MINIMAX ADAPTIVE SPECTRAL ESTIMATION FROM AN ENSEMBLE OF SIGNALS

FLORENTINA BUNEA[1,3], HERNANDO OMBAO[2] AND ANNA AUGUSTE[1]

Abstract. We develop a statistical method for estimating the spectrum from a data set that consists of several signals, all of which are realizations of a common random process. We first find estimates of the common spectrum using each signal, then we construct M partial aggregates. Each partial aggregate is a linear combination of $M-1$ of the spectral estimates. The weights are obtained from the data via a least squares criterion. The final spectral estimate is the average of these M partial aggregates. We provide an oracle inequality for the empirical risk of the partial aggregates which shows that aggregation via least squares yields risk optimal estimators up to a remainder term proportional to the ratio of M, the number of signals, to n, the number of frequencies at which we sample each signal. The ratio M/n is the price to pay for data adaptive linear aggregation and is optimal, in a minimax sense. As a consequence, we show that our final estimator is minimax rate adaptive, if at least two of the estimators per signal attain the optimal rate $n^{-2\alpha/2\alpha+1}$, for spectra belonging to a generalized Lipschitz ball with smoothness index α. Our simulation study strongly suggests that our procedure works well in practice, and in a large variety of situations is preferable to the simple averaging of the M spectral estimates.

Curve aggregation; Model averaging; Risk bounds; Minimax estimation; Periodogram; Spectrum; Stationary random process.

1. Introduction

We consider the following set up. We have a collection of signals, which are assumed to be independent realizations of a common random process. Our goal is to estimate the spectrum of this process. The problem of extracting information from several signals that are generated by the same process is encountered in many scientific fields. In neuroscience, for example, the power spectrum of electroencephalograms (EEGs) have been used to understand mental processes. In Buysse, et al (2001), the EEG power spectrum are used in characterizing sleep processes for normals and depressed. Harmony, et al (1999) used the power spectrum for discriminating between control and mental tasks. Ishihara and Yoshii (1972) used the EEG power spectrum to gain a deeper understanding of mental processes of juvenile delinquents. In seismology, the spectra of seismic signals have been used effectively in differentiating between earthquake and explosion

[1]Department of Statistics, Florida State University. Research partially supported by NSF-DMS 0406049.

[2]Department of Statistics and The Beckman Institute, University of Illinois. Research partially supported by NSF-DMS 0405243.

[3]Corresponding author.

seismic events (see Booker and Mitrovonas (1974), Blandford (1993) and Kakizawa, Shumway and Taniguchi (1998)).

We begin by establishing the statistical framework that will be considered throughout this article. Let $\{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(M)}\}$ be a collection of M independent signals, each having length T. We denote the m-th signal by $\mathbf{X}^{(m)} = [X_1^{(m)}, \ldots, X_T^{(m)}]$. Each signal is a recording from the same stationary random process with zero mean and an absolutely summable auto-covariance function defined by $\gamma(\tau) = E(X_{t+\tau}X_t), \quad \tau = 0, \pm 1, \pm 2, \ldots$. The auto-covariance structure of each signal is equivalently described by its spectrum $g(\nu) = \sum_{\tau=-\infty}^{\infty} \gamma(\tau) \exp(-i2\pi\nu\tau)$, where the frequency $\nu \in [0, 1]$. Since $g(\nu)$ is symmetric about $\nu = 1/2$, in practice it suffices to consider $g(\nu)$ only on the interval $[0, 1/2]$. This set up is frequently encountered in neuroscience experiments and seismic observational studies. The goal in these studies is to estimate the spectra of the groups (in neuroscience, this could mean normal and disease groups, in seismology, seismic events such as earthquake explosion). Empirical evidence suggests that there can be some variability in the spectra of signals within each group. Nevertheless, the variation between signals within a group does not preclude the scientist from estimating the group spectra by using the recorded signals. It is often the case that the variation between signals within a group is typically less significant than the variation between group spectra. This justifies treating the observed group signals as being generated by a common random process. This approach was adopted in Shumway (1996) and Stoffer, et al (1988) and provided scientists with valuable information.

We take the same approach in this article, and we study the problem of estimating the log spectrum $f(\nu) = \log g(\nu)$ from a collection of M independent signals. The classical non-parametric approach to estimating the log spectrum from one signal utilizes the log periodogram, which is a data analogue of the spectrum. The periodogram of the observed m-th signal $\mathbf{X}^{(m)}$ is defined to be

$$I_k^{(m)} = \frac{1}{T} \left| \sum_{t=1}^{T} X_t^{(m)} \exp(-i2\pi\nu_k t) \right|^2, \tag{1.1}$$

where $\nu_k = k/T, k = 1, \ldots, T/2-1$ are the fundamental Fourier frequencies. For ease in exposition, we ignored the periodograms at frequencies $\lambda_k = 0, 1/2$ (this is negligible for large T). It is well known that the periodogram $I_k^{(m)}$ is distributed approximately as $g(\nu) \times \epsilon_k^{(m)}$, where $\nu_k \to \nu$ as $T \to \infty$, and $\epsilon_k^{(m)}$ is a $\chi_2^2/2$ or Exponential(1) random variable that is independent across frequencies ν_k and signals m. The statistical model above can be reduced to an additive noise

model with stabilized variance across ν_k by applying the log transform. Let $n = T/2 - 1$ and denote $\gamma = -0.57221$ to be the Euler-Mascheronni constant. Then

$$Y_k^{(m)} = f(\nu_k) + \varepsilon_k^{(m)} \tag{1.2}$$

where $Y_k^{(m)} = \log I_k^{(m)} - \gamma$ is the bias-adjusted log periodogram; $f(\nu_k) = \log g(\nu_k)$; $\varepsilon_k^{(m)}$ are random variables with mean 0, variance $\sigma^2 = \pi^2/6$ and density $p_\varepsilon(x) = \exp[x - \exp(x)]$ (see Davis and Jones, 1968; Brockwell and Davis, 1991).

The problem of estimating the log spectrum f from one data set (one signal) has been thoroughly studied in the statistical and signal processing literature, see for example Blackman and Tukey (1958a and 1958b), Wahba (1980), Brillinger (1981), Brockwell and Davis (1991), Priestley (1981), Moulin (1994), Shumway and Stoffer (2000), Ombao, Raz, Strawderman and von Sachs (2001). In this paper we are interested in minimax adaptive estimation of f. It is well known that for a function f of smoothness $\alpha > 0$ (we make this precise in Section 2.3), the minimax optimal rate of convergence of an estimator $\widehat{f}$ is of the order $n^{-2\alpha/(2\alpha+1)}$. If, in addition, the construction of $\widehat{f}$ does not depend on α, the estimator $\widehat{f}$ is called minimax adaptive. There has been a large body of work devoted to minimax adaptive estimation in nonparametric regression, see for instance Donoho and Johnstone (1994, 1998), Lugosi and Nobel (1999), Barron, Birgé and Massart (1999), Baraud (2000, 2002), Antoniadis and Fan (2001), Wegkamp (2003), Bunea (2004) and the references therein. All these methods address the standard regression problem of estimating f from *one sample*.

Minimax adaptive estimation of f from a the collection of data sets $C^{(m)}$, $1 \le m \le M$, has not been studied. Since we assume that the data are generated independently from model (1.2), our procedure is based on the following natural strategy: find $\widehat{f}_1, \ldots, \widehat{f}_j, \ldots, \widehat{f}_M$ estimators of f, where each $\widehat{f}_j$ is obtained from the data set $C^{(j)}$, and then combine them into a final estimator, henceforth called the aggregate.

The simplest aggregate is the arithmetic average $\overline{\widehat{f}} = \frac{1}{M}\sum_{j=1}^{M} \widehat{f}_j$. This approach works well in the ideal situation where the signals are free of artifacts that could lead to bad estimates of the spectrum. Under these ideal conditions and if *all* $\widehat{f}_j$ are minimax adaptive estimators, the average estimator is also minimax adaptive, and there is no need for a more complicated strategy. However, in many practical situations some signals may contain artifacts. For example, Gotman (1982) discusses that EEGs may be contaminated by common noise due to eye blinks, electrode

and movement artifacts. In addition there may be physiological noise, such as those related to respiratory and cardiac activities, which may obscure the signal that is due to the experimental stimulus. There are already artifact rejection algorithms in place but these are not guaranteed to be free from error (Nuwer (1988)). Thus, the corresponding estimates from corrupted signals may be poor relative to others, even if we employ optimal methods of estimation. Note that despite that, an average will weigh them equally. Also, even if the data is not corrupted, it may become computationally expensive to construct M minimax adaptive estimates. We provide instances of these cases in Section 3.

In this paper we propose a procedure that addresses these problems at the same time: our aggregate is a weighted combination of the initial estimators, with data dependent weights. The weights are computed via a least squares criterion, and they are therefore expected to down-weigh the estimators with poor fit. In Sections 2.3 and 3 we also discuss how our method may lead to computational savings.

We describe our aggregation strategy in Section 2.1. Our method involves computing M partial aggregates and then averaging them. We quantify the performance of these aggregates in terms of their empirical risks in Section 2.2. The first part of Theorem 2.1, which is the main result of this section, shows that the empirical risk of the partial aggregates is smaller than the empirical risk of any other linear combination of the original estimators, up to the minimal aggregation price, which has the minimax optimal order M/n. Our theorem complements the one of Tsybakov (2003), who established this optimal bound in terms of theoretical risks in regression on random design of known marginal distribution.

