From Multiple Gaussian Sequences to Functional Data and Beyond: A Stein Estimation Approach

Mark Koudstaal

Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Fang Yao

Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Summary. We expand the notion of Gaussian sequence model to n experiments and propose a Stein estimation strategy which relies on pooling information across experiments. An oracle inequality is established to assess conditional risks given the underlying effects, based on which we are able to quantify the size of relative error and obtain a tuning-free recovery strategy that is easy to compute, produces model parsimony and extends to unknown variance. We show that the simultaneous recovery is adaptive to an oracle strategy, which also enjoys a robustness guarantee in a minimax sense. A connection to functional data is established, via Le Cam theory, for fixed and random designs under general regularity settings. We further extend the model projection to general bases with mild conditions on correlation structure, and conclude with potential application to other statistical problems. Simulated and real data examples are provided to lend empirical support to the proposed methodology and illustrate the potential for substantial computational savings.

Keywords: Functional data; Le Cam equivalence; Conditional oracle inequality; Nonparametric regression; Simultaneous recovery; Wavelets

1. Introduction

The Gaussian sequence model (GSM) occupies an important role in modern statistics, providing valuable insights on nonparametric estimation of an unknown function. See Candes (2006) and Johnstone (2015) for introductions and overviews. To be concrete, a sequence of signals or effects $\boldsymbol{\theta} = (\theta_k)_{k\in\mathbb{N}}$ are observed in Gaussian white noise,

$$Y_k = \theta_k + \sigma m^{-1/2} z_k, \;\; k \in \mathbb{N}, \tag{1}$$

where $z_k \overset{i.i.d.}{\sim} N(0,1)$ and σ is a constant with calibration $m^{-1/2}$. Despite simplicity, this model has broad implications for general estimation and testing problems. On one hand, it provides a framework which has minimal technical complication in various parameter spaces. On the other hand, results in this framework have an intrinsic connection to nonparametric estimation of an unknown function f from data

$$y_i = f(x_i) + \sigma z_i, \;\; z_i \overset{i.i.d.}{\sim} N(0,1), \;\; x_i \in [0,1], \;\; i = 1, \ldots, m, \tag{2}$$

in the limit $m \to \infty$. This follows from Le Cam equivalence between (2) and the white noise model

$$Y(dt) = f(t)dt + \sigma m^{-1/2} W(dt), \;\; t \in [0,1], \tag{3}$$

W being a standard brownian motion. This connection has been studied by Brown and Low (1996), Brown *et al.* (2002) and Reiß (2008), among others. The white noise model

(3) acts on an orthonormal basis $\{\psi_k\}_{k\in\mathbb{N}}$ of $L^2[0,1]$ to give rise to the GSM (1) with $\theta_k = \langle f, \psi_k\rangle$. For any estimator $\hat{f}$ of f, isometry leads to $\mathbf{E}||f-\hat{f}||_{L_2}^2 = \sum_{k=1}^{\infty}\mathbf{E}(\theta_k - \hat{\theta}_k)^2$, where $\hat{\theta}_k = \langle \hat{f}, \psi_k\rangle$. This reduces the problem of estimating f under L^2 loss to the problem of estimating a sequence of normal means $\boldsymbol{\theta} = (\theta_k)_{k\in\mathbb{N}}$ under ℓ_2 loss. The appeal of this framework is that a function of practical interest often has a natural characterization in terms of geometric constraints on its (generalized) Fourier coefficients θ_k in a suitable basis, and may be grouped into a collection $\mathcal{F}$ of possible generating mechanisms for (2). By distilling the central issues at play, reduction to (1) has inspired many estimation procedures with adaptivity properties. See Donoho (1993), Donoho *et al.* (1995), Cai (1999), Cavalier and Tsybakov (2002) and Zhang (2005) for original work in this direction and Candes (2006), Cai (2012) and Johnstone (2015) for comprehensive overviews. This framework has also facilitated understanding of frequentist and Bayesian properties of simple Bayesian nonparametric models. See Freedman (1999), Zhao (2000), Belitser and Ghosal (2003) and Szabó *et al.* (2013) for studies in this direction.

Modern scientific experiments are often conducted simultaneously with data sampled from multiple "similar" functions, such as images and voice signals, which motivates grouping together GSMs corresponding to individual experiments. In this paper, we expand on the notion of GSM (1) to study recovery of multiple sequences, namely the multiple GSMs of size n,

$$Y_{ik} = \theta_{ik} + \sigma m^{-1/2} z_{ik}, \;\; k\in\mathbb{N},\, i = 1,\ldots,n \tag{4}$$

where $z_{ik} \overset{i.i.d.}{\sim} N(0,1)$. This can be viewed as an idealization of observing n nonparametric experiments,

$$y_{ij} = f_i(x_{ij}) + \sigma z_{ij}, \;\; x_{ij}\in[0,1], \;\; i = 1,\ldots,n,\, j = 1,\ldots,m, \tag{5}$$

corresponding to the central model of functional data analysis (FDA) which has attracted considerable interest in the past decades. See Ramsay and Silverman (2005) for an introduction and examples.

The main contributions of this paper are two-fold. The first part focuses on simultaneous recovery of the effects $\{\boldsymbol{\theta}_i\}_{i\le n}$, where $\boldsymbol{\theta}_i = (\theta_{ik})_{k\in\mathbb{N}}$, from multiple GSMs (4). A form of Stein estimation, based on information pooling across experiments, is proposed and its properties are derived with the aid of new concentration results. The method is shown to attain the optimal rate for recovery of n experiments in a uniform manner, and enjoys a robustness guarantee in a minimax sense. Moreover, the theoretical analysis suggests an explicit tuning-free form which governs the amount of shrinkage parsimoniously under a general condition $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2} \to \infty$, for any $\gamma_2 \ge \gamma_1 > 0$, while prior knowledge of the speed and ordering of decay in the variances of the effects are not required. Here $\alpha_n \lesssim \beta_n$ denotes $\alpha_n/\beta_n = O(1)$ for real sequences $\{\alpha_n\}$ and $\{\beta_n\}$. Further, our development extends elegantly to the case of unknown variance, suggesting a simple yet effective estimator of σ^2 and revealing a risk transition phenomenon associated with model complexity. We emphasize that the traditional theory for Stein estimation does not lead to the same results for the proposed method, which controls the average risk $\mathbf{E}||\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i||_{\ell_2}^2$ and fails to provide individual recovery guarantees for a given sample of sequences.

The second thread of contributions establishes a rigorous connection between the multiple GSMs (4) and the functional data model (5) via Le Cam asymptotic equivalence for both fixed and random design, which covers standard cases considered in the functional data literature (Hall *et al.*, 2006; Cai and Yuan, 2011). We further show that the theoretical guarantees of the proposed recovery continue to hold for bases on which the projected

coefficients exhibit decaying correlations, and present concrete examples of process/basis pairs satisfying such assumptions. Although useful in many areas, common FDA methods based on nonparametric smoothing suffer from some drawbacks. Chief among these are complicated theoretical properties and reliance on computationally expensive procedures. Hence a key motivation for studying multiple GSMs (4) is to provide a simplified but foundational framework for functional data which might encourage development of new methods with properties that are both easy to establish and relate transparently to other areas of statistics. On the computational side, the proposed method operates in $O(nm \log m)$ time, in contrast to standard smoothing-based FDA techniques that typically operate in at least $O(nm^2 + m^3)$ time (Ramsay and Silverman, 2005; Yao *et al.*, 2005). This implies potentially significant computational savings and scalability to data of large dimensions/sizes. Further, our procedure may be employed in an online algorithm fashion: a new curve comes in, transform is performed, threshold weights are updated, which makes our method potentially useful in the context of realtime data collection and processing.

The rest of the article is organized as follows. In section 2 we draw inspiration from an oracle strategy known to achieve optimal recovery rates under the conditional ℓ_2 risk metric. Risks of the proposed Stein estimation are related to those of the oracle strategy via a set of concentration inequalities on the conditional measures. This leads to a theory of simultaneous recovery which gives a precise account of the shrinkage and extends seamlessly to the case of unknown variance. In Section 3, we make the connection to the functional data model (4) through Le Cam asymptotic equivalence and extrapolate our recovery theory to more general correlation settings that one might encounter in practice. In Section 4, we first present a simulation study to support the recovery method and its theoretical properties in the setting of multiple GSMs, then demonstrate its performance and computational gains using the Phoneme dataset (Hastie *et al.*, 1995). The computer codes and data for reproducing the results are available at *http://www.utstat.utoronto.ca/fyao/GSM-FDA-code.zip*. We conclude in Section 5 with discussion on potential application to other statistical problems that may deserve further investigation, such as multiple change-point detection for penal data. Due to space constraint, we collect proofs of all theoretical results and some additional simulation results in the Supplementary Material.

2. Multiple Gaussian Sequences and Stein Estimation

2.1. Problem setting and objective

Before moving forward, we outline some notation used throughout. For a function $f : \mathcal{D} \to \mathbb{R}$, mapping some domain $\mathcal{D}$ into $\mathbb{R}$, we let $\mathbf{Supp}(f)$ denote the support of f, i.e. $\mathbf{Supp}(f) = \{x \in \mathcal{D} : f(x) \neq 0\}$. We let '$\perp$" denote statistical independence, $(x)_+ = \max(0, x)$ for $x \in \mathbb{R}$ and for $x \in \mathbb{R}^m$ with positive components, we let $x_{(j)}$ denote the non-increasing order statistics of the coordinates, such that $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(m)}$. For $n \in \mathbb{N}$ we let $[n] = \{1, \ldots, n\}$ and for two sequences of real numbers, (α_n) and (β_n), $\alpha_n \approx \beta_n$ stands for $\alpha_n/\beta_n \to 1$, $\alpha_n \ll \beta_n$ for $\alpha_n/\beta_n \to 0$, $\alpha_n \gg \beta_n$ for $\alpha_n/\beta_n \to \infty$ and $\alpha_n \propto \beta_n$ denotes $0 < |\alpha_n/\beta_n| < \infty$ as $n \to \infty$. For $\boldsymbol{x} \in \ell_2$ and $f : \mathbb{R} \to \mathbb{R}$, let $f(\boldsymbol{x})$ denote $f(\boldsymbol{x}) = (f(x_1), f(x_2), \ldots) \in \mathbb{R}^\infty$, and for vectors $\boldsymbol{x}$, $\boldsymbol{y} \in \ell_2$, set $||\boldsymbol{x}/\boldsymbol{y}||_{n,\infty} = \max_{k \leq n} |x_k/y_k|$. For an array of $(Y_{ij})_{1 \leq i \leq n, j \in \mathbb{N}}$, let $Y_{\cdot j}$ and $Y_{i\cdot}$ denote the vectors $Y_{\cdot j} = (Y_{1j}, \ldots, Y_{nj})^T \in \mathbb{R}^n$ and $Y_{i\cdot} = (Y_{i1}, Y_{i2}, \ldots)^T \in \mathbb{R}^\infty$.

In the spirit of functional data, where f_i are taken as independent realizations from a stochastic process, we place a distributional structure on multiple GSMs (4) to impose

similarity of θ_{ik} across i for given k. A common model in the nonparametric literature on GSMs (1) takes $\lambda_k^{-1/2}\theta_k \overset{i.i.d.}{\sim} N(0,1)$ independent of z_k, with $\lambda_k = 2\alpha k^{-(2\alpha+1)}$ and $\alpha > 0$. We use this as our model for (4), with $\theta_{1k}, \ldots, \theta_{nk} \overset{i.i.d.}{\sim} \theta_k$ independent of the z_{ik}, but relax the variance decay to

$$\lambda_{(k)} \propto k^{-(2\alpha+1)},\ k \leq m; \quad \lambda_k \propto k^{-(2\alpha+1)},\ k > m. \tag{6}$$

Thus the bulk of each signal is contained in the first m coordinates, but the location and ordering of sizeable effects is unknown *a priori*. We give a graphical demonstration in Section 4.1 showing that this relaxation is suitable for modelling functions with striking local features from a smoothness perspective in a similar manner as wavelet estimation in nonparametric regression (Donoho and Johnstone, 1994). In what follows, we let $\mathbf{E}_i(\cdot) = \mathbf{E}_{\boldsymbol{\theta}_i}(\cdot) = \mathbf{E}(\cdot|\boldsymbol{\theta}_i)$ denote expectation conditional on the ith effect.

