000121)
	5	1.2560	.004652 (.000254)
	agg.		.003629 (.000292)
1000	1	-.1461	.040431 (.000074)
	7	1.1477	.002836 (.000146)
	agg.		.002472 (.000119)
27000	1	-.0542	.039811 (.000003)
	20	1.0544	.000283 (.000010)
	agg.		.000185 (.000006)
27000	2	-.0848	.009934 (.000035)
	20	1.0842	.000269 (.000009)
	agg.		.000212 (.000008)
27000	3	-.1285	.004485 (.000032)
	20	1.1316	.000293 (.000010)
	agg.		.000228 (.000009)
125000	1	-.0298	.039795 (.000001)
	33	1.0311	.000096 (.000006)
	agg.		.000059 (.000004)
125000	2	-.0528	.009884 (.000036)
	33	1.0503	.000092 (.000006)
	agg.		.000069 (.000004)
125000	3	-.0901	.004471 (.000029)
	33	1.0915	.000101 (.000005)
	agg.		.000073 (.000003)
125000	1	-.0240	.039793 (.000001)
	40	1.0227	.000096 (.000005)
	agg.		.000077 (.000005)
125000	3	-.0516	.004406 (.000030)
	40	1.0527	.000102 (.000005)
	agg.		.000089 (.000004)
125000	5	-.0825	.001601 (.000021)
	40	1.0852	.000098 (.000004)
	agg.		.000088 (.000004)
125000	1	-.0143	.039795 (.000001)
	60	1.0158	.000124 (.000005)
	agg.		.000115 (.000005)
125000	3	-.0261	.004442 (.000030)
	60	1.0283	.000120 (.000005)
	agg.		.000117 (.000005)
125000	5	-.0141	.001631 (.000024)
	60	1.0187	.000119 (.000004)
	agg.		.000119 (.000004)

Table 3: MA(1) $\theta = 0.5$ Bartlett aggregation results, two-bandwidth trapezoid discovery simulations.

In the theory of kernel spectral estimation, the so-called 'flat-top' lag windows have been shown to have very favorable asymptotic and finite-sample properties, especially when the autocovariance decays quite rapidly. The simplest flat-top lag-window is the trapezoid proposed by Politis and Romano (1995); for the definition and properties of general flat-top lag windows see Politis (2001, 2005 and 2011).

Since the trapezoid can be constructed as a linear combination of two triangular (Bartlett) kernels, we wanted to investigate the conditions under which conditions the aggregate estimator would tend to approximate a trapezoid. Note, however, that the aggregate estimator shoots for minimum MISE, and the

flat-top estimators only achieve optimal performance when their bandwidth is chosen to be sufficiently small. Hence, in Table 3 we investigate our aggregate's ability to outperform its near-optimal bandwidth component when a very low bandwidth component is also provided.

Indeed, the weight assignments chosen by the aggregate are trapezoid approximations, and the aggregate is able to achieve a MISE advantage of 20% with sample sizes in the hundreds, which rises to close to 40% in the 125k sample size case. However, the trapezoid's advantage appears to vanish as soon as the primary bandwidth reaches 2x optimal.

Length	Bandwidth	Avg. $\hat{K}$	MISE
100	1	-.7459	.046781 (.000725)
	2	4.5984	.022110 (.000866)
	3	-11.690	.018185 (.000945)
	4	8.9482	.017763 (.001020)
	agg.		.024602 (.001790)
500	1	-1.0408	.041107 (.000125)
	2	1.8388	.001223 (.000296)
	4	.5200	.005488 (.000264)
	8	-.3026	.005160 (.000290)
	agg.		.006807 (.000373)
4000	1	-.6016	.039950 (.000016)
	3	1.9790	.004779 (.000087)
	7	-.4626	.001348 (.000054)
	15	.0903	.001122 (.000047)
	agg.		.000830 (.000045)
27000	1	-.3707	.039813 (.000003)
	4	1.5363	.002495 (.000030)
	15	-.1786	.000299 (.000011)
	50	.0127	.000437 (.000011)
	agg.		.000135 (.000006)
125000	1	-.2439	.039796 (.000001)
	5	1.2143	.001604 (.000021)
	25	.0287	.000115 (.000006)
	125	.0036	.000230 (.000007)
	agg.		.000027 (.000004)
4000	1	-.4991	.039954 (.000015)
	3	1.4966	.004734 (.000086)
	agg.		.000386 (.000038)

Table 4: Geometric bandwidth spreads starting at 1.