The study of the convergence rates of data adaptive aggregates of arbitrary estimators received very little attention. To the best of our knowledge it has only been discussed in Yang (2004), for *one sample* regression on random design. He suggests a sequential procedure of aggregation that yields minimax adaptive estimators if *at least one* of the individual estimators is rate optimal. His procedure requires splitting the sample three times, and as such is not directly applicable to our problem. Also, he points out that the method may become very involved computationally. This therefore creates the need for procedures that address the *multi-sample* problem, are easily implementable and share theoretical properties that are similar to those established in the standard framework. Corollary 2.2 of Section 2.3, which is a consequence of the second part of Theorem 2.1,

shows that our algorithm yields rate optimal aggregates, if *at least two* of the original estimators are rate optimal, validating theoretically the use of our leave one out procedure. The practical implication is the following: one only needs to compute two minimax adaptive estimators, and use standard and less computationally involved methods for the remaining $M-2$ curves. If the processing of a large number of signals is of importance, this can lead to important computational savings, and the optimal convergence rate is still guaranteed.

We compare our methods with the simple averaging procedure in a variety of simulated scenarios in Section 3. The results are consistent: in all situations data adaptive aggregation is superior to averaging. The most notable differences can be seen when some of the M input estimators depart considerably from f. This supports very strongly the theoretical findings of Section 2. In Section 4 we consider a real data example. The last section contains our conclusions. The proof of Theorem 2.1 is given in the Appendix.

2. Least squares aggregation

2.1. **An aggregation algorithm.** In this section we describe our data adaptive procedure of aggregation. For each $m = 1, \ldots, M$ and $k = 1, \ldots, n$ we begin by computing the periodograms $I_k^{(m)}$ and the bias-corrected log periodograms $Y_k^{(m)} = \log I_k^{(m)} - \gamma$. We denote the resulting independent data sets by $C^{(m)} = \left\{(\nu_k, Y_k^{(m)}), 1 \le k \le n\right\}$. Then, in each case, we construct the individual spectral estimates at Step 1. Sections 2.3 and 3.1 contain a detailed discussion of this step.

If we could ensure that all signals are clean, i.e., artifact-free, then we can combine $M-1$ spectral estimates corresponding to, say, $C^{(1)}, \ldots, C^{(M-1)}$, by computing the weights on $C^{(M)}$. In practice, however, it is difficult to assess in advance which data set can play the role of $C^{(M)}$. We then adopt the following strategy: we leave out one data set at a time and use it for the aggregation of the estimators computed on the remaining data sets. Then, we take as the final aggregate the average of these M partial aggregates. We summarize this in the algorithm below. Let $J_{-m} = \{1, \ldots, m-1, m+1, \ldots, M\}$ and denote by $\mathbb{R}_{-m}^{M-1}$ the space obtained from $\mathbb{R}^M$ by deleting dimension m, $1 \le m \le M$.

1. For each $C^{(m)}$, $m = 1, \ldots, M$, compute the spectral estimate $\widehat{f}_j$.

2. For each $m = 1, \ldots, M$ and $j \in J_{-m}$ compute $\widehat{\lambda}_j^{(m)}$ by minimizing

$$\frac{1}{n}\sum_{k=1}^{n}\left[Y_k^{(m)} - \sum_{j\in J_{-m}} \lambda_j \widehat{f}_j(\nu_k)\right]^2$$

over $\lambda \in \mathbb{R}_{-m}^{M-1}$. Let $\widetilde{f}^{(m)}(\nu) = \sum_{j\in J_{-m}} \widehat{\lambda}_j^{(m)} \widehat{f}_j(\nu)$ be the m-th partial aggregate.

3. Form the final estimator $\widehat{f} = \frac{1}{M}\sum_{m=1}^{M} \widetilde{f}^{(m)}$.

2.2. **Optimal rates of aggregation.** In this section we discuss the statistical aggregation price of our method. One typically assesses the performance of a generic estimate $\hat{u}$ in terms of its risk $E\|\widehat{u} - f\|^2$, where $\| \ \|$ can be either the theoretical or the empirical norm, as defined below. For any function g of argument ν we denote by $\|g\|_n^2 = \frac{1}{n}\sum_{k=1}^{n} g^2(\nu_k)$ the empirical norm, and call $E\|\widehat{u} - f\|_n^2$ the empirical risk. For any function g of random argument Z we denote by $\|g\|_\mu^2 = \int g^2(z)d\mu(z)$ the theoretical $L_2(\mu)$ norm, where μ is the probability distribution of the design points Z. We call $E\|\widehat{u} - f\|_\mu^2$ the theoretical risk. The discussion that follows pertains to either risk, so we omit the indices μ and n until they become necessary.

We begin by discussing possible optimal targets for the estimators $\widetilde{f}^{(m)} = \sum_{j\in J_{-m}} \widehat{\lambda}_j^{(m)} \widehat{f}_j$ obtained in Step 2 of our procedure. The superscript (m) reminds us that the weights have been computed from data set $C^{(m)}$. Let now $f_\lambda = \sum_{j\in J_{-m}} \lambda_j \widehat{f}_j$ be an arbitrary aggregate with non-random weights. A way of judging the performance of our procedure is by comparing the risk of $\widetilde{f}^{(m)}$ with the best achievable risk of an arbitrary f_λ. If $\lambda \in \mathbb{R}_{-m}^{M-1}$ the comparison is done with an arbitrary linear combination of the estimators $\{\widehat{f}_j\}_{j\in J_{-m}}$. If $\lambda \in \Lambda_{-m}^{M-1}$, the simplex in $\mathbb{R}_{-m}^{M-1}$, then we look at arbitrary convex combinations. Finally, if λ is the set of the simplex' vertices, then we compare $\widetilde{f}^{(m)}$ with individual estimators. Then, as set forth in Nemirovski (2000), we may be interested in establishing the following bounds:

- The linear aggregation (L) bound $E\|\widetilde{f}^{(m)} - f\|^2 \le \inf_{\lambda\in\mathbb{R}_{-m}^{M-1}} E\|f_\lambda - f\|^2 + \Delta_L$.
- The convex aggregation (C) bound $E\|\widetilde{f}^{(m)} - f\|^2 \le \inf_{\lambda\in\Lambda_{-m}^{M-1}} E\|f_\lambda - f\|^2 + \Delta_C$.
- The model selection aggregation (MS) bound $E\|\widetilde{f}^{(m)} - f\|^2 \le \inf_{j\in J_{-m}} E\|\hat{f}_j - f\|^2 + \Delta_{MS}$.

The second term in the right hand side of each of the above inequalities is called the aggregation rate. The minimax optimal aggregation rate in each case has been established by Tsybakov (2003), for the theoretical risk, and Bunea, Tsybakov and Wegkamp (2004) for the empirical risk. For both

risks, they are : $\Delta_L = O(M-1/n)$, $\Delta_{MS} = O(\log(M-1)/n)$ and

$$\Delta_C = \begin{cases} (M-1)/n & \text{if } M-1 \leq \sqrt{n} \\ \sqrt{\{\log(1+(M-1)/\sqrt{n})\}/n} & \text{if } M-1 > \sqrt{n}, \end{cases}$$

in order. We call the constant multiplying the infima in the right hand side of these inequalities the leading constant. Note that it is 1 in the ideal situation.

The type of norm one uses is typically connected with the nature of the design points. If the design points can be assumed to be random and generated from a probability distribution μ, then the theoretical norm is the natural one to use in order to assess the performance of the estimator at a new data point Z, generated from μ independently of the data set on which the estimator was computed. If the design points are not random, then the empirical norm becomes of importance, as the theoretical norm loses its main interpretation. This is the framework we consider in this paper, since the design points are chosen by the scientist. In this context the theoretical norm can no longer be defined in terms of the distribution of the design points. One can however consider the $L_2(\mu^*)$ norm, where μ^* is the Lebesgue measure on the range $[a,b]$ of the design points. The associated $L_2(\mu^*)$ risk then measures the expected performance of the estimator when we average over all possible design points in $[a,b]$. Such measure is especially valuable when the pattern of the design points is not known and n is small. In contrast, in this paper we consider uniform design points in $[0,1]$, and n is large. In addition, the main focus here is the fit of the aggregate to the given data. This motivates the use of the empirical risk throughout this paper. We discuss the merits and limitations of using the $L_2(\mu^*)$ risk in this context in Remark 3 below.

Aggregation of *arbitrary* estimators obtained from *one data set* in regression models of *random design* and inequalities in the spirit of those above is becoming a growing area of research. Yang (2000, 2001, 2004) suggests several methods of convex aggregation, in particular ARM (adaptive regression by mixing). He establishes convex aggregation bounds with leading constants that are typically much larger than 1 and with aggregation rates that can be equal or approximately equal to the optimal rates when M is a power of n. Birgé (2003) suggests a convex aggregation method satisfying an analogue of the (C) bound with a leading constant that can be much greater than 1 and with a rate that is optimal for $M > \sqrt{n}$ and suboptimal for $M \leq \sqrt{n}$. Wegkamp (2003) suggests a data splitting aggregation strategy that achieves the (MS) bound, but with a leading constant greater than 1.

The study of the methods that lead to the optimal aggregation bounds above is still developing, and the bibliography is limited. See Catoni (2001) for a sequential aggregation method leading to the (MS) bound in Gaussian regression; Nemirovski (2000), Juditsky and Nemirovski (2000), for a stochastic approximation algorithm yielding the (C) bound, when $M > \sqrt{n}$; Tsybakov (2003), Koltchinskii (2004, Section 8) and Audibert(2003) for the (C) bound in both cases. Procedures achieving the (L) bound have received substantially less attention. Nemirovski (2000) discusses linear aggregation for Gaussian white noise models. Tsybakov (2003) suggests a procedure achieving the (L) bound for regression models with random design of known marginal distribution μ.

The literature on optimal aggregation bounds on the *empirical risk* of aggregates built from *arbitrary* estimators obtained from *one data set* in regression models of *fixed design* is also limited. Wegkamp (2003) discusses the (MS) bound, again with a constant greater than 1. Barron and Leung (2004) obtain the (C) bound via a convex aggregation method using exponential weights, but for the simplified model $Y = \theta + W$, where $\theta \in \mathbb{R}$ is an unknown mean and W is a Gaussian error term of mean zero and known variance.