REMARK 2.1. The regularity setting (6) can be regarded as randomization of weak ℓ_p decay conditions (up to m) with their origin in Donoho (1993) who noted that they are closely related to Besov smoothness in the context of wavelet coefficients. Standard properties of Gaussian variables lead to

$$\max_k k^{1/p}|\theta|_{(k)} \lesssim \max_k k^{1/p-(\alpha+1/2)}\{\log(1+k)\}^{1/2}$$

and so we are in every weak ℓ_p space for $p > 2/(2\alpha+1)$, with $|\theta_k|/(\log(1+k))^{1/2}$ lying in $p = 1/(2\alpha+1)$, where a rigorous argument is given in the Supplementary Material. These spaces are known to form important generalizations of the traditional Hölder and Sobolev type smoothness spaces. As discussed in Donoho (1993) and Candes (2006), once a suitable basis is specified, such decay conditions set the frontier of statistical recovery at nonlinear approximation spaces.

REMARK 2.2. We assume a centred model (5) for functional data that corresponds to mean-zero GSMs (4), as research in FDA mostly focuses on characterizing random realizations and covariance structure of an underlying process. Estimation of the mean function is usually considered an easier task by standard means of nonparametric regression, such as kernel-type (Yao *et al.*, 2005; Li and Hsing, 2010) or spline-type (Ramsay and Silverman, 2005) smoothing which attains a univariate convergence rate that is asymptotically negligible relative to simultaneous recovery or covariance estimation. Hence the proposed methodology can be applied by subtracting the estimated mean function (or simply the cross-sectional mean when observed on a common grid).

We now relate simultaneous recovery of $\boldsymbol{\theta}_i = (\theta_{ik})_{k\in\mathbb{N}}$ for $i = 1, \ldots, n$, to an oracle framework. Let $\mathbf{E}(\cdot|\boldsymbol{\theta}_i) = \mathbf{E}_{\boldsymbol{\theta}_i}(\cdot) = \mathbf{E}_i(\cdot)$ denote conditional expectation and define the conditional ℓ_2 risk of an estimator $\hat{\boldsymbol{\theta}}_i$ by

$$\mathcal{R}_{i,m}(\hat{\boldsymbol{\theta}}_i) \triangleq \mathbf{E}_i||\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i||_{\ell_2}^2. \tag{7}$$

Our goal is to devise an estimation strategy $\{\hat{\boldsymbol{\theta}}_i\}_{i\leq n}$ faithful to the effects $\{\boldsymbol{\theta}_i\}_{i\leq n}$ by maintaining small conditional ℓ_2 risks in a uniform manner. This corresponds to controlling "curve-wise" risks for recovering all functions in a given sample, which is a useful measure for functional data. It is known that the "ideal" linear oracle rule $\hat{\theta}_{ik}^{o;c} = \theta_{ik}^2 Y_{ik}/(\theta_{ik}^2 + 1/m)$ knowing the effects yields the minimal risk among diagonal linear rules. Since it is unrealistic

to mimic this ideal oracle beyond minimax performance, a simple calculation gives its conditional risks $\mathcal{R}^c_{i,m} = \sum_{k=1}^{\infty}(\theta_{ik}^2/m)/(\theta_{ik}^2 + 1/m)$, and a similar argument as showing Theorem 2 yields that $\max_{i\le n}\mathcal{R}^c_{i,m}$ concentrates at $\mathbf{E}\mathcal{R}^c_{i,m}$. A further calculation indicates that $\mathbf{E}\mathcal{R}^c_{i,m}$ is within a factor of the optimal average risk attained by the linear oracle $\hat{\theta}^{o,a}_{ik} = \lambda_k Y_{ik}/(\lambda_k + \sigma^2/m)$, i.e., $\mathbf{E}\|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i^{o,a}\|^2_{\ell_2} = \sum_{k=1}^{\infty}(\lambda_k/m)/(\lambda_k + 1/m) \propto m^{-2\alpha/(2\alpha+1)}$. We show in Section 2.3 that the conditional risks of this linear oracle, $\mathcal{R}^*_{i,m} \triangleq \mathbf{E}_i\|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i^{o,a}\|^2_{\ell_2}$, can be controlled uniformly near its average. From this perspective, no procedure can do significantly better than $\hat{\theta}^{o,a}_{ik}$ which is rate optimal among diagonal linear rules. This class includes strategies that may pool information from underlying distribution of the effects. Further, these rules are known to perform within a small factor of the minimax optimal estimators over a broad range of parameter spaces (Donoho *et al.*, 1990). Such considerations motivate the search for recovery strategy $\hat{\boldsymbol{\theta}}_i$ performing as well as the average case oracle $\hat{\boldsymbol{\theta}}_i^{o,a}$.

2.2. Stein estimation motivated by conditional concentration

The main idea guiding this paper is that models generated by similar experiments puts us in an empirical Bayes type setting where information pooling improves estimation. In this setting, concentration of measure can be used to guide design of estimators and quantify how information pooling improves estimation measured by risks conditioned on the effects. For the moment we take $\sigma^2 = 1$, treating the case of unknown variance in Section 2.6. Note that the average case oracle $\hat{\theta}^{o,a}_{ik}$ can be written equivalently as $\hat{\theta}^{o,a}_{ik} = [1 - (n/m)/\{n(\lambda_k + 1/m)\}]_+ Y_{ik}$. Thus, when the relative error incurred in approximating $n(\lambda_k + \sigma^2/m)$ by $||Y_{\cdot k}||^2$ is sufficiently small, it is natural to consider standard (positive part) Stein estimation

$$\hat{\theta}^S_{ik} = \begin{cases} \beta_{mn,k}Y_{ik}, & k \le m \\ 0, & k > m \end{cases}, \quad \beta_{mn,k} = \left(1 - \frac{n/m}{\|Y_{\cdot k}\|^2}\right)_+. \tag{8}$$

Similar ideas have been employed in the nonparametric setting (2) by Cai (1999), Cavalier and Tsybakov (2002) and Zhang (2005) from different perspectives.

In this and next subsection, we explore the intuition and theory to arrive at the proposed Stein estimation strategy that is different from (8),

$$\hat{\theta}^{RS}_{ik} = \begin{cases} \alpha_{mn,k}Y_{ik}, & k \le m \\ 0, & k > m \end{cases}, \quad \alpha_{mn,k} = \left\{1 - (1 + 2\sqrt{12\log m/n})\frac{n/m}{||Y_{\cdot k}||^2}\right\}_+. \tag{9}$$

We begin with a new concentration of measure result, conditional on the effects, which is used to assess the relative error of approximating $n(\lambda_k + \sigma^2/m)$ by $||Y_{\cdot k}||^2$, see the Supplementary Material for its proof. For $\delta \in (0,1)$, we consider collections of arrays whose components satisfy norm constraints indexed by δ. Specifically, given $\lambda_1, \dots, \lambda_m$, define the sets $A_{k,\delta}$, $k = 1, \dots, m$, and A^m_δ by complement as follows,

$$A^c_{k,\delta} = \{(1-\delta)n(\lambda_k + 1/m) \le ||Y_{\cdot k}||^2 \le (1+\delta)n(\lambda_k + 1/m)\} \tag{10}$$

and set $A^{m,c}_\delta = \bigcap_{k\le m} A^c_{k,\delta}$. Denote $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots)$ and $\mathbf{P}_i(\cdot) = \mathbf{P}(\cdot|\boldsymbol{\theta}_i)$.

LEMMA 1. *Consider multiple GSMs (4) with the decay assumption (6), for all* $\delta \in (0, 1/2)$ *and* $i = 1, \ldots, n$*, one has*

$$\begin{aligned} \mathbf{P}_i(A_{k,\delta}) &\leq 3\exp(\delta||\boldsymbol{\theta}_i^2/\boldsymbol{\lambda}||_{m,\infty}) \cdot \exp(-n\delta^2/6), \quad (11) \\ \mathbf{P}_i(A_\delta^m) &\leq 3\exp(\delta||\boldsymbol{\theta}_i^2/\boldsymbol{\lambda}||_{m,\infty}) \cdot \exp(-n\delta^2/6 + \log m). \quad (12) \end{aligned}$$

This result inspires a different Stein estimates (8) and sets the stage for bounding the conditional risks of the proposed estimator. Intuitively, for "most" effects, it is seen that $||\boldsymbol{\theta}_i^2/\boldsymbol{\lambda}||_{m,\infty}^2 \approx 2\log m$. Thus, if $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ for any $\gamma_2 \geq \gamma_1 > 0$, taking $\delta^2 = C\log m/n$ with $C > 6$ guarantees that the conditional probabilities in (12) are small while $\delta \to 0$ as $m, n \to \infty$.

To understand the behaviour of the traditional Stein estimate (8) in multiple GSMs, note that $\|Y_{\cdot k}\|^2 \sim (\lambda_k + 1/m)\chi^2(n)$ with $\mathbf{E}\|Y_{\cdot k}\|^2 = n(\lambda_k + 1/m)$. Since $\lambda_k \propto k^{-(2\alpha+1)}$ tends to zero quickly, once $k > m^{1/(2\alpha+1)}$ the signals decay rapidly below the noise level and in this range $||Y_{\cdot k}||^2 \approx \chi^2(n)/m$. Based on this intuition, we expect approximately half of the $\{||Y_{\cdot k}||^2\}_{k\leq m}$ to exceed n/m with corresponding standard Stein weights $\beta_{mn,k} > 0$. However, given the fast decay of λ_k, we expect that only on the order of $m^{1/(2\alpha+1)}$ of the $\{Y_{ik}\}_{k\leq m}$ contain sizeable signals, while the rest mainly contain noise. Further, when the norms fluctuate within the regime of high probability and $\lambda_k \ll 1/m$, many of the $\|Y_{\cdot k}\|^2 \approx (1+\delta)n(\lambda_k + 1/m)$, with $\delta = \sqrt{C\log m/n}$. This gives

$$\left|\beta_{mn,k} - \frac{\lambda_k}{\lambda_k + 1/m}\right| \Big/ \frac{\lambda_k}{\lambda_k + 1/m} \approx \frac{\delta}{m\lambda_k},$$

which can be arbitrarily large due to the rapid decay of λ_k. In other words, we do better by forcing most of $\beta_{mn,k}$ to zero in order to mimic the oracle weights in the range $\lambda_k \ll 1/m$.

This motivates a different Stein threshold in (8). Note that, on the event $A_\delta^{m,c}$, one has $||Y_{\cdot k}||^2 \leq (1+\delta)n(\lambda_k + 1/m)$. Thus for $\lambda_k \ll 1/m$, i.e., $k \gg m^{1/(2\alpha+1)}$, one expects that $||Y_{\cdot k}||^2 \lesssim (1+\delta)n/m < (1+2\delta)n/m$. In light of this, we propose to lift the threshold level to $(1+2\delta)n/m$ in order to force most weights in this range to zero, which leads to the new Stein weights $\alpha_{mn,k} = \left[1 - \{(1+2\delta)n/m\}/||Y_{\cdot k}||^2\right]_+$ for $k \leq m$. The impact of δ on estimation quality is precisely quantified in an extended version of Theorem 2, presented in the supplementary material, which suggests the explicit form in (9). Since Lemma 1 guarantees that, conditional on the effects $\boldsymbol{\theta}_i$, the events $A_\delta^{m,c}$ encompass most of the probability space with the right choice of $\delta \to 0$, we expect that the proposed strategy retains only important signals and forces the rest to zero, leading to desired model parsimony.