The particularly favorable performance of the aggregates including a bandwidth 1 component in the last batch of simulations suggested that geometric bandwidth spreads starting from 1 might significantly outperform the spreads investigated in Table 2. This is in fact the case; see Table 4. While previously the aggregate did not outperform the best individual component even with a length 27k time series, now we see outperformance at length 4k, and by 27k it is by more than a factor of 2. Note that, in the length 4k case, the two additional bandwidths roughly double the MISE compared to the simple trapezoid aggregate, but the procedure would still be worthwhile if one was not aware of the value of using trapezoidal kernels directly.

Length	Bandwidth	Avg. $\hat{K}$	MISE
100	1	.1070	.047035 (.000502)
	2	-19.102	.014780 (.000806)
	3	66.422	.015464 (.001022)
	4	-46.390	.018206 (.001225)
	agg.		.024315 (.002296)
500	1	-.4764	.040901 (.000109)
	2	1.8497	.004509 (.000274)
	4	-.2803	.003498 (.000240)
	8	-.0877	.006134 (.000332)
	agg.		.005989 (.000337)
4000	1	-.1417	.039942 (.000017)
	3	1.3346	.000836 (.000046)
	7	-.1769	.000726 (.000041)
	15	-.0177	.001411 (.000055)
	agg.		.001004 (.000048)
27000	1	-.0707	.039811 (.000002)
	4	1.1460	.000219 (.000010)
	15	-.0823	.000199 (.000009)
	50	.0074	.000658 (.000015)
	agg.		.000161 (.000008)
125000	1	-.0471	.039794 (.000001)
	5	1.2220	.000084 (.000006)
	25	-.1909	.000067 (.000005)
	125	.0151	.000349 (.000011)
	agg.		.000037 (.000003)

Table 5: Aggregation of Epanechnikov-Priestley kernels, MA(1) time series.

We also tried aggregating estimators using the optimal (among second order kernels) Epanechnikov-Priestley kernel, namely

$$w(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{for } |x| < 1; \\ 0 & \text{elsewhere.} \end{cases} \tag{22}$$

There is no exact result involving linear combinations of these kernels that is analogous to the relation between trapezoidal and Bartlett kernels. However, for the largest sample sizes our aggregate was able to significantly outperform all the individual components, and across all sample sizes the aggregate never had MISE bigger than twice the best individual component—see Table 5.

Finally, we reran the last two simulations with AR(1) instead of MA(1) time series, with similar results—see Tables 6-7.

Length	Bandwidth	Avg. $\hat{K}$	MISE
100	1	-13.7843	.102601 (.000768)
	2	50.5226	.056486 (.001362)
	3	-32.9616	.041702 (.001642)
	4	-2.7238	.036646 (.001776)
	agg.		.047189 (.003437)
500	1	.0788	.096973 (.000239)
	2	-1.0643	.044617 (.000554)
	4	1.4626	.018105 (.000703)
	8	.5251	.012500 (.000795)
	agg.		.012815 (.000993)
4000	1	.0161	.094792 (.000019)
	3	-.7424	.022092 (.000210)
	7	1.4975	.005092 (.000184)
	15	.2279	.002384 (.000126)
	agg.		.001691 (.000120)
27000	1	.0020	.094598 (.000004)
	4	-.4071	.012932 (.000096)
	15	1.5566	.001182 (.000050)
	50	-.1517	.000739 (.000029)
	agg.		.000420 (.000020)
125000	1	-.0007	.094567 (.000001)
	5	-.2547	.008363 (.000081)
	25	1.2785	.000398 (.000026)
	125	-.0218	.000324 (.000013)
	agg.		.000130 (.000012)
4000	1	-.5904	.094800 (.000025)
	3	1.5906	.022217 (.000223)
	agg.		.010948 (.000087)

Table 6: AR(1) ($\rho = 0.5$) time series Bartlett aggregation results.