The methods mentioned above for aggregating estimators of f share a common feature: one observes only *one* data set, and this is split in two independent parts. The first part is used to construct, say, M estimators, and the second part is used for aggregation. For a pre-specified m, Step 2 of our algorithm can be regarded as a natural extension of this principle from one data set to multiple data sets. The limitation is that only $M - 1$ estimators can be aggregated in this way. Nevertheless, for that fixed m, the theoretical results obtained in the classical framework would transfer unchanged here. To the best of our knowledge, the optimal empirical risk (L) bound for aggregation of arbitrary estimators in regression on fixed design has not been established in the classical context, and therefore not in ours. We show that the estimators obtained at Step 2 of our procedure achieve the (L) bound.

We also remark that the data splitting strategies for *one sample* have the common problem of introducing extra variability by using a random split. Although we do not have this problem here directly, we can induce it by deciding in advance on the set $C^{(m)}$ in Step 2 of our procedure. Our remedy to this is the leave-one-out strategy. To the best of our knowledge, there are no theoretical results on the rates of convergence of this type of estimators. The next subsection provides sufficient conditions under which our method yields minimax adaptive estimates.

We explain in what follows why the (L) bound is important to our paper. First notice that an interesting feature of the (L), (C) and (MS) bounds is that they are not, in general, comparable. We always have

$$\inf_{\lambda\in\mathbb{R}^{M-1}} E\|f_\lambda - f\|^2 \le \inf_{\lambda\in\Lambda^{M-1}} E\|f_\lambda - f\|^2 \le \inf_{j\in J_{-m}} E\|\hat{f}_j - f\|^2, \tag{2.1}$$

since the infima are taken over nested sets. On the other hand, the aggregation rates satisfy a somewhat reversed inequality: Δ_L is the largest and Δ_{MS} is the smallest, while Δ_C plays an intermediate role, depending on how large M is relative to n.

Thus, in general, one cannot declare one bound superior to the others. However, in the applications we consider here, M does not depend on n, and is typically much smaller than n. See, for instance, Kakizawa et al. (1998), Buysse et al. (2001), Harmony et al. (1999) for studies in seismology and neuroscience that use spectral analysis. If $M \le \sqrt{n}$, it is clear from above that the linear bound (L) is smaller than the convex bound (C). Thus, if one can construct an aggregate achieving the linear bound, then this aggregate will automatically have a risk that is smaller than the risk of any other convex combination of the estimators, up to the remainder term $(M-1)/n$. The arithmetic average is an instance of such a convex combination. Also, by (2.1) and since the terms $\log(M-1)/n$ and $(M-1)/n$ become comparable for large n and fixed M, an estimator achieving the (L) bound may also be preferable to one meeting the (MS) target.

The first part of Theorem 2.1 guarantees that each estimator obtained at Step 2 achieves the (L) bound in terms of empirical risks. The second part shows that the risk of the estimator $\hat{f}$, obtained in Step 3 is bounded by the average of the smallest achievable risks of the partial aggregates, up to the aggregation price. This is an intermediate result that will allow us to obtain, in the next section, optimal convergence rates for $\hat{f}$ under minimal conditions on the individual estimators.

Theorem 2.1.

(*a*) *For each* $m = 1, \dots, M$ *and every* n

$$E\|\widetilde{f}^{(m)} - f\|_n^2 \le \inf_{\lambda\in\mathbb{R}^{M-1}_{-m}} E\|f_\lambda - f\|_n^2 + \sigma^2\frac{M-1}{n}.$$

(*b*)
$$E\|\widehat{f} - f\|_n^2 \le \frac{1}{M}\sum_{m=1}^{M} \inf_{\lambda\in\mathbb{R}^{M-1}_{-m}} E\|f_\lambda - f\|_n^2 + \sigma^2\frac{M-1}{n}.$$

The proof of the first inequality is presented in the Appendix. The second inequality is immediate, since by Jensen's inequality we have

$$E\|\widehat{f}-f\|_n^2 = E\|\frac{1}{M}\sum_{m=1}^{M}(\widetilde{f}^{(m)}-f)\|_n^2 \leq \frac{1}{M}\sum_{m=1}^{M}E\|\widetilde{f}^{(m)}-f\|_n^2,$$

and we can then invoke part (a) of the theorem.

Remark 1. Notice that $\bar{f}_{-m} = \frac{1}{M-1}\sum_{j\in J_{-m}}\hat{f}_j$ is a particular instance of f_λ, $\lambda \in \mathbb{R}_{-m}^{M-1}$, corresponding to all λ_j's equal to $\frac{1}{M-1}$. Then (a) implies that $E\|\widetilde{f}^{(m)}-f\|_n^2 \leq E\|\bar{f}_{-m}-f\|_n^2 + \sigma^2\frac{M-1}{n}$, for all m. This relation shows that the data adaptive aggregate is superior to the average, when the remainder term becomes negligible.

Remark 2. Notice that if we took our final estimator to be $\bar{\hat{f}}$, the arithmetic average of all M initial estimators, the sharpest bound on its risk, in terms of individual risks, is simply given by Jensen's inequality. Thus, we always have $E\|\bar{\hat{f}}-f\|_n^2 \leq \frac{1}{M}\sum_{j=1}^{M}E\|\hat{f}_j-f\|_n^2$. Part (b) of Theorem 2.1 implies that

$$E\|\widehat{f}-f\|_n^2 \leq \frac{1}{M}\sum_{m=1}^{M}\inf_{\lambda\in\mathbb{R}_{-m}^{M-1}}E\|f_\lambda-f\|_n^2+\sigma^2\frac{M-1}{n} \leq \frac{1}{M}\sum_{j=1}^{M}E\|\hat{f}_j-f\|_n^2+\sigma^2\frac{M-1}{n}. \tag{2.2}$$

This suggests that there exists a trade-off between the two estimators: if we use $\bar{\hat{f}}$ we do not have to pay the aggregation price, but if we are willing to pay it, we can guarantee that the risk of $\hat{f}$ has a sharper bound, given by the first inequality in (2.2), than the risk of $\bar{\hat{f}}$. It is this bound that allows us to obtain optimal rates of convergence on $\hat{f}$ under much weaker conditions than for $\bar{\hat{f}}$, as we illustrate in the next subsection.

Remark 3. Theorem 2.1 discusses the empirical risk $E\|\widetilde{f}^{(m)}-f\|_n^2$ of our aggregates and holds in full generality, under no assumptions on the estimators or the underlying regression function. One can alternatively study $\widetilde{f}^{(m)}$ in terms of its $L_2(\mu^*)$ risk $E\|\widetilde{f}^{(m)}-f\|_{\mu^*}^2 = E\int_0^1(\widetilde{f}^{(m)}(\nu)-f(\nu))^2d\nu$. Here $\|\ \|_{\mu^*}$ denotes the $L_2(\mu^*)$ norm with respect to the Lebesgue measure μ^* on [0, 1]. However, obtaining similar oracle inequalities for the theoretical risk of the estimates requires further assumptions.

A standard assumption made in this context is that the design points are *random*. If in addition we assume that their distribution is *uniform* on [0, 1], the following modification of Step 2 yields the theoretical risk analogue of Theorem 2.1.

- For each $m = 1, \ldots, M$, let S^{-m} be the subspace of $L_2(\mu^*)$ generated by $\{\hat{f}_j\}_{j \in J_{-m}}$. Find a basis $B = \{\psi_1, \ldots, \psi_{M'}\}$ of S^{-m}, orthonormal with respect to μ^*, with $M' \leq M - 1$.
- For each $j = 1, \ldots, M'$, compute $\hat{\beta}_j = \frac{1}{n} \sum_{k=1}^{n} Y_k \psi_j(\nu_k)$. Define $\widetilde{g}^{(m)}(\nu) = \sum_{j=1}^{M'} \hat{\beta}_j^{(m)} \psi_j(\nu)$, for $\nu \in [0, 1]$.

We can then take as the final estimator $\hat{g} = \frac{1}{M} \sum_{m=1}^{M} \widetilde{g}^{(m)}$. Since μ^* is the Lebesgue measure on [0, 1], B can be computed in practice via a Gram-Schmidt procedure. We remark that $\hat{\beta}_j$ is not the ordinary least squares estimator, since B is a basis in a function space, and there is no guarantee that the corresponding vectors $(\psi_j(\nu_1), \ldots, \psi_j(\nu_n))$, $1 \leq j \leq M'$, form an orthonormal system in $\mathbb{R}^n$. This strategy is presented in Tsybakov (2003) and his Theorem 4 shows that it yields the analogue of Theorem 2.1 with $\| \ \|_{\mu^*}$ in place of $\| \ \|_n$, and with the aggregation price modified to $(\sigma^2 + L^2)(M - 1)/n$, where $L > 0$ is an assumed common bound on the regression function and the estimators.