2.3. Adaptive conditional risks of simultaneous recovery

Based on Lemma 1 we are able to derive a new oracle inequality that relates the componentwise conditional risks of the Stein estimates $\hat{\theta}_{ik}^{RS}$ to those attainable by the oracle strategy $\hat{\theta}_{ik}^{o,a}$. The derivation is given in the Supplementary Material, employing techniques of technical interest for assessing Stein or other shrinkage estimates in similar settings. Denote the componentwise conditional risks of the oracle strategy $\hat{\theta}_{ik}^{o,a}$ by $\mathcal{R}_{i,m}^*(k) \triangleq \mathbf{E}_i(\theta_{ik} - \hat{\theta}_{ik}^{o,a})^2$, where simple calculation yields

$$\mathcal{R}_{i,m}^*(k) = \frac{\lambda_k\sigma^2/m}{\lambda_k + \sigma^2/m} + \frac{\sigma^4/m^2}{\lambda_k + \sigma^2/m}(\theta_{ik}^2 - \lambda_k), \quad (13)$$

where $\sigma^2 = 1$ is written explicitly for generality, and isometry gives $\mathcal{R}^*_{i,m} = \sum_{k=1}^{\infty} \mathcal{R}^*_{i,m}(k)$. Similarly, for any estimator $\hat{\boldsymbol{\theta}}_i$, denoting $\mathcal{R}_{i,m}(\hat{\boldsymbol{\theta}}_i, k) = \mathbf{E}_i(\theta_{ik} - \hat{\theta}_{ik})^2$ gives $\mathcal{R}_{i,m}(\hat{\boldsymbol{\theta}}_i) = \sum_{k=1}^{\infty} \mathcal{R}_{i,m}(\hat{\boldsymbol{\theta}}_i, k)$.

THEOREM 1. *Consider multiple GSMs (4) with the decay assumption (6), for* $i = 1, \ldots, n$ *and* $k = 1, \ldots, m$,

$$\mathbf{E}_i(\theta_{ik} - \hat{\theta}^{RS}_{ik})^2 = \mathcal{R}^*_{i,m}(k) + e_{ik}, \tag{14}$$

where $\mathcal{R}^*_{i,m}(k) = \mathbf{E}_i(\theta_{ik} - \hat{\theta}^{o,a}_{ik})^2$ *are the conditional oracle risks in (13), and the discrepancies are quantified by*

$$e_{ik} \leq \max\left(1, \frac{\theta^2_{ik}}{\lambda_k}\right) \left\{ C_\delta \min(\lambda_k, \delta/m) + C\mathbf{P}_i^{1/2}(A^m_\delta)(\lambda_k + 1/m) \right\}, \tag{15}$$

where $C_\delta = 3(6+\delta)/(1-\delta)$, $C = (\sqrt{24} + 4\sqrt{12})$ *are bounded and* $\mathbf{P}_i(A^m_\delta)$ *satisfy the probability bounds in (12).*

This oracle inequality sets the theoretical basis for evaluating the conditional risks of simultaneously recovering the effects $\{\boldsymbol{\theta}_i\}_{i \leq n}$. Note that the componentwise conditional oracle risks, $\mathcal{R}^*_{i,m}(k)$, differ from the optimal avarage risk by random perturbations. When we sum these risks to obtain $\mathcal{R}^*_{i,m}$, the decay condition on λ_k (6) with concentration argument ensures that cancellations keep $\mathcal{R}^*_{i,m}$ on the same order of the average risk. Owing to space constraints, we provide a condensed version of our theorem quantifying the maximal conditional risks, which shows that the proposed recovery strategy adapts to the average case oracle. An expanded version and its proof are given in the Supplementary Material.

THEOREM 2. *Consider multiple GSMs (4) with the decay assumption (6), and suppose that* $n, m \to \infty$ *with* $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ *for any* $\gamma_2 \geq \gamma_1 > 0$. *Then*

$$\max_{i \leq n} \mathcal{R}^*_{i,m} = \{1 + o_{a.s.}(1)\} \sum_{k=1}^{\infty} \frac{\lambda_k/m}{\lambda_k + 1/m} \propto m^{-2\alpha/(2\alpha+1)}, \;\; i = 1, \ldots, n. \tag{16}$$

Further, the conditional risks $\mathcal{R}_{i,m}(\hat{\boldsymbol{\theta}}^{RS}_i) = \mathbf{E}_i||\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}^{RS}_i||^2_{\ell_2}$ *adapt simultaneously to the oracle risks* $\mathcal{R}^*_{i,m}$ *for any choice of* $\kappa \geq 2$ *with* $\delta = \sqrt{4(\kappa+1)\log m/n} \to 0$,

$$\max_{i \leq n} \mathcal{R}_{i,m}(\hat{\boldsymbol{\theta}}^{RS}_i)/\mathcal{R}^*_{i,m} = 1 + o_{a.s.}(1). \tag{17}$$

It follows from the expanded version of Theorem 2 in the Supplementary Material that, without prior knowledge of the α governing decay of the $\{\lambda_k\}_{k \in \mathbb{N}}$, $\kappa = 2$ is the smallest value such that $\max_{i \leq n} e_i = o_{a.s.}\{m^{-2\alpha/(2\alpha+1)}\}$, which yields an explicit form of the Stein estimates defined in (9).

2.4. Risk comparisons with individual blocking

To appreciate the advantages of information pooling by the proposed method, we compare with the risk performance of standard "pathwise" blocking estimators that use the data

from that individual only, in the context of multiple GSMs, see Tsybakov (2009); Johnstone (2015) for details. To construct a pathwise blocking Stein estimation, given an increasing sequence of numbers in $[m] = \{1, \dots, m\}$, $1 \le j_1 < j_2 < \cdots < j_{K_m} = m$ with $K_m \to \infty$ and setting $j_0 = 0$, we form a partition of $[m]$ into blocks $\mathcal{B} = \{B_1, \dots, B_{K_m}\}$, where $B_k = \{j_{j-1} + 1, \dots, j_k\}$. Denote the cardinality of the kth block by $|B_k|$, set $\boldsymbol{a}_{B_k} = (a_{j_{k-1}+1}, \dots, a_{j_k})^T$ for $\boldsymbol{a} \in \mathbb{R}^m$, one estimates the components of the kth block by Stein shrinkage,

$$\hat{\boldsymbol{\theta}}_{i,B_k} = \left(1 - \frac{|B_k|/m}{||\boldsymbol{Y}_{i,B_k}||_2^2}\right)_+ \boldsymbol{Y}_{i,B_k}, \tag{18}$$

and correspondingly $\boldsymbol{\theta}_i$ by $\hat{\boldsymbol{\theta}}_i^{\mathcal{B}} = (\hat{\boldsymbol{\theta}}_{i,B_1}, \dots, \hat{\boldsymbol{\theta}}_{i,B_{K_m}}, 0, \dots)^T$, with the conditional risks denoted by $\mathcal{R}_i(\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}) \triangleq \mathbf{E}_{\boldsymbol{\theta}_i} ||\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i^{\mathcal{B}}||_{\ell_2}^2$.

THEOREM 3. *Consider multiple GSMs (4) with the decay assumption (6), and suppose that $n, m \to \infty$ with $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ for any $\gamma_2 \ge \gamma_1 > 0$. For any blocking scheme $\mathcal{B}$ satisfying $K_m = o\{m^{1/(\alpha+1/2)}/\log m\}$ and containing at most $O(\log m)$ blocks of size less than $\log m$, the blocking estimator (18) cannot outperform the average case oracle. The lower bound of the conditional risks is given by*

$$\min_{i \le n} \mathcal{R}_i(\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}) \ge \{1 + o_{a.s.}(1)\} \sum_{k=1}^{\infty} \frac{\lambda_k/m}{\lambda_k + 1/m} \propto m^{-2\alpha/(2\alpha+1)}.$$

Further, in this setting, when $K_m = o\{m^{1/(2\alpha+1)}/\log m\}$ and $\alpha \ge 1/2$, there always exist permutations of $(\lambda_1, \dots, \lambda_m)^\top$ under which $\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}$ performs poorly such that $\mathcal{R}_i(\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}) \gg m^{-2\alpha/(2\alpha+1)}$ with high probability for each $i = 1, \dots, n$.

Standard blocking schemes, such as dyadic and weak geometric systems, satisfy conditions in the theorem and have a fundamental limit at performance of the average case oracle. The crucial drawback of these blocking estimators is that they require knowledge of the decay ordering, as there must be some block of size at least $m^{2\alpha/(2\alpha+1)} \log m$. If a permutation places too many large effects in this block, then, with high probability, the realized effects will be larger than $|B|/m \gtrsim m^{-1/(2\alpha+1)} \log m$, resulting in a crude lower bound $m^{-1/(2\alpha+1)} \log m \gg m^{-2\alpha/(2\alpha+1)}$. The restriction $\alpha \ge 1/2$ is not stringent for two reasons. First, from the constructed "bad" permutation in the proof, a smaller number of blocks K_m leads to a slower rate with a crude lower bound $O(1/K_m)$. If K_m is much smaller than stated in the theorem, as for dyadic or weakly geometric systems, the rate can be substantially worse for any $\alpha > 0$. Second, from a smoothness perspective, envisioning $\lambda_{(k)}$ as eigenvalues of a covariance kernel satisfying Sacks-Ylvisacker conditions of order $r = \alpha - 1/2$ (Ritter, 2000), the constraint $\alpha \ge 1/2$ even includes the case of Brownian motion. This provides a rationale of not pooling across indices k, e.g., Cai (1999); Zhang (2005), because performance of these strategies relies on coefficients in a given block being of similar size, which generally needs an implicit assumption on decay. Although pooling over k would alleviate the condition $n \gtrsim m^{\gamma_1}$ in the case of standard decay, $\lambda_k \propto k^{-(2\alpha+1)}$, in the more general setting $\lambda_{(k)} \propto k^{-(2\alpha+1)}$, such strategies can be highly suboptimal. Nevertheless, it remains an open question whether pooling over the experiments $i = 1, \dots, n$ is strictly necessary, which deserves further investigation.

2.5. Robustness guarantee in minimax sense

In order to provide a robustness guarantee for the Stein estimates (9) in a minimax sense, we specify a sequence of parameter spaces which account for the increasing number of experiments we need to control. When we restrict to the first m coefficients of the $\{\boldsymbol{\theta}_i\}_{i\le n}$, it is reasonable to expect these decay on the order of $\sqrt{\lambda_k \log(mn)}$, which suggests the scaling of the spaces. Thus we fix $a, b > 0$ and define

$$A_{mn,k}(\lambda_k) = \big\{x \in \ell_2 : x_k^2/\lambda_k \le a\log(mn)\big\}, B_{mn,k}(\lambda_k) = \big\{x \in \ell_2 : x_k^2/\lambda_k \le b\log(nk)\big\},$$
$$A_{mn}(\boldsymbol{\lambda}) = \cap_{k\le m} A_{mn,k}(\lambda_k), B_{mn}(\boldsymbol{\lambda}) = \cap_{k>m} B_{mn,k}(\lambda_k), \boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda}) = A_{mn}(\boldsymbol{\lambda}) \cap B_{mn}(\boldsymbol{\lambda}).$$

Then with proper choices of a and b, one can guarantee that eventually all the θ_{ik} lie in $\boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})$ and this becomes void if we substantially shrink the space. As a benchmark, we calculate the classical minimax risk,

$$\mathcal{R}_m\{\boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})\} \propto m^{-2\alpha/(2\alpha+1)}\{\log(mn)\}^{1/(2\alpha+1)},$$

which follows from the fact that the $\boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})$ are hyper-rectangles and the bounds for minimax rates over these geometric regions given by Donoho *et al.* (1990). Theorem 2 indicates that our estimation strategy recovers the "within-sample" signals $\boldsymbol{\theta}_i$ simultaneously below the minimax risk over a sequence of parameter spaces which "just" contains them. In the next theorem, we quantify the probabilistic assertion $\{\boldsymbol{\theta}_i\}_{i\le n} \subseteq \boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})$ and provide a robustness guarantee to deviations from the distributional assumption in the following minimax sense. Suppose we are given a fixed $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})$ and noisy observations on $\boldsymbol{\theta}^*$ corresponding to (1),

$$Y_k^* = \theta_k^* + m^{-1/2} z_k^*, \qquad k \in \mathbb{N},$$

independent of $\{Y_{ik}\}_{i\le n}$. Then we may construct an estimate $\hat{\boldsymbol{\theta}}^{*RS}$ using the weights $\alpha_{mn,k}$ (9) calculated from $\{Y_{ik}\}_{i\le n}$ via $\hat{\theta}_k^{*RS} = \alpha_{mn,k} Y_k^*$. The following theorem asserts that the largest risk we incur by this procedure comes within a logarithmic factor of the minimax risk over $\boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})$.