Length	Bandwidth	Avg. $\hat{K}$	MISE
100	1	.6500	.103243 (.000724)
	2	-6.8104	.047702 (.001502)
	3	3.7080	.036470 (.001861)
	4	3.6089	.035628 (.002121)
	agg.		.045341 (.004222)
500	1	.0070	.096433 (.000167)
	2	-.5379	.032949 (.000489)
	4	1.2218	.011537 (.000589)
	8	.3231	.011149 (.000691)
	agg.		.011892 (.000609)
4000	1	-.0033	.094815 (.000027)
	3	-.2538	.011080 (.000153)
	7	1.1741	.001740 (.000114)
	15	.0929	.002056 (.000116)
	agg.		.002087 (.000114)
27000	1	-.0016	.094592 (.000003)
	4	-.0348	.004298 (.000050)
	15	1.1718	.000346 (.000018)
	50	-.1360	.000988 (.000028)
	agg.		.000467 (.000020)
125000	1	-.0032	.094568 (.000002)
	5	.0113	.001988 (.000031)
	25	1.0596	.000093 (.000007)
	125	-.0671	.000485 (.000016)
	agg.		.000116 (.000008)

Table 7: Aggregation of Epanechnikov-Priestley kernels, AR(1) time series.

3. Conclusions

We presented an aggregation procedure for kernel spectral density estimators with asymptotically optimal performance. Our simulations verified that the aggregate consistently performed within a factor of two (in MISE terms) of its best component, and that it was capable of discovering nontrivial optimal linear combinations such as the trapezoid kernel.

The procedure works best with large sample sizes (> 1000), but reasonable results were obtained with a sample size as small as 500. It is particularly important to minimize the number of aggregate components (preferably to two) in the latter case, since there is a large error term linear in the number of components; however, this term has favorable asymptotics, so very large sample sizes allow diverse aggregates to be employed at minimal cost.

In contexts where this is too limiting, we note that the density estimator aggregation result of Yang (2000) has an error term which grows as $\log M$, instead of linearly, in the number of aggregate components (at the price of stronger assumptions on the true density); it may be possible to adapt that strategy to the kernel density estimation context.

Acknowledgement. The authors are grateful to the Editor, Hira Koul, and two anonymous reviewers for their helpful comments. Research of the second author was partially supported by NSF grants DMS 13-08319 and DMS 12-23137.

References

[1] Brillinger, D., 1981. *Time Series: Data Analysis and Theory.* Holden Day, San Francisco.

[2] Parzen, E., 1957. On consistent estimates of the spectrum of a stationary time series. *Ann. Math. Stat.* **28** 157–214.

[3] Parzen, E., 1962. On estimation of a probability density function and mode. *Ann. Math. Stat.* **33** 1065–1076.

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

[5] Politis, D. N., Romano, J.P., and Wolf, M., 1999. *Subsampling*. Springer, New York.

[6] Politis, D. N., 2001. On nonparametric function estimation with infinite-order flat-top kernels. In Charalambides Ch. et al., editors, *Probability and Statistical Models with applications*, pages 469–483. Chapman and Hall/CRC, Boca Raton.

[7] Politis, D. N., 2003. Adaptive bandwidth choice. *J. Nonparam. Statist.* **15** 517–533.

[8] Politis, D. N., 2005. Complex-valued tapers. *IEEE Signal Processing Letters* **12** 512–515.

[9] Politis, D. N., 2011. Higher-order accurate, positive semidefinite estimation of large-sample covariance and spectral density matrices. *Econometric Theory* **27** 703–744.

[10] Priestley, M., 1981. *Spectral Analysis and Time Series*. Academic Press, London.

[11] Rigollet, P., Tsybakov, A., 2007. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics* **16** 260–280.

[12] Rosenblatt, M., 1956. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27** 832–837.

[13] Wang, Z., Paterlini S., Gao F., Yang Y., 2011. Adaptive Minimax Estimation over Sparse ℓ_q-Hulls. http://arxiv.org/pdf/1108.1961.pdf

[14] Yang, Y., 2000. Mixing strategies for density estimation. *Ann. Stat.* **28** 75–87.