Although regression on *uniform random design* is closely related to regression on *fixed design on a uniform grid*, we discussed above why we cannot use the random design assumption here. The literature on bounds on the theoretical risk of estimates for regression models on fixed design is very limited, and mainly developed for aggregation of elements of various orthogonal bases of $L_2(\mu^*)$. Although devoted to this latter situation, Corollary 3.2 page 474 in Baraud (2000) can be adapted to aggregation of *arbitrary* estimates and implies that the following $L_2(\mu^*)$ risk bound holds for a least squares estimator $\widetilde{f}^{(m)}$ in nonparametric regression on *fixed design*: $E\|\widetilde{f}^{(m)} - f\|_{\mu^*}^2 \leq C_n \left(\inf_{\lambda \in \mathbb{R}_{-m}^{M-1}} E\|f_\lambda - f\|_{\mu^*}^2 + d_\infty^2(f, S_N) + \sigma^2 \frac{M-1}{n} \right)$. Here S_N is a N-dimensional space of $L_2(\mu^*)$ that includes S_{-m} and $d_\infty^2(f, S_N) = \inf_{g \in S_N} \|g - f\|_\infty$, where $\| \ \|_\infty$ is the supremum norm. The constant C_n depends on the smallest and largest eigen values of the matrix $\Phi_n = \left(\frac{1}{n} \sum_{k=1}^{n} \phi_j(\nu_k) \phi_l(\nu_k) \right)_{1 \leq j, l \leq N}$, where $\{\phi_j\}_{1 \leq j \leq N}$ form an orthonormal system in S_N with respect to μ^*. He showed that this bound cannot be essentially improved. The quantities C_n and $d_\infty^2(f, S_N)$ can only be evaluated on a case to case basis, as they will depend on the choice of S_N, which in turn depends on the original estimators. Further investigation of this issue is beyond the scope of this article, as our focus here is on aggregation of *arbitrary* estimators. Moreover, in the applications we consider here n is typically larger than 256 (512 and 1024 are common values), and so the design

points $\nu_k = k/n$ form a fine mesh of [0, 1]. Thus the empirical risk offers an accurate measure for the performance of our estimators on this range.

2.3. **Rates of convergence of the aggregated estimator.** Theorem 2.1 and our discussion so far are independent of the method we employ at Step 1 for estimating the spectrum of each curve. We turn now to investigating this step and its impact on the rate of convergence of our final estimators.

Let $Lip^*(\alpha, 2)$ denote a generalized Lipschitz space, for some smoothness parameter $\alpha > 0$ and let $|\ |_{\alpha,2}$ be the seminorm in this space (see, e.g., DeVore and Lorentz, 1993, page 51, for definition and properties). The subscript 2 indicates that we consider square integrable functions. For some positive constant $A > 0$, define $\mathcal{D}_{\alpha,2}(A) = \{g \in Lip^*(\alpha, 2),\ |g|_{\alpha,2} \leq A\}$. As we mentioned in the Introduction, an adaptive minimax rate optimal estimator of $f \in \mathcal{D}_{\alpha,2}(A)$ is an estimator whose construction does not depend on α and whose risk satisfies $E\|\widehat{f} - f\|_n^2 = O(n^{-2\alpha/(2\alpha+1)})$, which immediately implies that $\|\widehat{f} - f\|_n^2 = O_P(n^{-2\alpha/(2\alpha+1)})$, by Markov's inequality. An estimator with this property is called optimal rate consistent.

Corollary 2.2 below shows that our aggregate is minimax optimal if *at least two* of the M estimators are rate optimal. This contrasts with the performance of a simple average aggregate, which is minimax optimal if *all* the estimators have this property.

Corollary 2.2. *Assume that $f \in \mathcal{D}_{\alpha,2}(A)$, $\alpha > 0$. If there exist $l_1, l_2 \in \{1, \ldots, M\}$ such that $E\|\widehat{f}_{l_j} - f\|_n^2 = O(n^{-2\alpha/(2\alpha+1)})$, $j = 1, 2$, then $E\|\widehat{f} - f\|_n^2 = O(n^{-2\alpha/(2\alpha+1)})$.*

Proof: Notice that for each m we have $\inf_{\lambda \in \mathbb{R}^{M-1}_{-m}} E\|f_\lambda - f\|_n^2 \leq \inf_{j \in J_{-m}} E\|\widehat{f}_j - f\|_n^2$, since the infimum is taken over a smaller set. Then

$$
\begin{aligned}
E\|\widehat{f} - f\|_n^2 &\leq \frac{1}{M}\sum_{m=1}^{M} \inf_{j \in J_{-m}} E\|\widehat{f}_j - f\|_n^2 + \sigma^2 \frac{M-1}{n} \\
&= \frac{1}{M} \inf_{j \in J_{-l_1}} E\|\widehat{f}_j - f\|_n^2 + \frac{1}{M} \sum_{m=1; m \neq l_1}^{M} \inf_{j \in J_{-m}} E\|\widehat{f}_j - f\|_n^2 + \sigma^2 \frac{M-1}{n} \\
&\leq \frac{1}{M} E\|\widehat{f}_{l_2} - f\|_n^2 + \frac{M-1}{M} E\|\widehat{f}_{l_1} - f\|_n^2 + \sigma^2 \frac{M-1}{n} \qquad (2.3) \\
&= O(n^{-2\alpha/(2\alpha+1)}) + O((M-1)/n) = O(n^{-2\alpha/(2\alpha+1)}).
\end{aligned}
$$

The first inequality holds by part (b) of Theorem 2.1. For the third inequality, we notice that $l_2 \in J_{-l_1}$, since this set contains all indices from 1 to M except for l_1, by definition. Also, since

$l_1 \in J_{-m}$, for all $m \neq l_1$, we can bound each infima by $E\|\widehat{f}_{l_1} - f\|_n^2$. The last equality holds because $O((M-1)/n) = O(1/n)$, since M is independent of n, $n^{-2\alpha/(2\alpha+1)} > n^{-1}$ and by our hypothesis on $\widehat{f}_{l_j}$, $j = 1, 2$. ■

Remark. This corollary provides further insight into our algorithm. It shows that, in practice, we only need to compute two minimax adaptive estimators in order to ensure that $\hat{f}$ is minimax adaptive. For the other $M-2$ estimators we can therefore choose any method that renders reasonable estimators per signal. We investigate this in detail in Section 3. Furthermore, notice the factors $1/M$ and $M - 1/M < 1$ in (2.3). This indicates that we should expect increased accuracy for a larger M, as long as the ratio $(M-1)/n$ remains small. It also supports the intuitive belief that a larger number of signals will contribute to a more accurate estimate, as further illustrated throughout our simulation section.

Since Corollary 2.2 is independent of the minimax adaptive method used in Step 1, one can choose any of the tecniques mentioned in the Introduction. Typical methods include wavelet-based estimation via thresholding, as pioneered by Donoho and Johnstone (1994, 1998), penalized least squares selection of the best approximating basis for f, as in Baraud (2000) or locally linear smoothing combined with a data-splitting procedure for adaptively selecting the bandwidth h, as suggested by Hentgartner, Matzner-Lober and Wegkamp (2002), henceforth HMW. For computational simplicity, we have opted here for the HMW method. We describe their method in detail in the next section. We briefly recall their result here, for completeness. If $\widehat{f}_j$, $j = 1, \ldots, M$, are obtained by the HMW method, their Theorem 3, page 794, and the comments following it imply that, for n large and some positive constant $C > 0$, we have $E\|\widehat{f}_j - f\|_n^2 \leq C \inf_{1/n \leq h \leq 1}(\frac{1}{nh} + h^{2\alpha}) = O(n^{-2\alpha/(2\alpha+1)})$. The sum $\frac{1}{nh} + h^{2\alpha}$ reflects the usual variance-bias decomposition, up to multiplicative constants, see for example Korostelev and Tsybakov (1993). Thus the estimator $\widehat{f}_j$, which corresponds to a data adaptive choice of h, and hence is constructed independently of α, achieves the minimax rate. The next section presents our aggregation algorithm in connection with the HMW method, in the context of a simulation study. Our results strongly suggest that the combined algorithm performs very well in practice.

3. Simulation Studies

The goal in these numerical investigations is to compare, using the mean squared error MSE as a criterion, the performance of our proposed least squares aggregate (LSA) to the performance of the arithmetic average (Av), under different scenarios.

In Set up I, some of the M signals are corrupted. This scenario may be encountered in neuroscience experiments, for instance, where some EEG signals may contain artifacts (corruptions) due to muscle movement and cardiac/respiratory activities. Such artifacts can be removed by algorithms for signal processing but these are not guaranteed to be free from error. We investigated the effect of having $k = 0, 2$ corrupted signals for data sets with $M = 8, 20$ signals each having log periodogram curve of lengths of $n = 128, 256, 512$.

While in Set up I we controlled the corruption, in practice we do not know where and how it may occur. We expect, however, that a high level of corruption will yield estimators that are far from the true log spectrum. We can mimic this situation in simulations by deliberately constructing estimators that are far from the truth. This is the rationale behind Set up II below.

Finally, in Set up III, we investigate the Remark following Corollary 2.2: we compare the average of M minimax estimators to our aggregate, when only two of the input estimates are minimax and the rest are computed via an ad-hoc smoothing method.

In the simulation studies, we used ARMA(p, q) processes. A time series $X_t, t = 1, \ldots, T$ is said to be generated from an ARMA(p, q) if it has the representation $X_t = \beta_1 X_{t-1} + \ldots + \beta_p X_{t-p} + \epsilon_t + \alpha_1 \epsilon_{t-1} + \ldots \alpha_q \epsilon_{t-q}$ where ϵ_t are iid with zero mean and variance σ^2. The spectrum of an ARMA(p, q) process defined above is

$$g(\nu) \quad = \quad \sigma^2 \frac{|1 + \sum_{\ell=1}^{q} \alpha_\ell \exp(-i2\pi\ell\nu)|^2}{|1 - \sum_{\ell=1}^{p} \beta_\ell \exp(-i2\pi\ell\nu)|^2}$$

We conducted simulations for the process ARMA$(0, 4)$ with $\alpha = [-0.3, -0.6, -0.3, 0.6]$. We also investigated other processes, in particular autoregressive models. Due to the similarity of the results and space limitation, we only we use this model here as the basis of our simulations.