THEOREM 4. *Consider multiple GSMs (4) with the decay assumption (6), and suppose that $n, m \to \infty$ with $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ for any $\gamma_2 \ge \gamma_1 > 0$.*

(i) It holds that $\max_{i\le n} ||\boldsymbol{\theta}_i^2/\boldsymbol{\lambda}||_{m,\infty} = \{1 + o_{a.s.}(1)\}2\log(nm)$, thus, if $a < 2$, eventually some $\boldsymbol{\theta}_i \notin \boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})$. Further, For any $a > (\gamma_1+2)/(\gamma_1+1)$ and $b > (2\gamma_1+3)/(\gamma_1+1)$, we have $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \notin \boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})$ only finitely often.

(ii) Let $\hat{\boldsymbol{\theta}}^{*RS}$ denote the procedure outlined above, and $\mathcal{R}_m(\hat{\boldsymbol{\theta}}^{*RS}) = \mathbf{E}_{\boldsymbol{\theta}^*}\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}^{*RS}\|_{\ell^2}^2$ for any $\boldsymbol{\theta}^*$. Then we have

$$\sup_{\boldsymbol{\theta}^* \in \boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})} \mathcal{R}_m(\hat{\boldsymbol{\theta}}^{*RS}) \propto \{\log(mn)\}^{2\alpha/(2\alpha+1)} \mathcal{R}_m\{\boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})\}.$$

By coupling the arguments in the proofs of Theorem 2 and Theorem 4 in the Supplementary Material, we also see that, if $\{\boldsymbol{\theta}_{n+1}, \ldots, \boldsymbol{\theta}_{n+N}\}$ are independent draws from the same GSM (4) with $m^{\gamma_1} \lesssim N \lesssim m^{\gamma_2}$, owing to information pooling over $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$, we have almost surely,

$$\max_{j\le N} \mathbf{E}_{\boldsymbol{\theta}_{n+j}} \|\boldsymbol{\theta}_{n+j} - \hat{\boldsymbol{\theta}}_{n+j}^{RS}\|_{\ell^2}^2 \propto m^{-2\alpha/(2\alpha+1)} = o[\mathcal{R}_m\{\boldsymbol{\Theta}_{mn}(\boldsymbol{\lambda})\}],$$

where $\hat{\boldsymbol{\theta}}^{RS}_{n+j}$ are obtained by applying $\alpha_{mn,k}$ (9) calculated from $\{Y_{ik}\}_{i\leq n}$ to the noisy sequence $\{Y_{n+j,k}\}_{k\leq m}$ via $\hat{\theta}^{RS}_{n+j,k} = \alpha_{mn,k}Y_{n+j,k}$. Moreover, part (i) in Theorem 4 applies to various weakly dependent variables (e.g. Pickands, 1969) and part (ii) holds regardless of dependence structures, which can be seen from its proof.

2.6. The case of unknown variance

We now consider the case of unknown variance σ^2 in multiple GSMs,

$$Y_{ik} = \theta_{ik} + \sigma m^{-1/2} z_{ik}, \;\; i = 1, \ldots, n, \;\; k \in \mathbb{N},$$

with the goal of maintaining the risk properties of the proposed Stein estimates while using a data-based estimator $\hat{\sigma}^2$. Variance estimation has been extensively studied in nonparametric regression (2), mostly based on localized squared residuals (Hall and Carroll, 1989; Hall and Marron, 1990; Fan and Yao, 1998) or difference sequences (Müller and Stadtmüller, 1987; Hall *et al.*, 1990; Brown and Levine, 2007), among many others. It is rarely discussed in the GSM (1) which mainly serves as a theoretical device to study nonparametric regression problems. A relevant case is the robust estimation using median of the finest scale coefficients in a wavelet-transformed model (Donoho and Johnstone, 1994). In the multiple GSMs model (4), we propose a natural means of estimating σ^2, based on concentration of measure.

Derivation of the key oracle inequality of Theorem 1 relies on the sets $A^c_{k,\delta}$ containing most of the probability mass. In the case of unknown variance, the definitions extend to

$$A^c_{k,\delta} = \{(1-\delta)(\lambda_k + \sigma^2/m) \leq ||Y_{\cdot k}||^2/n \leq (1+\delta)(\lambda_k + \sigma^2/m)\},$$

and $A^{m,c}_\delta = \bigcap_{k\leq m} A^c_{k,\delta}$. Conditional concentration of measure guarantees that the probability bounds in Lemma 1 continue to hold, thus these sets capture "most" realizations. Algebra shows that for realizations in this range, an amendment to the Stein weights (9) guarantees they remain close to the linear oracle weights. Specifically, let $Q^m_p(\cdot)$, $p \in (0,1)$, denote the quantile function retrieving the element in a vector $x \in \mathbb{R}^m$ that is greater than or equal to pm elements. Then on $A^{m,c}_\delta$, denoting $\boldsymbol{\lambda}_m = (\lambda_1, \ldots, \lambda_m)^\top$ and $\|\boldsymbol{Y}_m\|^2 = (\|Y_{\cdot 1}\|^2, \ldots, \|Y_{\cdot m}\|^2)^\top$,

$$(1-\delta)\{Q^m_p(\boldsymbol{\lambda}_m) + \sigma^2/m\} \leq Q^m_p(||\boldsymbol{Y}_m||^2/n) \leq (1+\delta)\{Q^m_p(\boldsymbol{\lambda}_m) + \sigma^2/m\}.$$

If p is fixed, by the decay assumption (6), we have $Q^m_p(\boldsymbol{\lambda}_m) \asymp \{(1-p)m\}^{-(2\alpha+1)}$, i.e., $mQ^m_p(\boldsymbol{\lambda}_m) \propto m^{-2\alpha}$. Hence, on A^c_δ for some $0 < c \leq C$,

$$(1-\delta)(\sigma^2 + cm^{-2\alpha}) \leq mQ^m_p(||\boldsymbol{Y}_m||^2/n) \leq (1+\delta)(\sigma^2 + Cm^{-2\alpha}).$$

This motivates an estimator of σ^2 with small relative error on sets of high probability,

$$\hat{\sigma}^2_p = mQ^m_p(||\boldsymbol{Y}_m||^2/n), \tag{19}$$

Further, we show that this holds for any p satisfying $\{(1-p)m\}^{-(2\alpha+1)} = o(m^{-1})$. Thus we may let p vary to reveal an interesting phenomenon stated in the following theorem, with the proof deferred to the Supplementary Material. To state the theorem more precisely, for $p \in (0,1)$, we take $\hat{\sigma}^2_p = mQ^m_p(||\boldsymbol{Y}_m||^2/n)$ and amend the Stein weights in (9),

$$\alpha_{mn,k}(\hat{\sigma}^2_p) = \left[1 - \left\{1 + (q_\delta + 2)\sqrt{12\log m/n}\right\} \frac{n\hat{\sigma}^2_p/m}{||Y_{\cdot k}||^2}\right]_+, \tag{20}$$

where

$$q_\delta = (1+2\delta)/(1-\delta) \gtrsim 1, \quad \delta = \sqrt{12\log m/n}, \tag{21}$$

and form the corresponding estimates $\hat{\theta}_{ik}^{ST}(\hat{\sigma}_p^2) = \alpha_{mn,k}(\hat{\sigma}_p^2)Y_{ik}$ for $k \le m$ and 0 otherwise.

THEOREM 5. *Consider multiple GSMs (4) with decay assumption (6) and error variance σ^2 unknown and suppose that $n, m \to \infty$ with $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ for any $\gamma_2 \ge \gamma_1 > 0$. Then there exists a sequence $p_m^* \to 1$ as $m \to \infty$, such that the conditional oracle inequality of Theorem 1 continues to hold, with slight modification, for the estimates $\hat{\theta}_{ik}^{ST}(\hat{\sigma}_p^2)$ for all $p \le p_m^*$. Consequently, any choice of $p \le p_m^*$ leads to a consistent estimator which enjoys the optimal rate of Theorem 2 and robustness guarantees of Theorem 4.*

For the threshold sequence $p_m^* \to 1$, below which we always have consistency and above which the oracle inequality breaks down, the intuition is that if p is too large, our estimate of σ^2 becomes confounded by signal. From a practical perspective, it is not necessary to precisely quantify the threshold value p_m^*. A safe choice is to adopt the smallest order statistic

$$\hat{\sigma}_{\min}^2 = (m\|\boldsymbol{Y}_m\|^2/n)_{(m)}. \tag{22}$$

Larger choices of p yield more parsimonious recoveries with the potential price of breaking consistency. See Section 4.1 for a demonstration of an "elbow" type risk transition associated with model complexity.

3. Connection to Functional Data and Recovery in General Basis

Recall the functional data model (5) in additive Gaussian noise,

$$y_{ij} = f_i(x_{ij}) + \sigma z_{ij}, \;\; x_{ij} \in [0,1], \;\; i = 1,\dots,n, \; j = 1,\dots,m,$$

where $z_{ik} \overset{i.i.d.}{\sim} N(0,1)$. Here, without loss of generality, we assume the observation points are either fixed, $x_{ij} = j/m$, or random, $x_{ij} \overset{i.i.d.}{\sim} U(0,1)$, which covers the main cases of interest in sampling design for dense functional data. We show in this section, via Le Cam asymptotic equivalence, that conditional risk problems concerning the underlying functions f_i, for both fixed and random designs, can be directly related to their corresponding problems in the following white noise models,

$$Y_i(dt) = f_i(t)dt + \sigma m^{-1/2} B_i(dt), \qquad i = 1,\dots,n, \quad t \in [0,1]. \tag{23}$$

In the case of L^2 recovery, these are equivalent to the multiple GSMs (4) when projected onto the Karhunen-Loève (KL) basis (i.e., eigenfunctions) of the covariance function $C = \mathbf{E} f \otimes f$. Noting that the variances of the effects, λ_k, are nondeceasingly ordered eigenvalues in this setting, the decay condition $\lambda_k \propto k^{-(2\alpha+1)}$ in (6) is satisfied under fairly general assumptions. For instance, it is sufficient that C satisfies the Sacks-Ylvisacker conditions of order $r = \alpha - 1/2 \ge 0$ (Ritter *et al.*, 1995).

We see that, to apply the proposed recovery strategy to functional data, an obstacle is the unknown KL basis. Although this basis may be estimated from data, it is not our

purpose to employ traditional functional principal component procedures, which are computationally expensive and introduce data-dependent uncertainty. Instead, we further extend the proposed Stein estimates and their theoretical guarantees to more general settings where the effects θ_{ik} may be correlated across k. This allows projection of the white noise models (23) onto other bases, such as wavelet or Fourier, provided that the correlation among projected coefficients satisfies some mild conditions.