3.1. **Set up I.** We performed 500 simulations. Each simulation contained M signals of which $k = 0, 2$ may be corrupted. For every signal, we computed the log periodogram at n distinct Fourier frequencies. We have considered separately $M = 8, 20$ and for each M we considered, in succession, $k = 0, 2$ corrupted signals, $n = 128, 256, 512$. We thus have 12 combinations (M, k, n) for

the model. The signals are corrupted by artificially adding values of log(20) to the log periodogram in neighborhoods around the peaks (local maxima). This approach to corrupting data is patterned after what we typically see in practice (in fMRI for instance) where the signal of interest (i.e., changes in the hemodynamic response that is related to the stimulus in the experimental design) can be corrupted by non-relevant physiological noise (such as respiratory or cardiac activities). For the $b-$th simulation ($b = 1, \ldots, 500$), we computed the the mean squared error between the true log spectrum and the aggregate estimate

$$MSE_b = \frac{1}{n}\sum_{k=1}^{n}[f(\nu_k) - \widehat{f}_b(\nu_k)]^2.$$

We describe here the locally smoothing method we employed. Given a signal $\left(\nu_k, Y_k^{(m)}\right)_{k=1}^{n}$ assumed to follow the regression model (1.2), the locally linear smoother is $\widehat{\beta}_0 \equiv \widehat{f}_{m,h}(\nu)$ which minimizes for each ν the weighted sum of squares

$$\sum_{k=1}^{n} K(\frac{\nu - \nu_k}{h})(Y_k^{(m)} - \beta_0 - \beta_1'(\nu - \nu_k))^2$$

over (β_0, β_1), where K is a bounded probability density and h is the bandwidth. In our simulations we considered the Epanechnikov kernel $K(\nu) = 3(1 - \nu^2/5)I(\nu^2 \leq 5)/4\sqrt{5}$. We use the method of HTW (2002) for a data adaptive choice of h. We first randomly split $\left(\nu_k, Y_k^{(m)}\right)_{k=1}^{n}$ into a training sample of size n_1 and a testing sample of size n_2. The conditions for the choice of these sub sample sizes are that $\frac{n_2}{n_1} \longrightarrow 0$, $\frac{n_2}{n_1^{\beta}} \longrightarrow \infty$ for some $\beta > \frac{4}{5}$, and finally $n_1 + n_2 = n$. We used $n_2 \approx n_1^{\frac{9}{10}}$ for our simulation study. For convenience, let us denote the training sample by $\left\{(\nu_1, Y_1^{(m)}), \ldots, (\nu_{n_1}, Y_{n_1}^{(m)})\right\}$ and the testing sample by $\left\{(\nu_{n_1+1}, Y_{n_1+1}^{(m)}), \ldots, (\nu_n, Y_n^{(m)})\right\}$. For each h in the geometric grid $\mathcal{H} = \left\{a, a(1+\delta), \ldots, a(1+\delta)^{n_1-1}, 1\right\}$ $\quad a = \delta = \frac{1}{n_1}$ we find the locally linear smoother $\widehat{f}_{m,h}(\nu)$, as defined above, using the training sample. We then select the estimator $\widehat{f}_{m,H}(\nu)$ which has the smallest empirical prediction error $\sum_{j=n_1+1}^{n}\left\{Y_j^{(m)} - \widehat{f}_{m,h}(\nu_j)\right\}^2$, where this is now evaluated on the testing sample.

We present the tables of the percentile values (10, 20, 50, 80, 90-th precentiles) of the MSEs as well the mean and standard deviation, based on 500 simulated datasets, for the LSA and Av methods. In Table 1, one observes that the distribution of the MSEs of LSA and Av share some overlap. However, in all cases, for either $k = 0$ or $k = 2$ the median and mean MSE of the LSA is smaller than that of the Av. Thus, on average, LSA, which is data-adaptive, outperforms the Av.

		P_{10}	P_{20}	P_{50}	P_{80}	P_{90}	EMSE	σ
$M = 8, n = 128$								
$(k = 0)$	Av	50.3	58.5	81.3	111.1	129.7	86.2	(30.9)
	LSA	27.6	32.5	43.6	56.1	63.8	45.3	(14.8)
$(k = 2)$	Av	59.2	67.7	89.3	114.0	129.2	92.2	(27.5)
	LSA	42.7	49.5	62.8	80.2	90.5	64.9	(19.6)
$M = 8, n = 256$								
$(k = 0)$	Av	24.5	28.4	38.6	50.8	58.9	41.0	(15.5)
	LSA	15.1	17.0	22.4	29.2	33.7	23.7	(7.7)
$(k = 2)$	Av	27.4	31.5	41.6	52.7	59.9	43.2	(13.9)
	LSA	19.7	22.6	29.7	37.9	43.0	30.7	(9.4)
$M = 8, n = 512$								
$(k = 0)$	Av	12.7	15.1	19.8	25.2	30.1	20.6	(6.8)
	LSA	8.3	9.7	12.3	15.7	17.5	12.8	(3.7)
$(k = 2)$	Av	13.4	15.3	19.9	25.6	28.7	20.8	(6.5)
	LSA	9.5	11.1	13.9	17.4	19.1	14.3	(3.9)
$M = 20, n = 128$								
$(k = 0)$	Av	52.2	59.0	73.7	89.5	98.7	75.0	(18.3)
	LSA	14.0	16.2	21.5	28.2	32.4	22.6	(7.3)
$(k = 2)$	Av	51.8	58.1	72.0	87.6	96.0	73.7	(17.8)
	LSA	18.0	20.8	26.2	33.5	37.8	27.3	(7.9)
$M = 20, n = 256$								
$(k = 0)$	Av	24.1	26.6	33.3	41.6	45.5	34.3	(9.0)
	LSA	7.3	8.2	10.4	13.3	15.1	10.9	(3.2)
$(k = 2)$	Av	23.8	26.4	32.5	40.2	45.3	33.8	(8.9)
	LSA	8.4	9.6	12.0	15.1	16.7	12.5	(3.6)
$M = 20, n = 512$								
$(k = 0)$	Av	11.3	12.7	16.1	19.9	21.8	16.4	(4.3)
	LSA	3.6	4.2	5.2	6.6	7.2	5.4	(1.5)
$(k = 2)$	Av	10.8	12.3	15.6	19.2	21.4	15.9	(4.2)
	LSA	4.0	4.5	5.7	7.1	7.8	5.9	(1.5)

TABLE 1. Set up: data set contains $k = 0, 2$ number corrupted signals. The values shown are the percentile values (P_v denotes the $v-$th percentile) of the MSEs, the mean and standard deviation of the MSEs based on 500 simulated data sets for the arithmetic average (Av) and the proposed method (LSA). The numerical values are in units of 10^{-3}.

In addition, the standard deviation of the LSA is smaller than that of the arithmetic average. This indicates that LSA is more stable, i.e., gives more consistent results, than the arithmetic average. Moreover, in most of these cases, the 80-th percentile of the MSE for the LSA is comparable to (or sometimes even smaller) than the 20-th percentile of the Av. Thus, even in the situation where LSA performs "poorly" (consider the 80-th percentile MSE), it is still competitive to the "best" performance (say 20-th percentile) of the Av. Also, the MSE's decrease for both methods when M increases.

3.2. **Set up II.** In the previous set up we controlled the corruption and we only considered 2 corruptions. In practice however we do not know where and how the corruption occurs. We expect however that a high level of corruption will yield estimators that are far from the true log spectrum. We considered two different cases. In Case A, all initial estimates except for two are minimax, and they obtained by the HMW method. The two non-minimax estimators are linear and polynomial of order 5, respectively. This complements the previous case, in which we corrupted two signals, but we still used the HMW method for *all* input estimates.

Case B illustrates the extreme situation in which only two initial estimates are minimax (computed here by the HMW method). For the non-minimax estimators, one is linear and the rest are polynomials of order 5. In Table 2, the results clearly show that the distribution of the MSE values for the LSA and Av methods do not share an overlap. In particular, we draw the attention of the reader to the tails of the MSE values. The "worst" behavior of the LSA ($90-$th percentile of MSEs) is clearly better than the "best" behavior of the arithmetic average ($10-$th percentile). Thus, in the situation where some initial curve estimates are not optimal (and even bad such as fitting linear estimates to more complicated data), the proposed LSA method is markedly superior to the naive arithmetic average. Unlike the arithmetic average, the LSA procedure, being data adaptive, is able to overcome the effect of having some bad initial curve estimates. Moreover, in Figure 1, the plots show that a typical LSA estimate (curve corresponds to the median MSE for LSA) is able to capture the peaks and troughs better than a typical arithmetic average estimate (curve corresponds to the median MSE for Av). As in Set up I, the MSE's decrease for both methods as M increases.

3.3. **Set up III.** If all estimators are minimax, then the average is minimax. However, if the ultimate goal is to compute a minimax estimator, Corollary 2.2 indicates that only two minimax

M=8		P_{10}	P_{20}	P_{50}	P_{80}	P_{90}	EMSE	σ
$n = 128$								
(Case A)	Av	109.5	117.5	147.8	178.6	206.0	150.4	(38.7)
	LSA	31.5	36.7	49.8	67.1	76.7	52.5	(18.2)
(Case B)	Av	190.7	200.2	220.1	254.7	280.6	228.0	(32.2)
	LSA	44.3	50.1	73.1	95.4	111.0	75.9	(25.6)
$n = 256$								
(Case A)	Av	68.2	77.9	95.8	115.4	123.4	97.0	(23.9)
	LSA	16.0	20.2	25.6	30.7	35.0	25.9	(7.4)
(Case B)	Av	161.7	175.6	191.7	210.9	224.9	193.5	(26.1)
	LSA	23.8	29.2	37.6	55.0	59.5	41.4	(14.9)
$n = 512$								
(Case A)	Av	51.4	56.9	69.1	77.5	85.6	68.4	(12.2)
	LSA	9.0	10.9	13.4	16.4	18.4	13.6	(3.7)
(Case B)	Av	154.2	160.5	175.6	187.5	195.2	174.7	(15.2)
	LSA	13.3	15.8	21.7	30.9	39.2	23.4	(9.2)

M=20		P_{10}	P_{20}	P_{50}	P_{80}	P_{90}	EMSE	σ
$n = 128$								
(Case A)	Av	72.1	80.0	98.1	112.4	128.4	98.5	(22.0)
	LSA	14.4	17.8	23.7	29.2	32.0	24.3	(9.2)
(Case B)	Av	448.7	456.3	475.7	496.2	511.7	477.1	(24.0)
	LSA	15.0	17.6	24.3	30.6	35.8	25.7	(11.2)
$n = 256$								
(Case A)	Av	36.0	41.6	48.4	57.5	62.1	49.7	(11.1)
	LSA	7.0	8.1	10.1	11.9	13.2	10.3	(2.6)
(Case B)	Av	429.2	433.4	450.3	465.2	471.6	450.9	(17.4)
	LSA	8.4	9.5	11.9	14.1	15.5	12.0	(2.9)
$n = 512$								
(Case A)	Av	21.6	24.3	28.6	33.0	36.0	29.2	(6.0)
	LSA	3.4	4.1	5.3	6.3	7.2	5.4	(1.4)
(Case B)	Av	423.0	427.1	434.7	444.0	450.9	436.1	(11.2)
	LSA	4.5	5.3	6.7	8.4	9.3	6.8	(1.7)

TABLE 2. Set up: some initial estimates are not minimax. In Case A all but 2 are minimax. In case B, only two are minimax . The values shown are the percentile values (P_v denotes the $v-$th percentile) of the MSEs, the mean and standard deviation of the MSEs based on 500 simulated data sets for the arithmetic average (Av) and the proposed method (LSA). The numerical values are in units of 10^{-3}.

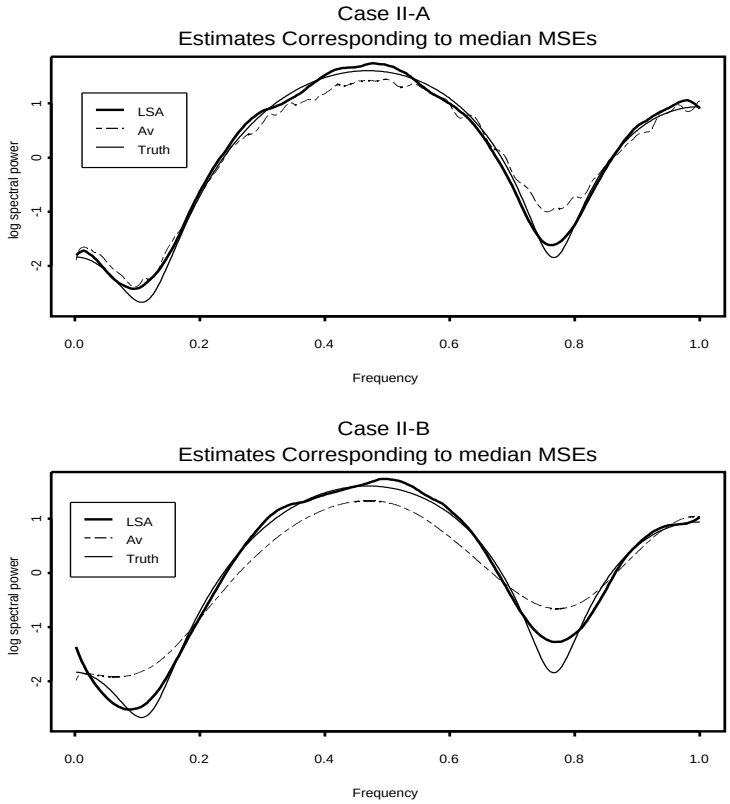

FIGURE 1. True log spectrum; the LSA estimate and the arithmetic average estimate that corresponds to the median of the MSEs for each of the methods. The x-axis is the frequency where 1 corresponds to the Nyquist frequency. The y-axis is the log spectral power.

estimates are needed if our algorithm is used. In Set up II we considered extreme cases for the remaining $M-2$ estimates. Here we design a procedure that can be used in practice in connection with a minimax adaptive method of choice. We use again the HMW method. We outline the step below. [Step 1]. Use the HMW method to get initial estimates for two of the $M = 8, 20$ log periodogram data. [Step 2]. Compute h, the average of the two bandwidths from Step 1. [Step 3]. Use this h as the bandwidth to smooth the remaining $M-2$ log periodograms. [Step 4]. Use these resulting M initial estimates as the input of our algorithm.

This is compared to the arithmetic average (which is the intuitive but computationally more expensive) approach: [Step 1]. Use the HMW method to get initial estimates for all of the M log periodograms, each of which is denoted as $\widehat{f}_j$. [Step 2]. Average these to get the "intuitive" estimator $\overline{\widehat{f}}$.

$M = 8$		P_{10}	P_{20}	P_{50}	P_{80}	P_{90}	EMSE	σ
$(n = 256)$	Av	24.6	28.1	39.2	54.8	62.4	41.8	(15.7)
	LSA	15.1	17.7	24.5	34.7	39.7	26.4	(10.5)
$(n = 512)$	Av	12.3	14.7	19.2	25.7	29.5	20.4	(6.6)
	LSA	8.1	9.3	12.7	16.8	20.4	13.6	(5.2)
$M = 20$								
$(n = 256)$	Av	22.5	25.7	32.5	41.0	46.2	33.6	(9.3)
	LSA	7.7	8.7	11.3	15.2	17.7	12.2	(4.4)
$(n = 512)$	Av	11.4	12.9	16.0	20.3	22.9	16.7	(4.8)
	LSA	4.0	4.4	5.6	7.3	8.3	6.0	(1.9)

TABLE 3. Set up: least squares aggregate (LSA) method that is constructed with using only two initial estimates that are minimax is compared with the arithmetic average approach where all initial estimates are minimax. The values shown are the percentile values (P_v denotes the $v-$th percentile) of the MSEs, the mean and standard deviation of the MSEs based on 500 simulated data sets for the arithmetic average (Av) and the proposed method (LSA). The numerical values are in units of 10^{-3}.

Figure 2 reveals that a "typical" LSA estimate (i.e., curve estimate that correponds to the median MSE for LSA) is able to attend to the local minima better than a "typical" arithmetic average estimate. In Table 3, the simulation studies clearly indicates that LSA is superior to the arithmetic average of minimax estimates. Again, for $n = 512$, the "worst" behavior of the LSA is better than the "best" behavior of the Av. Moreover, as anticipated, we also gain in terms of the computing time. In the case where $M = 8, n = 256$, it took 37.6 seconds for the LSA that uses only two initial curves that are minimax versus 46.1 seconds for the arithmetic average that uses all initial estimates that are minimax. In fact, the LSA procedure kills two birds with one stone here: faster computing time and lower MSE than the arithmetic average.

4. Data example: explosion seismic waves

The seismic data set consists of the P-phases of 8 explosion recordings measured by stations located in Scandinavia (see Table 4). These explosion events are in the range of 2.13 to 2.19 on the

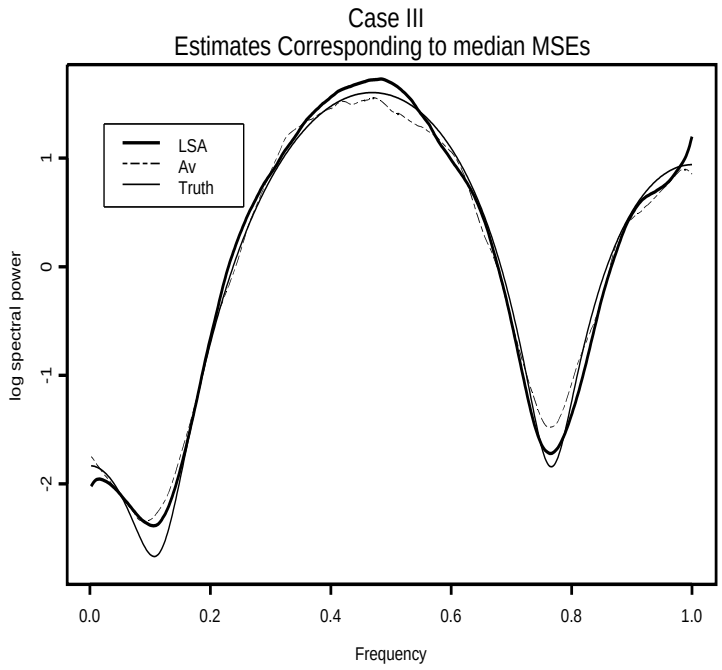

FIGURE 2. True log spectrum; the LSA estimate and the arithmetic average estimate that corresponds to the median of the MSEs for each of the methods. The x-axis is the frequency where 1 corresponds to the Nyquist frequency. The y-axis is the log spectral power.

Number	Type	Date	Magnitude	Latitude	Longitude
1	EXP	23 March 1991	2.85	69.2	34.3
2	EXP	13 April 1991	2.60	61.8	30.7
3	EXP	26 April 1991	2.95	67.6	33.9
4	EXP	03 August 1991	2.13	67.6	30.6
5	EXP	05 September 1991	2.32	67.1	21.0
6	EXP	10 December 1991	2.59	59.5	24.1
7	EXP	29 December 1991	2.96	69.4	30.8
8	EXP	25 March 1992	2.94	64.7	55.2

TABLE 4. P-phases of the 8 explosion seismic signals.