Besides theoretical advantages, a major benefit of recovering functional data by GSMs is potentially significant computational savings. Typical pre-smoothing (Ramsay and Silverman, 2005) or post-smoothing (Yao *et al.*, 2005) of individual functions for estimation of covariance/eigen-structure operates with $O(nm^2 + m^3)$ complexity for common design, which scales poorly with data, and $O(n^3m^3)$ complexity for random design, which is intractable for large datasets. A recent proposal by Xiao *et al.* (2016) dealing with covariance estimation for functional data in common design settings using penalized splines operates at the order of $O(nm^{1+\rho})$, where $0 < \rho \leq 1$ depends on smoothness and can be lifted to $\rho = 1$ to capture striking local features (Donoho and Johnstone, 1994), such as in situations illustrated in the top right panel of Figure 1. In contrast, by our method one can take advantage of fast transforms, such as fast wavelet or Fourier transform, to obtain recoveries in $O(nm \log m)$ time with spatial adaptation. A simple calculation indicates that our recovery strategy retains at most

$$d_* \lesssim \left(m\sqrt{n/\log m}\right)^{1/(2\alpha+1)}$$

nonzero weights with high probability. One may in turn run principal components analysis on these at $O(nd_*^2 + d_*^3)$ cost to attain estimates of the covariance/eigen structure. The entire procedure may then potentially scale at $O(nm \log m)$. Further, our method may be employed in an online algorithm fashion: a new curve comes in, transform is performed, threshold weights updated. This could potentially extend the FDA applications to realtime data collection and processing.

3.1. Review of Le Cam equivalence for nonparametric experiments

We begin with the notion of Le Cam equivalence used in Brown and Low (1996); Brown *et al.* (2002); Reiß (2008), following an amalgam of notations from these sources. Denote two sequences of experiments by standard probability triples, indexed by an identical parameter space Θ_m, changing with m (or n), by

$$\mathbb{E}_m = \{(\mathcal{X}_1^m, \mathcal{B}_1^m, P_{1,\theta}^m), \theta \in \Theta_m\}, \qquad \mathbb{G}_m = \{(\mathcal{X}_2^m, \mathcal{B}_2^m, P_{2,\theta}^m), \theta \in \Theta_m\}.$$

Given randomized decision rules, δ^i, $i = 1, 2$, on a common action space $\mathcal{A}$ and loss functions $L = L_m : \Theta_m \times \mathcal{A} \to [0, \infty)$, the corresponding risks, $R^i(\delta^i, L, \theta) = \mathbf{E}_{P_{i,\theta}^m} L(\delta^i, \theta)$, take the form

$$R^i(\delta^i, L, \theta) = \int_{\mathcal{X}_i^m} \int_{\mathcal{A}} L(\theta, a)\delta^i(da|y)\mathbf{P}_\theta^i(dy), \quad i = 1, 2.$$

With the pseudo norm $||L|| = \sup\{L(\theta, a) : \theta \in \Theta_m, a \in \mathcal{A}\}$, the experiments $\mathbb{E}_m$ and $\mathbb{G}_m$ are considered asymptotically equivalent in the Le Cam sense as $m \to \infty$, if

$$\Delta(\mathbb{E}_m, \mathbb{G}_m) := \max\left[\inf_{\delta^1}\sup_{\delta^2}\sup_{\theta\in\Theta_m, ||L||=1}|R^1(\delta^1, L, \theta) - R^2(\delta^2, L, \theta)|,\right.$$
$$\left.\inf_{\delta^2}\sup_{\delta^1}\sup_{\theta\in\Theta_m, ||L||=1}|R^1(\delta^1, L, \theta) - R^2(\delta^2, L, \theta)|\right] \longrightarrow 0. \quad (24)$$

To unify results on equivalence for fixed and random designs of nonparametric experiments, we need to introduce some background and notations on piecewise constant approximations. For a given positive integer k, let $I_{k,j} = [j/k, (j+1)/k)$ for $j = 0, \ldots, k-2$ and $I_{k,k-1} = [1 - 1/k, 1]$ so that the $I_{k,j}$ form a partition of $[0, 1]$. Let $\phi_{k,j} = k^{1/2}\mathbf{1}_{I_{k,j}}$, such that for a given k these functions form an orthonormal basis for the subspace S_k of $L^2[0, 1]$ consisting of functions constant on each of the $I_{k,j}$. The functions $H_{l,j} = 2^{-l/2}(\phi_{2^{l+1},2j} - \phi_{2^{l+1},2j+1})$, $l \geq 1$ and $k = 0, \ldots, 2^l - 1$, together with $H_{0,0} = \mathbf{1}_{[0,1]}$, form the orthonormal Haar wavelet basis of $L^2[0, 1]$ and $w_{l,j}(f) = \langle f, H_{l,j}\rangle$ the Haar wavelet coefficients for a function $f \in L^2[0, 1]$. We introduce a class of norms defined on the Haar wavelet coefficients, for a given $\alpha > 1/2$,

$$||f||_{(\alpha)} = \left\{\sum_{k=0}^{\infty} 2^{2k\alpha}\sum_{l=0}^{2^k-1} w_{k,l}^2(f)\right\}^{1/2},$$

which are closely related to a specific instance of Besov norms and provide a generalization of various types of smoothness, e.g. Hölder continuity, Sobolev smoothness. With these norms in place, we define the parameter spaces of interest. For some $B_n \to \infty$, we take

$$\Theta_m = \{f \in L^2[0, 1] : ||f||_{(\alpha)} < B_m\}.$$

Now let $\mathbb{F}_m$ and $\mathbb{R}_m$ denote the nonparametric experiments in (2) for fixed and random designs, respectively, and let $\mathbb{W}_m$ denote the white noise model in (3), as f ranges over Θ_m. The lemma below provides a generalization of existing results that may be applied to functional data models, see Supplementary Material for its proof.

LEMMA 2. *For fixed and random designs of the nonparametric experiments in (2), we have a unified bound on the Le Cam distance*

$$\max\{\Delta(\mathbb{F}_m, \mathbb{W}_m), \Delta(\mathbb{R}_m, \mathbb{W}_m)\} \lesssim B_m^2 m^{-\left(\frac{2\alpha-1}{2\alpha+1}\right)}.$$

Consequently, if $B_m = o\left(m^{(2\alpha-1)/(4\alpha+2)}\right)$, *we have the asymptotic equivalence in Le Cam sense between the nonparametric model (2) under both designs and the white noise model (3).*

3.2. Functional data and white noise representations

Given a single function f as a realization of a sufficiently smooth stochastic process, we may form the nonparametric regression experiment $y_i = f(x_i) + z_i$ and corresponding white noise model $Y(dt) = f(t)dt + m^{-1/2}W(dt)$, $i = 1, \ldots, m$. It is not obvious that we may

approximate the random variables $\mathbf{E}_f L(f,\cdot)$ from one model by those from the other as $m \to \infty$. Modulo proper assumptions on the smoothness of f, this is indeed possible and can be shown using Le Cam equivalence theory for fixed functions, e.g., Lemma 2 from above. Intuitively, this works because for reasonable stochastic processes, the smoothness of f measured in $||\cdot||_{(\alpha)}$ has null probability of growing fast enough as $m \to \infty$ to dominate the convergence rate of Le Cam equivalence of the experiments over Θ_m. By extension, it is not obvious that, given n nonparametric regression models sampled at m points generated from the functional data model (5), the risks $\mathbf{E}_{f_i} L(f_i,\cdot)$, $i = 1,\ldots,n$, from the nonparametric regression models are approaching those from the corresponding white noise models (23).

Recall that n, m are asymptotically linked through the constraint $m^{\gamma_1} \lesssim n \lesssim n^{\gamma_2}$. The main appeal of Lemma 2 is that proving bounds on wavelet coefficients $|w_{l,j}(f)|$ under various smoothness constraints commonly used in practice is a relatively simple task, while in FDA smoothness constraints are typically imposed through moment conditions on a norm $||f||^*$ which often dominates $||f||_{(\alpha)}$ for some $\alpha > 1/2$. If the growth of the norms $\max_{i\leq n} ||f_i||^*$ can be controlled a.s. by a B_m satisfying the conditions of Lemma 2, we may construct a Θ_m over which we have Le Cam equivalence between the experiments of the previous section while ensuring that eventually $f_1,\ldots,f_n \in \Theta_m$. This means that for any estimator δ^1 in the nonparametric regression experiment there is an estimator δ^2 in the white noise model, and visa versa, so that eventually

$$\begin{aligned}\max_{1\leq j\leq n} \sup_{||L||\leq 1} |R^1(\delta^1, L, f_j) - R^2(\delta^2, L, f_j)| &\leq \sup_{f\in\Theta_m} \sup_{||L||\leq 1} |R^1(\delta^1, L, f) - R^2(\delta^2, L, f)| \\ &\lesssim B_m^2 m^{-(2\alpha-1)/(2\alpha+1)} = o(1).\end{aligned}$$

As this allows for randomized estimators, we are guaranteed that we can model functional data (5), under risks $\mathbf{E}_{f_i} L(f_i,\cdot)$, by the corresponding risks under white noise models (23), and visa versa. This follows on noting that $\mathbf{E}_{f_i}(\cdot) = \mathbf{E}(\cdot|f_i)$ is the expectation that averages over all data in $j \neq i$ experiments and the noise in the ith experiment, and thus any estimator which pools over $j = 1,\ldots,n$ may be viewed as a randomized estimator with respect to $\mathbf{E}_{f_i}$. We first provide a useful lemma and some concrete examples then summarize in a general theorem.

LEMMA 3. *Suppose that $f \in L^2[0,1]$ has a generalized derivative, $f' \in L^2[0,1]$. Then $w_{0,0}^2(f) \leq ||f||_2^2$ and for $k \geq 1$, $l = 0,\ldots,2^k-1$, the Haar wavelet coefficients obey the decay*

$$w_{k,l}^2(f) \leq 2^{-2k} \int_{I_{k,l}} |f'(s)|^2 ds.$$

Consequently, $||f||_{(\alpha)} < \infty$ for any $\alpha \in (1/2, 1)$ and may be bounded by a factor of $||f||_2 + ||f'||_2$ which may, in turn, be bounded by a factor of $|f(0)| + ||f'||_2$.

Based on this lemma and the discussion above, if the growth of the norms $\max_{i\leq n}(||f_i|| + ||f_i'||_2)$ can be controlled a.s. by $m^{-(2\alpha-1)/(2\alpha+1)}$, we may construct a Θ_m over which we have Le Cam equivalence between the experiments while ensuring eventually $f_1,\ldots,f_n \in \Theta_m$.

EXAMPLE 3.1. A general method for forming processes is to smooth white noise by integrating against a kernel. Regularity assumptions on the underlying kernel result in regularity of the sample paths. As a concrete example, suppose that all second order partial derivatives of $R : [0,1]^2 \to \mathbb{R}$ exist and are of bounded variation with either argument taken as fixed. Define the process f on $[0,1]$ by $f(s) = \int_0^1 R(s,t)B(dt)$, where B is a Brownian

motion. First notice that $f(0)$ is gaussian with mean 0 and, by Ito Isometry, variance $\int_0^1 R^2(0,t)dt$ which gives $\max_{i\leq n}|f_i(0)| \lesssim (\log m)^{1/2}$ a.s. under the assumptions on n, m. Further, the assumptions on R allow us to integrate by parts, $f'(s) = B(1)\partial_s R(s,1) - \int_0^1 B(t)\partial_s R(s,dt)$. Denoting the variation of a function g by $V(g,[0,1])$ gives

$$||f'||_\infty \leq \sup_{0\leq s\leq 1}(|\partial_s R(s,1)| + V(\partial_s R(s,\cdot),[0,1])) \sup_{0\leq t\leq 1} B(t).$$

If the first supremum term is bounded, the reflection principle puts sub-Gaussian tails on $||f'||_\infty$. Hence for a sample $f_1,\ldots,f_n \overset{i.i.d.}{\sim} f$, we have $\max_{i\leq n}||f_i'||_\infty \lesssim (\log m)^{1/2}$, a.s. Since $m^{-(2\alpha-1)/(2\alpha+1)}\log m = o(1)$ for any choice of $\alpha \in (1/2,1)$, Lemma 2 gives Le Cam equivalence over $\Theta_m = \Theta_m(\alpha)$ with $B_m \lesssim (\log m)^{1/2}$. Lemma 3 implies $||f||_{(\alpha)} \lesssim |f(0)| + ||f'||_2$ which is bounded by $|f(0)| + ||f'||_\infty$. This guarantees that, in view of $\max_{i\leq n}(|f_i(0)| + ||f_i'||_\infty) \lesssim (\log m)^{1/2}$ a.s., we eventually have $f_1,\ldots,f_n \in \Theta_m$. Thus we may model the risks of recovering the f_i from functional data by white noise models. Sufficiently regular convolutions fall under this model which generalizes to processes formed by taking linear combinations of this form. □

Example 3.2. Slightly more general assumptions in FDA literature include, for all $C > 0$,

$$\max_{j=0,1,2}\mathbf{E}||f^{(j)}||_\infty^C < \infty,$$

as in Hall *et al.* (2006). A weakening of this requirement which might be seen as a strengthening of the condition in Cai and Yuan (2011) requires, for all $C > 0$,

$$\max_{j=0,1,2}\mathbf{E}||f^{(j)}||_2^C < \infty.$$

Under these conditions, for any choice of $\alpha \in (1/2,1)$ and a sample $f_1,\ldots,f_n \overset{i.i.d.}{\sim} f$, we have

$$\max_{i\leq n}(||f_i||_2 + ||f_i'||_2) = o_{a.s.}(m^{-(2\alpha-1)/(2\alpha+1)}).$$

Thus, as in the previous example, we may model the risks of recovering the f_i from functional data by white noise models. □

These examples indicate a general approach to establishing white noise equivalence for functional data, as stated in the following theorem with proof in the Supplementary Material, which includes the commonly adopted settings in FDA literature as special cases. Recall the general setting $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ for any $\gamma_2 \geq \gamma_1 > 0$, and the decay parameter $\alpha > 0$ in (6).