Richter scale. All the events chosen were on or near land and were distributed almost uniformly over Scandinavia. Each time series has length $T = 1024$, recorded for about 26 seconds and then sampled at 40 hertz. The motivation behind collecting these seismic recordings is to identify spectral features of explosion recordings and how they may differ from other types of seismic events such as earthquakes. These recordings are typical of the base data set in Blandford (1993) and is well studied and has been used to illustrate time series methods in Kakizawa, Shumway and

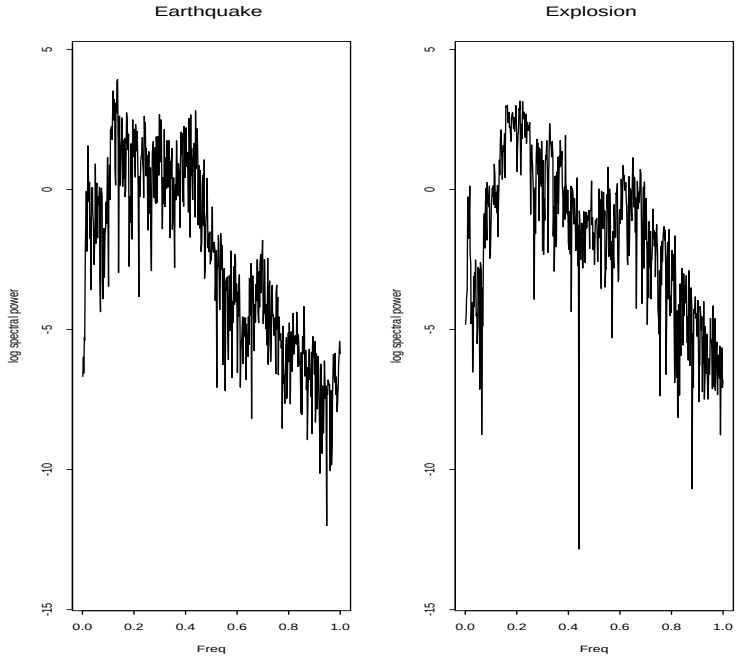

FIGURE 3. Log periodogram curves for the P-phases of the two (out of the 8) explosion events. The x-axis is the frequency. The actual range of $0 - 20$ hertz (Nyquist frequency) is rescaled to $[0, 1]$. The y-axis is the log spectral power.

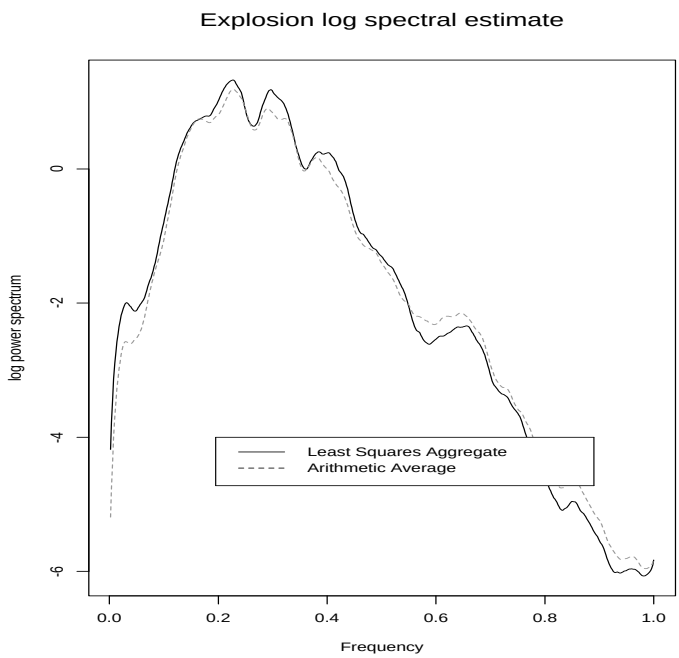

FIGURE 4. Estimates from the least squares aggregate estimate and the arithmetic average. The x-axis is the frequency. The actual range of $0 - 20$ hertz (Nyquist frequency) is rescaled to $[0, 1]$. The y-axis is the log spectral power.

Stoffer (1998) and Shumway and Stoffer (2000) and Huang, Ombao and Stoffer (2004). The log periodogram curves for the 8 seismic events are in Figure 3. As in Blandford (1993) and Kakizawa at al. (1998), seismic signals from the same event will be assumed to have been generated by a common process. We note that these data are observational and hence were not recorded under controlled experimental conditions. However, it is completely reasonable to assume that explosion signals share common spectral features. This approach has been shown by Shumway (1996) and Kakizawa, Shumway and Taniguchi (1998) to yield useful results. The least squares aggregation estimates and the arithmetic average estimates (Figure 4) are very similar. The purpose of this exercise was to compare the estimates of the arithmetic average of eight minimax initial estimates with the LSA estimate that was computed using two initial curve estimates that are minimax and six that are non-minimax (the bandwidth chosen for these six was an average of the optimal bandwidths of the two minimax estimates). It is interesting to note that the LSA, which demanded less computational time, gave similar global results with the more intensive arithmetic average. Moreover, the LSA estimate displays sharp peaks at frequencies $\lambda = 0.05, 0.1, 0.20$ which correspond to $1, 2, 4$ hertz respectively. Shumway and Stoffer (2000) discuss that the differences between types of seismic activities is most prominent at the frequency range of under 8 hertz. Thus, it is necessary for any method to be able to accurately estimate the power at these frequencies. Given how well the LSA performed in the simulations, we believe that these peaks are estimated well by the LSA procedure.

5. Conclusion

In this paper, we developed a statistical method for estimating the spectrum from a data set that consists of several signals, all of which are realizations of the same random process. Our method aggregates the spectral estimates from each signal by using weights that are obtained adaptively from the data via a least squares criterion. The procedure is independent of the way in which the input estimators are constructed. Theorem 2.1 and Corollary 2.2 establish theoretical properties of our aggregates. Theorem 2.1*(a)* shows that least squares aggregation yields estimators the empirical risk of which achieves the optimal linear aggregation bound. This result is a finite sample result, and it holds for *arbitrary* estimators, under no boundedness or smoothness assumptions; also, no conditions on f are needed. In Section 2.3 we investigate the rate of convergence of our

aggregates. Corollary 2.2 shows that if a minimum of *two* of the M input spectral estimates are minimax adaptive, then the aggregate is minimax adaptive. In contrast, the arithmetic average of M estimates is minimax rate optimal if *all* M estimates are minimax. This validates our algorithm in a variety of scenarios, as further illustrated by our simulations section. In particular, Section 3.2 above shows that our aggregate exhibits an excellent fit even in the extreme case when only 2 estimators are minimax adaptive and the rest are deliberately chosen to render extremely poor fits. In such cases, our proposed aggregate estimator is markedly superior to a naive estimate such as the arithmetic average. Also, the MSE of the least squares aggregate estimator has a lower standard deviation and is, therefore, a more stable procedure than the arithmetic average. Moreover, in many cases, even the "worst" behavior of the aggregate estimator is better than the "best" behavior of the arithmetic average. In addition, as illustrated in Section 3.3, our algorithm can be combined with minimax adaptive methods to yield estimates that are faster to compute and have lower mean squared errors than a simple average.

Appendix

Proof of Theorem 2.1. For clarity of exposition we shall drop the subscript (m) in the course of this proof, and we will re-denote $\widetilde{f}^{(m)}$ by $\widetilde{f}$, $\widehat{\lambda}^{(m)}$ by $\widehat{\lambda}$ and $Y^{(m)}$ by Y. Let

$$D_{-m} = C^{(1)} \cup \ldots \cup C^{(m-1)} \cup C^{(m+1)} \cup \ldots \cup C^{(M)}.$$

We always have $E\|\tilde{f} - f\|_n^2 = E\left(E\left(\|\tilde{f} - f\|_n^2 | D_{-m}\right)\right)$. In what follows, we bound the inner expectation. Notice that conditionally on D_{-m} the estimators $\widehat{f}_1, \ldots, \widehat{f}_{m_1}, \widehat{f}_{m+1}, \ldots, \widehat{f}_M$ can be treated as fixed functions. Thus, still conditionally on D_{-m}, only the weights $\widehat{\lambda}$ are random in $\widetilde{f}(\nu_k) = \sum_{j \in J_{-m}} \widehat{\lambda}_j \widehat{f}_j(\nu_k)$, since they have been obtained on data set C^m. Let now $\hat{\mathbf{f}}_j = (\hat{f}_j(\nu_1), \ldots, \hat{f}_j(\nu_n)) \in \mathbb{R}^n$, for $j \in J_{-m}$. Denote by S_{-m} the subspace of $\mathbb{R}^n$ generated by $\left\{\hat{\mathbf{f}}_j\right\}_{j \in J_{-m}}$. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_{M'}\}$, $M' \leq M - 1$, be a basis in S_{-m} that is orthonormal with respect to the Euclidean inner product in $\mathbb{R}^n$, that is $\sum_{k=1}^n \mathbf{v}_{kj}\mathbf{v}_{kl} = \delta_{jl}$, where $\delta_{jl} = 1$ if $j = l$ and zero otherwise. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$. By definition (2.1), our estimator satisfies

$$\sum_{k=1}^n \left[Y_k - \widetilde{f}(\nu_k)\right]^2 \leq \sum_{k=1}^n [Y_k - \sum_{j \in J_{-m}} \lambda_j \widehat{f}_j(\nu_k)]^2,$$

for any $\lambda \in \mathbb{R}^{M-1}$, thus $\tilde{\mathbf{f}} = (\tilde{f}(\nu_1), \ldots, \tilde{f}(\nu_n))$ is the projection of $\mathbf{Y}$ onto S_{-m} with respect to the Euclidean inner product. Thus

$$\tilde{\mathbf{f}} = \sum_{j=1}^{M'} \hat{\gamma}_j \mathbf{v}_j, \qquad \text{where } \hat{\gamma}_j = \sum_{k=1}^{n} Y_k \mathbf{v}_{kj}. \tag{5.1}$$