Theorem 6. *Suppose that for a norm $||\cdot||^*$ which dominates $||\cdot||_{(\alpha)}$ for some $\alpha \in (1/2,1)$, as in the examples above, the process f satisfies*

$$\mathbf{P}(||f||^* > x) \lesssim (1+x)^{-\beta},$$

for some $\beta > 2(\alpha+1)\max(1,\gamma_2)/(2\alpha-1)$. Then we can model the recovery of functional data in both fixed and random designs, by white noise models (23).

3.3. Recovery of functional data in general basis and extension to general decay

A remaining issue of applying the recovery strategy to functional data is that the multiple GSMs (4) assume independence among θ_{ik} across k, which corresponds to projecting white noise models (23) onto the unknown KL basis of $C = \mathbf{E}f \otimes f$. We elaborate that, for projections onto other basis with the coefficients displaying to some extent decaying correlations, the recovery results continue to hold modulo constants and yield estimators adaptive to the average case oracle strategies.

Now the coefficients θ_{ik} are projections of f_i onto a given basis ψ_k, $\theta_{ik} = \langle f_i, \phi_k \rangle$. Given the independence across curves, the oracle inequality of Theorem 1 continues to hold and the main difficulty lies in generalizing the results of Theorem 2 to deal with dependence among the projected coefficients of the underlying process. With $\lambda_k = \mathbf{Var}(\theta_{ik})$, set

$$\Delta_m = \mathrm{diag}(\lambda_1, \ldots, \lambda_m), \quad \Sigma_m = (\mathbf{Cov}(\theta_{ij}, \theta_{ik}))_{j,k \le m}, \quad \Gamma_m = \Delta_m^{-1/2} \Sigma_m \Delta_m^{-1/2}. \tag{25}$$

The matrix Γ_m provides term-wise correlations between projected coefficients in the chosen basis, $\{\psi_k\}_{k=1}^{\infty}$ and reduces to identity for KL basis.

(A) The following conditions, collectively called Conditions (A), are sufficient to guarantee that our procedure recovers $\boldsymbol{\theta}_i$ at the optimal rate.

(A.1) The ordered variances $\lambda_k = \mathbf{Var}(\theta_{ik})$ decay as $\lambda_{(k)} \propto k^{-(2\alpha+1)}$.

(A.2) The correlations between θ_{ik} decay sufficiently fast that

$$\max_{i \le m} \sum_{j=1}^{m} |\Gamma_{m,ij}| \le B_m. \tag{26}$$

with $B_m^2 = O\{m^{1/(2\alpha+1)}/(\log m)^p\}$ for some $p \ge 2$.

REMARK 3.1. Regarding (A.1), it is known that covariance functions satisfying the Sacks-Ylvisacker conditions of order $r = \alpha - 1/2$ generate reproducing kernel Hilbert spaces lying within a polynomial translation of the Sobolev space $H^{r+1}([0,1])$. There are many comparable smoothness classes which share similar decay when expressed in efficient bases. Thus it is reasonable to expect that in such bases $\lambda_{(k)}$ decay at the Karhunen-Loève rate, which does not know the ordering *a priori*. Condition (A.2) is fairly unrestrictive, only requiring that the average correlation with any given coefficient decay at sufficient polynomial order as the space increases.

THEOREM 7. *Consider multiple GSMs (4) with Conditions (A) hold, and suppose that* $n, m \to \infty$ *with* $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ *for any* $\gamma_2 \ge \gamma_1 > 0$. *Then the results of Theorem 2 continue to hold and the procedure is adaptive to the average case oracle.*

Next we exhibit some process/basis pairs for which Conditions (A) are satisfied, suggesting the applicability of our recovery method when a suitable basis is chosen for the underlying process.

EXAMPLE 3.3. STATIONARY PROCESS AND PERIODIZED MEYER-TYPE WAVELET BASIS Suppose that $K : \mathbb{R} \to \mathbb{R}$ is a kernel function on $\mathbb{R}$ satisfying sufficient regularity conditions and set $k(s) = \sum_{m \in \mathbb{Z}} K(s - m)$ to be the periodization of K. In particular, assume that $K \in H^{2\alpha+2}(\mathbb{R})$ satisfies the decay $K(x) \lesssim (1 + |x|)^{-l}$ for $l > 1$ as $|x| \to \infty$. Now let f be

the stationary Gaussian process on $[0,1]$ with covariance $C(s,t) = k(t-s)$ and consider the coefficients, $\theta_{jk} = \langle f, \psi_{jk} \rangle$, in a periodized Meyer-type wavelet basis as in Walter and Shen (2000). The proposition below asserts that θ_{jk} are close enough to independence to guarantee the recovery. Denote the "warped" distance at scale j by $|p-q|_j = \inf_{n\in\mathbb{Z}} |p - q + n2^j|$.

PROPOSITION 1. *In the setting outlined above, the wavelet coefficients θ_{jk} and $\theta_{j'k'}$ (i) are uncorrelated for $|j-j'|>1$; (ii) have correlation $O\{(1+|k-2^{j-j'}k'|_j)^{-l}\}$ for $|j-j'|\le 1$. Consequently, for fixed p,q and some $B>0$,*

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \frac{\mathbf{Cov}(\theta_{jk},\theta_{pq})}{\sqrt{\mathbf{Var}(\theta_{jk})\mathbf{Var}(\theta_{pq})}} \le B.$$

EXAMPLE 3.4. SOBOLEV REPRODUCING KERNEL AND SMOOTH WAVELET BASIS
Let K_r, $r\in\mathbb{N}$, be the reproducing kernel of order r, given by $K_r(x,y) = A(x,y)+B(x,y)$,

$$A(x,y) = \sum_{p=0}^{r} \frac{x^p y^p}{(p!)^2}, \qquad B(x,y) = \int_0^{\min(x,y)} \frac{(x-u)^r (y-u)^r}{(r!)^2} du,$$

which is a polynomial of order $(2r+1)$ in x and y. This is the canonical example of covariance structure for a process satisfying Sacks-Ylvisacker conditions of order r. Let f be a mean zero Gaussian process with the covariance kernel K_r, and $\{\psi_{jk}\}_{j,k}$ be a compactly supported wavelet basis orthogonal to polynomials of degree $(2r+1)$. The proposition below indicates that the recovery holds for such processes with $B_m \lesssim \log m$, noting the scales $p \lesssim \log m$ in wavelet bases.

PROPOSITION 2. *In the setting outlined above, the covariance structure of the wavelet coefficients θ_{jk} satisfies*

$$\frac{\mathbf{Cov}(\theta_{jk},\theta_{j'k'})}{2^{-(r+1)(k+k')}} = \begin{cases} 0, & \mathbf{Supp}(\psi_{jk}) \cap \mathbf{Supp}(\psi_{j'k'}) = \emptyset \\ O(2^{-(r+1/2)|k-k'|}), & \textit{otherwise} \end{cases}.$$

Consequently, for fixed p,q and some $B>0$,

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \frac{\mathbf{Cov}(\theta_{jk},\theta_{pq})}{\sqrt{\mathbf{Var}(\theta_{jk})\mathbf{Var}(\theta_{pq})}} \le B\max(1,p).$$

REMARK 3.2. We expect that the above examples hold in greater generality. For instance, in the stationary example, given the local nature of ψ_{jk} for general wavelet bases of $L^2[0,1]$, it is reasonable to expect that similar decay conditions are satisfied if f is taken as a "snapshot" over $[0,1]$ of a process over $\mathbb{R}$ corresponding to K of a locally stationary process as in Mallat *et al.* (1998). Similarly, the derivation for Sobolev-type kernels extends to $A(x,y)$ with reasonable coefficients, and $B(x,y)$ may be expanded to include integrals against more general functions of u.

EXAMPLE 3.5. CALDERON-ZYGMUND AND PSUEUDO-DIFFERENTIAL TYPE INTEGRAL OPERATORS AND SUFFICIENTLY REGULAR WAVELET BASIS

More generally, in sufficiently regular wavelet bases $\{\psi_{jk}\}_{j,k}$, integral operators C such as Calderon-Zygmund and psueudo-differential type, which correspond to broad classes of covariance structures, are known to satisfy decay conditions of the form

$$|\langle \psi_{jk}, C\psi_{j'k'}\rangle| \lesssim \frac{2^{-(r+1)(j+j')}2^{-\kappa|j-j'|}}{(1+d((j,k),(j',k')))^{\gamma}},$$

where $d((j,k),(j',k')) = 2^{\min(j,j')}|k2^{-j} - k'2^{-j'}|$ provides a measure of the distance between supports of the wavelets. See e.g. Cohen (2003) for results in this direction. If such a C, $\kappa > 1$ and $\gamma \geq 1$, corresponds to the covariance of operator of a mean zero Gaussian process f on $[0,1]$ with $\langle \psi_{jk}, C\psi_{j'k'}\rangle \propto 2^{-(r+1)(j+j')}$, then we can show the assertion below regarding the covariance structure of the wavelet coefficients θ_{jk}, and thus $B_m \lesssim \log m$ with $p \lesssim \log m$ in wavelet bases.

PROPOSITION 3. *In the setting outlined above, the covariance structure of the wavelet coefficients θ_{jk} satisfies, for fixed p and q,*

$$\sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1} \frac{\mathbf{Cov}(\theta_{jk},\theta_{pq})}{\sqrt{\mathbf{Var}(\theta_{jk})\mathbf{Var}(\theta_{pq})}} \lesssim p.$$

Further, if $\gamma > 1$ this bound is constant and independent of m.

We conclude this section by pointing out that the results of this paper can be extended to more general decay settings. Denote by $\mathcal{R}$ the average oracle risk, $\mathcal{R} = \mathcal{R}(m) = \sum_{k=1}^{\infty}(\lambda_k/m)(\lambda_k + 1/m)$. The following conditions, collectively called Conditions (B), are needed for the proposed method continue to hold.

(B.1) The signal in the components larger than noise, $\lambda_k > 1/m$, dwarfs the operator norm of the correlations, i.e., $(m\mathcal{R})/||\Gamma||^2 \gtrsim (\log m)^2$.

(B.2) The bulk of signal components are larger than noise, $\lambda_k > 1/m$, remain in the first m components; i.e., $\sum_{k>m}\lambda_k = o(\mathcal{R}/\log m)$.

(B.3) The risk is slowly varying in m so that $\mathcal{R}(m/\delta) \lesssim \delta^{\gamma}\mathcal{R}(m)$ for some $\gamma \in (0,1)$.