Define now f_γ such that

$$\|f - f_\gamma\|_n^2 = \inf_{\lambda \in \mathbb{R}^{M-1}_{-m}} \|f - f_\lambda\|_n^2,$$

where we recall that $f_\lambda = \sum_{j \in J_{-m}} \lambda_j \hat{f}_j$. Thus, $\mathbf{f}_\gamma = (f_\gamma(\nu_1), \ldots, f_\gamma(\nu_n))$ is the Euclidean projection of $\mathbf{f} = (f(\nu_1), \ldots, f(\nu_n))$ onto S_{-m}. Then we can write

$$\mathbf{f}_\gamma = \sum_{j=1}^{M'} \gamma_j \mathbf{v}_j, \qquad \text{where } \gamma_j = \sum_{k=1}^{n} f(\nu_k) \mathbf{v}_{kj}. \tag{5.2}$$

By Pythagoras theorem we have

$$n\|f - \tilde{f}\|_n^2 = n\|f - f_\gamma\|_n^2 + n\|\tilde{f} - f_\gamma\|_n^2, \tag{5.3}$$

by noticing that $n\|g\|_n^2 = \|\mathbf{g}\|_2^2$, for any function g and corresponding vector $\mathbf{g} = (g(\nu_1), \ldots, g(\nu_n))$, and where $\| \ \|_2$ is the Euclidean norm in $\mathbb{R}^n$. Notice further that we have $\|\tilde{\mathbf{f}} - \mathbf{f}_\gamma\|_2^2 = \sum_{j=1}^{M'} (\hat{\gamma}_j - \gamma_j)^2$, since the basis is orthonormal. By (5.3) we then have

$$\begin{aligned} E\left(\|\tilde{f} - f\|_n^2 | D_{-m}\right) &= E\left(\inf_{\lambda \in \mathbb{R}^{M-1}} \|f - f_\lambda\|_n^2 | D_{-m}\right) + \frac{1}{n} \sum_{j=1}^{M'} E\left((\hat{\gamma}_j - \gamma_j)^2 | D_{-m}\right) \\ &\leq \inf_{\lambda \in \mathbb{R}^{M-1}} E\left(\|f - f_\lambda\|_n^2 | D_{-m}\right) + \frac{1}{n} \sum_{j=1}^{M'} E\left((\hat{\gamma}_j - \gamma_j)^2 | D_{-m}\right) \\ &= \inf_{\lambda \in \mathbb{R}^{M-1}} E\left(\|f - f_\lambda\|_n^2 | D_{-m}\right) + \frac{1}{n} \sum_{j=1}^{M'} \mathrm{Var}\left(\hat{\gamma}_j | D_{-m}\right), \end{aligned} \tag{5.4}$$

using (5.1) and (5.2), in connection with the fact that $E(Y_k) = f(\nu_k)$, to deduce that $E(\hat{\gamma}_j | D_{-m}) = \gamma_j$. Notice that

$$\mathrm{Var}\left(\hat{\gamma}_j | D_{-m}\right) = \mathrm{Var}\left(\sum_{k=1}^{n} Y_k \mathbf{v}_{kj} | D_{-m}\right) = \sigma^2 \sum_{k=1}^{n} \mathbf{v}_{kj}^2 = \sigma^2,$$

and the proof is concluded by taking the outer expectation and recalling that $M' \leq M$. ■

References

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, **96** , 939-967.

Audibert, J.-Y. (2003) Aggregated estimators and empirical complexity for least square regression.*Prépublication, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2003.*

Baraud, Y. (2000). Model selection for regression on a fixed design. *Probability Theory and Related Fields* , **117** , 467 - 493.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probability & Statistics*, **7**, 127 – 146.

Barron, A., Birgé, L., Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113**, 301 – 413.

Birgé, L. (2003). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Prépublication n.862, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7, http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2003.*

Blackman, R. and Tukey, J. (1958a). The measurement of the power spectra from the point of view of communications engineering, Part I. *Bell System Technical Journal*, **37**, 185-282.

Blackman, R. and Tukey, J. (1958b). The measurement of the power spectra from the point of view of communications engineering, Part II. *Bell System Technical Journal*, **37**, 485-569.

Blandford, R. (1993). Discrimination of Earthquakes and Explosions at Regional Distances Using Complexity. *AFTAC-TR-93-044 HQ, Air Force Technical Applications Center, Patrick Air Force Base, Florida.*

Booker, A. and Mitrovonas, W. (1964). An Application of Statistical Discrimination to Classify Seismic Events. *Bulletin of the Seismological Society of America,* **54**, 951-971.

Brillinger, D. (1981). *Time Series Analysis: Data Analysis and Theory.* New York: Holt, Rinehart and Winston.

Brockwell, P. and Davis, R. (1991). *Time Series: Theory and Methods.* New York: Springer.

Bunea, F. (2004). Penalty choices and consistent covariate selection in semiparametric regression. *Annals of Statistics*, **32**, 898-927.

Buysse, D., Hall, M., Begley, A., Cherry, C., Land, S., Ombao, H., Kupfer, E. and Frank, E. (2001). REM Sleep and Treatment Response in Depression: New Findings Using Power Spectral Analysis. *Psychiatry Research*, **103**, 51-67.

CATONI, O. (2001). *Statistical Learning Theory and Stochastic Optimization.* Ecole d'Eté de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics, Springer, New York.

DAVILA, C. AND MOBIN, M. (1992). Weighted Averaging of Evoked Potentials. *IEEE Transactions on Biomedical Engineering*, **39**, 338-345.

DAVIS, H. AND JONES, R. (1968). Estimation of the variance of the a stationary time series. *Journal of the American Statistical Association,* **63**, 141-149.

DEVORE, R. A. AND LORENTZ, G. G. (1993) Constructive Approximation, Springer-Verlag.

DONOHO, D. L., JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425 - 455.

DONOHO, D. L., JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics* , **26** , 879-921.

GOTMAN, J. (1982). Automatic Recognition of Epileptic Seizures in the EEG. *Electroencephalography and Clinical Neurophysiology,* **54**, 530-540.

HARMONY, T., FERNANDEZ, T., SILVA, J., BOSCH, J., VALDES, P., FERNANDEZ-BOUZAS, A., GALAN, L., AUBERT, E. AND RODRIGUEZ, D. (1999). Do Specific EEG Frequencies Indicate Different Processes During Mental Calculation? *Neuroscience Letters,* **266**, 25-58.

HENGARTNER, N.W., MATZNER-LØBER, E., AND WEGKAMP, M.H. (2002). Bandwidth selection for local linear regression. *Journal of the Royal Statistical Society, Series B* , **64** , 1 - 14.

HUANG, H-Y., OMBAO, H. AND STOFFER, D. (2004). Classification and Discrimination of Non-Stationary Time Series Using the SLEX Model. *Journal of the American Statistical Association,* **99**, 763-774.

ISHIHARA, T. AND YOSHII, N. (1972). Multivariate analytic study of EEG mental activity in juvenile delinquents. *Electroencephalography and Clinical Neurophysiology,* **33**, 71-80.

KAKIZAWA, Y., SHUMWAY, R. AND TANIGUCHI, M. (1998). Discrimination and Clustering for Multivariate Time Series. *Journal of the American Statistical Association*, **93**, 328-340.

KOLTCHINSKII, V. (2004). Local Rademacher complexities and oracle inequalities in risk minimization. *Manuscript.*

JUDITSKY, A. AND NEMIROVSKI, A. (2000). Functional aggregation for nonparametric regression. *Annals of Statistics*, **28**, 681–712.

Leung, G. and Barron, A.R. (2004) Information theory and mixing least-squares regressions. *Manuscript.*

Leski, J. (2002). Robust weighted averaging. *IEEE Transactions on Biomedical Engineering,* **49**, 796-804.

Lugosi, G. and Nobel, A. (1999). Adaptive model selection using empirical complexities. *Annals of Statistics* , **27** , 1830–1864 .

Nemirovski, A. (2000). Topics in non-parametric statistics, Ecole d'Eté de Probabilités de Saint-Flour - 1998, *Lecture Notes in Mathematics*, Springer-Verlag, New York, **XXVIII** .

Ombao, H., Raz, J., Strawderman, R., and von Sachs, R. (2001). A simple GCV method of span selection for periodogram smoothing. *Biometrika*, **88** , 1186-1192.

Shumway, R. (1996). Statistical Approaches to Seismic Discrimination. In *Monitoring a Comprehensive Test Ban Treaty,* pp. 791-803. A.M. Dainty and E.S. Husebye eds. Doordrecht, The Netherlands: Kluwer Academic.

Shumway, R. and Stoffer, D. (2000). *Time Series Analysis and Its Applications.* New York: Springer.

Stoffer, D., Scher, M., Richardson, G., Day, N. and Coble, P. (1988). A Walsh-Fourier Analysis of the Effects of Moderate Maternal Alcohol Consumption on Neonatal Sleep-State Cycling. *Journal of the American Statistical Association,* **83**, 954-963.

Tsybakov, A.B. (2003) Optimal rates of aggregation . *Lecture Notes in Artificial Intelligence, Proceedings of 16th Annual Conference on Learning Theory (COLT) and 7th Annual Workshop on Kernel Machines*, Springer-Verlag, Heidelberg , **2777** .

Wegkamp, M.H. (2003). Model selection in nonparametric regression. *Annals of Statistics* , **31** , 252 - 273.

Wahba, G. (1980). Automatic Smoothing of the Log Periodogram. *Journal of the American Statistical Association.* **75**, 122-132.

Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, **74**, 135 – 161.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of American Statistical Association*, **96**, 574 – 588.

Yang, Y. (2004). Aggregating regression procedures for a better performance. *Bernoulli*, **10**, 25 – 47.