It is easy to verify that a more general decay $k^{-(2\kappa+1)} \lesssim \lambda_{(k)} \lesssim k^{-(2\alpha+1)}$ with λ_k slowly varying for $k \leq m$ and $k^{-(2\kappa+1)} \lesssim \lambda_k \lesssim k^{-(2\alpha+1)}$ for $k > m$, where $0 < \alpha < \kappa$, satisfies these conditions.

THEOREM 8. *Consider multiple GSMs (4) with Conditions (B) hold, and suppose that $n, m \to \infty$ with $m^{\gamma_1} \lesssim n \lesssim m^{\gamma_2}$ for any $\gamma_2 \geq \gamma_1 > 0$. Then the proposed method continues to adapt to the average case oracle, i.e.,* $\max_{i\leq n}\mathbf{E}_{\boldsymbol{\theta}_i}||\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i||_{\ell_2}^2/\mathcal{R} = 1 + o_{a.s.}(1)$.

4. Simulated and Real Data Examples

In this section, we first report simulation experiments which highlight advantages of information pooling in multiple GSMs. The performance of our recovery procedure is compared against the linear oracle as well as individual-blocking and soft-thresholding estimators. Robustness to deviation from distributional assumptions and model complexity associated with estimation of σ^2 are also examined, with results supporting our theoretical findings. We then apply the proposed method to the Phoneme dataset studied in Hastie *et al.* (1995). Classification performance and computational times of our method are compared against those attained by common FDA methods based on pre/post-smoothing individual functions.

4.1. Simulation studies

The simulated data are generated from the multiple GSMs with the first m effects,

$$Y_{ik} = \theta_{ik} + m^{-1/2} z_{ik}, \qquad k = 1, \ldots, m; \quad i = 1, \ldots, n, \tag{27}$$

where $z_{ik} \overset{i.i.d.}{\sim} N(0,1)$, $\theta_{ik} \sim N(0, \lambda_j) \perp z_{ik}$, and two scenarios are considered. The first scenario follows the model where variances λ_k are decreasing in k with Sobolev type decay $\lambda_k = \lambda_{(k)} = 2\alpha k^{-(2\alpha+1)}$. In the second scenario, we permute $\{\lambda_k\}_{k\le m}$ uniformly at random, then generate data from the model according to the permuted sequence. In each scenario, 1000 Monte Carlo runs are performed for all combinations of sample sizes $n = 10, 100, 1000$, sampling rates $m = 50, 500, 5000$, and decay speeds $\alpha = 2/3$ and 1. The benchmark is the linear oracle $\hat{\theta}^{o,a}_{ik} = \lambda_k Y_{ik}/(\lambda_k + \sigma^2/m)$. For the proposed method, we calculate the Stein weights using both true and estimated variances, i.e., $\alpha_{mn,k}$ in (9) with the true value $\sigma^2 = 1$; $\alpha_{mn,k}(\hat{\sigma}^2_{\min})$ in (20) with q_δ in (5) and the safe choice $\hat{\sigma}^2_{\min} = (m\|\boldsymbol{Y}_m\|^2/n)_{(m)}$ (22), respectively. To highlight advantages of information pooling, we also compare with the "individual" blocking and soft-thresholding estimators, denoted by $\hat{\boldsymbol{\theta}}^{\mathcal{B}}_i$ and $\hat{\boldsymbol{\theta}}^{ST}_i$, respectively, using $\sigma^2 = 1$. Note that both methods only use the data $\{Y_{ik}\}_{k\le m}$ from the ith experiment to estimate the effects $\boldsymbol{\theta}_i$. Specifically, we use the weakly geometric blocking scheme for $\hat{\boldsymbol{\theta}}^{\mathcal{B}}_i$ and the threshold level $\sqrt{2\log m}$ for $\hat{\boldsymbol{\theta}}^{ST}_i$. See Tsybakov (2009) and Donoho and Johnstone (1994) for explicit formula.

Shown in Table 1 are the average and maximal ℓ^2 errors over n recoveries, $\{\|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i\|^2_{\ell_2}\}_{i\le n}$, for both decreasing and permuted $\{\lambda_k\}_{k\le m}$ using different methods with the decay parameter $\alpha = 2/3$. The results provide empirical evidence for the assertions in Theorem 2, and demonstrate the advantage of information pooling across experiments by our method, using both true and estimated variance, even when the sample size is moderate at $n = 100$. It is expected that the blocking method performs well for the case of decreasing $\{\lambda_k\}_{k\le m}$ with large sampling rate, for instance $m = 5000$, but degrades substantially when λ_k are randomly permuted. Although soft-thresholding is adaptive to permutation, it incurs much larger errors. We remark that, even improved by SURE-chosen thresholds, the performance of soft-thresholding is still inferior to the blocking method and does not alter the pattern of comparison (thus not reported). Moreover, we consider it fair to compare our tuning-free method using a universal threshold to those requiring similar computation. The results of $\alpha = 1$ illustrate a similar performance pattern (not reported) with overall smaller errors, as faster decay corresponds to smoother functions that are easier to be recovered. To appreciate the adaptivity to permutation, we depict a realization $\boldsymbol{\theta}_i$ and its random permutation in top panels of Figure 1, expanded in a smooth wavelet basis, specifically, the symlets of order 6 that is nearly symmetric with a minimum support size for that order corresponding to the number of vanishing moments. It is evident that the function corresponding to permuted effects exhibits striking local features and thus presents a more challenging pattern for recovery from a smoothness regularity perspective. We also inspect the robustness of the proposed method in the case of recovering new observations generated distributions that violate the Gaussian assumption. The results, reported in Table S1 of the Supplementary Material, provide an empirical support to Theorem 4.

It is of also interest to examine the influence of quantile on the estimation of σ^2 which manifests in a tradeoff between quality of recovery and model complexity. We use data of size $n = m = 100$ for enhanced visualization with $\lambda_k = 2\alpha k^{-(2\alpha+1)}$, and calculate recoveries of $\{\boldsymbol{\theta}_i\}_{i\le n}$ by Stein estimation (20) using $\hat{\sigma}^2_p$ (19) over a range of percentages. From bottom

Table 1. Average and maximal ℓ^2 errors ($\times 10^2$) over n recoveries, $\{\|\boldsymbol{\theta}_i - \hat{\boldsymbol{\theta}}_i\|_{\ell^2}^2\}_{i\leq n}$, for both decreasing and permuted sequences of $\{\lambda_k\}_{k\leq m}$ using different methods, when the decay parameter $\alpha = 2/3$ and the sampling rate m varies.

		Decreasing Sequence $\{\lambda_k\}_{k\leq m}$					
		$n = 10$		$n = 100$		$n = 1000$	
		Avg	Max	Avg	Max	Avg	Max
$m = 50$	Oracle $\hat{\boldsymbol{\theta}}_i^{o,a}$	15.4	25.6	15.2	35.3	15.2	44.8
	$\hat{\boldsymbol{\theta}}_i^{RS}(\sigma^2)$	32.4	60.7	21.3	50.3	16.9	48.2
	$\hat{\boldsymbol{\theta}}_i^{RS}(\hat{\sigma}^2_{\min})$	21.5	36.9	20.9	49.2	17.4	49.3
	Block $\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}$	21.5	37.2	21.4	51.1	21.4	64.7
	Soft $\hat{\boldsymbol{\theta}}_i^{ST}$	46.3	77.2	46.1	101	46.1	123
$m = 5000$	Oracle $\hat{\boldsymbol{\theta}}_i^{o,a}$	1.19	1.46	1.19	1.66	1.19	1.83
	$\hat{\boldsymbol{\theta}}_i^{RS}(\sigma^2)$	2.88	3.78	1.84	2.64	1.39	2.11
	$\hat{\boldsymbol{\theta}}_i^{RS}(\hat{\sigma}^2_{\min})$	9.88	10.5	1.60	2.25	1.39	2.11
	Block $\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}$	1.58	1.99	1.58	2.28	1.58	2.54
	Soft $\hat{\boldsymbol{\theta}}_i^{ST}$	5.97	7.23	5.97	8.08	5.97	8.76
		Permuted Sequence $\{\lambda_k\}_{k\leq m}$					
		$n = 10$		$n = 100$		$n = 1000$	
		Avg	Max	Avg	Max	Avg	Max
$m = 50$	Oracle $\hat{\boldsymbol{\theta}}_i^{o,a}$	15.3	26.0	15.2	35.5	15.2	44.7
	$\hat{\boldsymbol{\theta}}_i^{RS}(\sigma^2)$	32.3	61.0	21.3	50.3	16.9	47.9
	$\hat{\boldsymbol{\theta}}_i^{RS}(\hat{\sigma}^2_{\min})$	21.4	36.8	20.9	49.0	17.4	49.1
	Block $\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}$	73.0	182	70.0	320	71.1	475
	Soft $\hat{\boldsymbol{\theta}}_i^{ST}$	46.0	76.7	46.2	101	46.1	122
$m = 5000$	Oracle $\hat{\boldsymbol{\theta}}_i^{o,a}$	1.19	1.46	1.19	1.66	1.19	1.83
	$\hat{\boldsymbol{\theta}}_i^{RS}(\sigma^2)$	2.88	3.78	1.84	2.65	1.39	2.11
	$\hat{\boldsymbol{\theta}}_i^{RS}(\hat{\sigma}^2_{\min})$	9.57	10.2	1.60	2.26	1.39	2.12
	Block $\hat{\boldsymbol{\theta}}_i^{\mathcal{B}}$	53.7	146	53.7	270	51.5	376
	Soft $\hat{\boldsymbol{\theta}}_i^{ST}$	5.97	7.25	5.97	8.09	5.97	8.77

left panel in Figure 1 showing the pattern of maximal recovery error as a function of p, as dictated by Theorem 5, we observe an "elbow" type transition from consistency to a sudden risk hike, when p exceeds certain threshold p^* approaching one. Together with the bottom right panel showing the corresponding model complexity measured by the number of retained variables, we see that exercising caution in choosing a larger p may be worthwhile for balancing model complexity and recovery quality.

4.2. Application to Phoneme data

We apply our recovery method to the Phoneme dataset studied in Hastie *et al.* (1995). The data consists of $n = 4509$ equally-spaced log-periodogram sequences of length $m = 256$ derived from continuous speech of male subjects. Each sequence of log-periodograms belongs to one of the five categories: "sh", "dcl", "iy", "aa" and "ao". To assess the classification performance, we randomly split the data into a training sample of 1000 trajectories versus a testing sample of 3509 for each of 100 monte-carlo iterations. In each run, we perform three

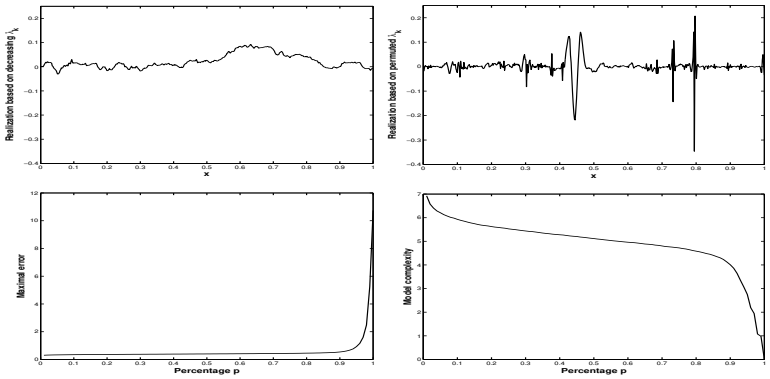

Fig. 1. Top row displays a randomly chosen realization generated from the decreasing (left panel) and permuted (right panel) sequences of $\{\lambda_k\}_{k\leq m}$ with $m = 500$, expanded in a smooth wavelet basis. Bottom row shows the maximum error ($\times 10^2$) (left panel) and model complexity measured by number of retained variables (right panel) as functions of p used in $\hat{\sigma}_p^2$ (19), respectively, when $n = m = 100$.

procedures on the training sample: penalized spline smoothing of individual functions (i.e., pre-smoothing, Ramsay and Silverman, 2005) followed by eigen-decomposition of covariance, denoted by RAMS; the Principal Analysis through Conditional Expectation (PACE) method (Yao *et al.*, 2005) based on kernel smoothing to raw covariances formed from noisy data (i.e., post-smoothing); the proposed Stein estimation in a smooth wavelet basis, denoted by STEIN. For recoveries by this method, a periodized Battle-Lemarie wavelet system (spline based) with 6 vanishing moments was used. Then a simple linear discriminant classifier is fit in the reduced model spaces, the FPC scores resulting from RAMS and PACE, and the retained coefficients from our method. For a comprehensive comparison over various model dimensions, we train the classification rules for RAMS and PACE methods by retaining FPCs that explain 90%, 95% and 99% of total variance, respectively, while different percentages for $\hat{\sigma}^2(p)$ are used in our method, specifically $p = 0.60, 0.70, 0.75, 0.80, 0.85, 0.90$. Results in Table 2 shows that, while classification by each method appears indistinguishable in its optimal case and is also comparable to the benchmark in Hastie *et al.* (1995), our recovery method is seen much more stable over different complexities.

To illustrate computational savings we use synthetic data, with recovered curves and estimated variance $\hat{\sigma}^2$ used to extrapolate the original data to larger sampling rates m. To be specific, we add noise following $N(0, \hat{\sigma}^2(p))$ with $p = 0.9$ to the curves recovered curves by our method to generate larger synthetic datasets with $m = 512, 1024, 2048, 4096, 8192, 16384$, respectively. We use the public Matlab packages at http://www.psych.mcgill.ca/misc/fda and http://www.stat.ucdavis.edu/PACE for RAMS and PACE methods with default selections for smoothing parameters, on a Mac Mini with a 2.3 GHz Intel Core i7 chip and 8 GB of DDR3 RAM. The average computation times for one full sample of $n = 4509$ reported in Table 2 indicates significant time savings for large m regimes by the proposed method, which approximately agree with the computation complexity: $O(nm^2 + m^3)$ for RAMS, $O(nm^2 + m^6)$ for PACE, and $O(nm + m \log m)$ for our method. Note that the PACE method is designed for sparse functional data with random design, thus encounters computational challenge in dense designs when data are not binned prior to smoothing.

Table 2. Classification error (CE) is based on 100 random partitions of Phoneme data into training ($n = 1000$) and testing ($N = 3509$) samples. Model complexity is determined by the percentage p in $\hat{\sigma}^2(p)$ for the proposed method (STEIN), and by the total variance explained (TVE) for Ramsay's (RAMS) and PACE methods. The average computation time is for a full sample ($n = 4509$) of extrapolated data by adding noise to recovered curves at increased sampling rates m.

	Comparison of Performance					
	STEIN					
$\hat{\sigma}^2(p)$	$p = .60$	$p = .70$	$p = .75$	$p = .80$	$p = .85$	$p = .90$
CE (%)	7.66 (.32)	7.68 (.33)	7.85 (.38)	7.82 (.35)	7.48 (.30)	7.65 (.30)
	RAMS			PACE		
TVE	90%	95%	99%	90%	95%	99%
CE (%)	10.9 (.36)	8.15 (.46)	7.72 (.31)	12.2 (.60)	9.25 (.33)	7.68 (.33)
	Comparison of Computation time (minute)					
m	512	1024	2048	4096	8192	16384
STEIN	.068	.078	.113	.169	.328	.860
RAMS	.108	.240	.664	3.76	26.5	221
PACE	6.95	30.2	98.9	426	—	—

5. Potential Application to Other Statistical Problems

The statistical principles and mathematical techniques explored in this article may be lent to broad classes of problems involving information pooling across similar experiments. An example is the change-point problem that has been traditionally treated individually (Fryzlewicz, 2014, 2016). More recently, this problem has been considered in the context of panel data where common structure of multiple lends strength to multiple change-point detection (Cho, 2016). We briefly outline a treatment of multiple change-points for panel data through lens of multiple GSMs, which may deserve a further study.

Assume that the data consist of $Y_{ij} = \theta_{ij} + z_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m+1$, with θ_{ij} varying independently across i (individuals) and being piecewise constants across j (time) with T (unknown) change-points at $k_l \in \{2, \ldots, m+1\}$, $l = 1, \ldots, T$. At the unknown change-points, assume that the $\theta_{ij} \sim N(0, \lambda)$ are generated independently of each other and are also independent of the z_{ij} which are i.i.d. $N(0, 1)$. We difference the data to form GSMs, denoting $Y^*_{ij} = Y_{i,j+1} - Y_{ij}$, similarly θ^*_{ij} and z^*_{ij},

$$Y^*_{ij} = \theta^*_{ij} + z^*_{ij}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, m.$$

Note that $\theta^*_{ij} = 0$ except at the unknown change points in which case $\theta^*_{ij} \sim N(0, 2\lambda)$ and $z^*_{ij} \overset{i.i.d.}{\sim} N(0, 2)$. Thus we have $||Y^*_{\cdot j}||^2 \approx 2n(1+\lambda)$ for $j \in \{k_1, \ldots, k_T\}$ and $2n$ otherwise. Applying the proposed Stein procedure to estimate θ^*_{ij}, we find that $\hat{\theta}^*_{ij}$ are non-zero, with high probability, at the locations satisfying

$$||Y^*_{\cdot j}||^2/(2n) > 1 + 2\delta,$$

whenever $\lambda > (1+2\delta)(1+\delta)/(1-\delta) - 1 = o(1)$. Shown in Figure 2 is an illustration of $||Y^*_{i\cdot}||^2/(n\hat{\sigma}^2)$ with $n = 100, m = 64$ and $\lambda = 3$ in a Monte Carlo sample, where we take the median $\hat{\sigma}^2 = n^{-1}\text{med}(||Y^*_{\cdot j}||^2)$. Moreover, we have successfully detected the three change-points in nearly all of the 100 Monte Carlo samples.

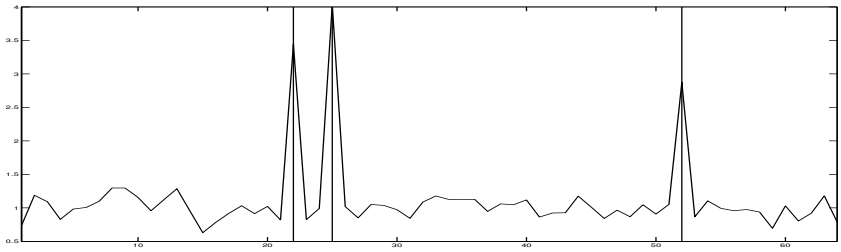

Fig. 2. Shown are the ratios $||Y_i^*||^2/\text{med}(||Y_i^*||^2)$ for one Monte Carlo sample with $n = 100$, $m = 64$, $\lambda = 3$ and change-points located at $j = 22, 25, 52$. The vertical lines indicate successful detection of these change-points.

Supplementary Material

Due to space constraint, we collect some additional simulation and theoretical results, and the proofs of main lemmas, theorems and propositions in the Supplementary Material.

Acknowledgements

This research is partially supported by Natural Science and Engineering Research Council of Canada.

References

Belitser, E. and Ghosal, S. (2003) Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *Annals of Statistics*, **31**, 536–559.

Brown, L. D., Cai, T. T., Low, M. G. and Zhang, C.-H. (2002) Asymptotic equivalence theory for nonparametric regression with random design. *Annals of Statistics*, **30**, 688–707.

Brown, L. D. and Levine, M. (2007) Variance estimation on nonparametric regression via the difference sequence method. *Annals of Statistics*, **35**, 2219–2232.

Brown, L. D. and Low, M. G. (1996) Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, **24**, 2384–2398.

Cai, T. T. (1999) Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Annals of Statistics*, **27**, 898–924.

— (2008) On information pooling, adaptability and superefficiency in nonparametric function estimation. *Journal of Multivariate Analysis*, **99**, 412–436.

— (2012) Minimax and adaptive inference in nonparametric function estimation. *Statististical Science*, **27**, 31–50.

Cai, T. T. and Yuan, M. (2011) Optimal estimation of the mean function based on discretely sampled functional data: phase transition. *Annals of Statistics*, **39**, 2330–2355.

Candes, E. J. (2006) Modern statistical estimation via oracle inequalities. *Acta Numerica*, **15**, 1–69.

Cavalier, L. and Tsybakov, A. (2002) Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields*, **123**, 323–354.

Cho, H. (2016) Change-point detection in panel data via double CUSUM statistic. *Electronic Journal of Statistics*, **10**, 2000–2038.

Cohen, A. (2003) *Numerical Analysis of Wavelet Methods*. Elsevier, first edn.

Donoho, D. L. (1993) Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, **1**, 100–115.

Donoho, D. L. and Johnstone, I. M. (1994) Ideal spatial adaption by wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society, Series B*, 301–369.

Donoho, D. L., Liu, R. C. and MacGibbon, B. (1990) Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, **18**, 1416–1437.

Fan, J. and Yao, Q. (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645–660.

Freedman, D. (1999) Wald lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Annals of Statistics*, **27**, 1119–1141.

Fryzlewicz, P. (2014) Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, **42**, 2243–2281.

— (2016) Tail-greedy bottom-up data decompositions and fast multiple change-point detection. Available at http://stats.lse.ac.uk/fryzlewicz/tguh/tguh.pdf.

Hall, P. and Carroll, R. J. (1989) Variance function estimation in regression: the effect of estimating the mean. *Journal of the Royal Statistical Society, Series B*, **51**, 521–528.

Hall, P., Key, J. and Titterington, D. (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.

Hall, P. and Marron, J. S. (1990) On variance estimation in nonparametric regression. *Biometrika*, **77**, 415–419.

Hall, P., Müller, H.-G. and Wang, J.-L. (2006) Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, **34**, 1493–1517.

Hastie, T., Buja, A. and Tibshirani, R. (1995) Penalized discriminant analysis. *Annals of Statistics*, **23**, 73–102.

Johnstone, I. M. (2015) *Gaussian Estimation: Sequence and Multiresolution Models*. Unpublished Mongraph.

Li, Y. and Hsing, T. (2010) Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, 3321–3351.

Mallat, S., Papanicolaou, G. and Zhang, Z. (1998) Adaptive covariance estimation of locally stationary processes. *Annals of Statistics*, **26**, 1–47.

Müller, H.-G. and Stadtmüller, U. (1987) Estimation of heteroscedasticity in regression analysis. *Annals of Statistics*, **15**, 610–625.

Pickands, J. (1969) An iterated logarithm law for the maximum in a stationary gaussian sequence. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **12**, 344–353.

Ramsay, J. O. and Silverman, B. W. (2005) *Functional Data Analysis*. Springer Series in Statistics. New York: Springer, second edn.

Reiß, M. (2008) Asymptotic equivalence for nonparametric regression with multivariate and random design. *Annals of Statistics*, **36**, 1957–1982.

Ritter, K. (2000) *Average-Case Analysis of Numerical Problems*. Springer Verlag, Berlin.

Ritter, K., Wasilkowski, G. W. and Woźniakowski, H. (1995) Multivariate integration and approximation for random fields satisfying sacks-ylvisaker conditions. *The Annals of Applied Probability*, 518–540.

Szabó, B. T., van der Vaart, A. W. and van Zanten, J. H. (2013) Empirical Bayes scaling of Gaussian priors in the white noise model. *Electronic Journal of Statistics*, **7**, 991–1018.

Tsybakov, A. B. (2009) *Introduction to Nonparametric Estimation*. Springer.

Walter, G. G. and Shen, X. (2000) *Wavelets and other orthogonal systems*. Chapman & Hall/CRC.

Xiao, L., Zipunnikov, V., Ruppert, D. and Crainiceanu, C. (2016) Fast covariance estimation for high-dimensional functional data. *Statistics and computing*, **26**, 409–421.

Yao, F., Müller, H.-G. and Wang, J.-L. (2005) Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100**, 577–590.

Zhang, C.-H. (2005) General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Annals of Statistics*, **33**, 54–100.

Zhao, L. H. (2000) Bayesian aspects of some nonparametric problems. *Annals of Statistics*, **28**, 532–552.