

Journal of the American Statistical Association 8

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

# **Kernel Meets Sieve: Post-Regularization Confidence Bands for Sparse Additive Model**

Junwei Lu, Mladen Kolar & Han Liu

To cite this article: Junwei Lu , Mladen Kolar & Han Liu (2020) Kernel Meets Sieve: Post-Regularization Confidence Bands for Sparse Additive Model, Journal of the American Statistical Association, 115:532, 2084-2099, DOI: 10.1080/01621459.2019.1689984

To link to this article: <u>https://doi.org/10.1080/01621459.2019.1689984</u>

View supplementary material



Published online: 11 Mar 2020.

Submit your article to this journal 🗹

Article views: 678



🜔 View related articles 🗹

View Crossmark data 🗹



Citing articles: 1 View citing articles 🗹

# Kernel Meets Sieve: Post-Regularization Confidence Bands for Sparse Additive Model

Junwei Lu<sup>a</sup>, Mladen Kolar<sup>b</sup>, and Han Liu<sup>c</sup>

<sup>a</sup>Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA; <sup>b</sup>Booth School of Business, The University of Chicago, Chicago, IL; <sup>c</sup>Department of Computer Science, Northwestern University, Evanston, IL

### ABSTRACT

We develop a novel procedure for constructing confidence bands for components of a sparse additive model. Our procedure is based on a new kernel-sieve hybrid estimator that combines two most popular nonparametric estimation methods in the literature, the kernel regression and the spline method, and is of interest in its own right. Existing methods for fitting sparse additive model are primarily based on sieve estimators, while the literature on confidence bands for nonparametric models are primarily based upon kernel or local polynomial estimators. Our kernel-sieve hybrid estimator combines the best of both worlds and allows us to provide a simple procedure for constructing confidence bands in high-dimensional sparse additive models. We prove that the confidence bands are asymptotically honest by studying approximation with a Gaussian process. Thorough numerical results on both synthetic data and real-world neuroscience data are provided to demonstrate the efficacy of the theory. Supplementary materials for this article are available online.

#### **ARTICLE HISTORY**

Received March 2018 Accepted October 2019

Taylor & Francis

Check for updates

Tavlor & Francis Group

### **KEYWORDS**

Confidence band; Kernel method; Sieve estimator; Sparse additive model

# 1. Introduction

Nonparametric regression investigates the relationship between a target variable *Y* and many input variables  $\mathbf{X} = (X_1, \dots, X_d)^T$ without imposing strong assumptions. Consider a model

$$Y = f(X) + \varepsilon, \tag{1.1}$$

where  $X \in \mathcal{X}^d \subseteq \mathbb{R}^d$  is a *d*-dimensional random vector in  $\mathcal{X}^d$ ,  $\varepsilon$  is random error satisfying  $\mathbb{E}[\varepsilon | X] = 0$ , and *Y* is a target variable. The goal is to estimate the unknown function  $f : \mathcal{X}^d \mapsto \mathbb{R}$ . When *d* is small, fitting a fully nonparametric model (1.1) is feasible (Wasserman 2006). However, the interpretation of such a model is challenging. Furthermore, when *d* is large, consistently fitting  $f(\cdot)$  is only possible under additional structural assumptions due to the curse of dimensionality.

A commonly used structural assumption on  $f(\cdot)$  is that it takes an additive form

$$Y = \mu + \sum_{j=1}^{a} f_j(X_j) + \varepsilon$$
, and  $\mathbb{E}_{X_j}[f(X_j)] = 0$ , (1.2)

where  $\mu$  is a constant and  $f_j(\cdot)$ ,  $j = 1, \ldots, d$ , are smooth univariate functions (Friedman and Stuetzle 1981; Stone 1985; Hastie and Tibshirani 1990). Under an additional assumption that only *s* components are nonzero ( $s \ll d$ ), significant progress has been made in understanding additive models in high dimensions (Sardy and Tseng 2004; Lin and Zhang 2006; Meier, van de Geer, and Bühlmann 2009; Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010; Koltchinskii and Yuan 2010; Kato 2012; Lou et al. 2014; Petersen, Witten, and Simon 2014). These articles establish theoretical results on the estimation rate of sparse additive models, however, it remains unclear how to perform statistical inference for the model. Confidence bands can provide uncertainty assessment for components of the model and have been widely studied in the literature with dimension fixed (Härdle 1989; Sun and Loader 1994; Fan and Zhang 2000; Claeskens and Van Keilegom 2003; Zhang and Peng 2010). However, it remains an open question how to construct confidence bands in high-dimensional setting, primarily because the direct generalization of those ideas is challenging. Confidence bands proposed in the classical literature with fixed dimensionality d are mostly built upon kernel or local polynomial methods (Opsomer and Ruppert 1997; Fan and Jiang 2005), while existing estimators for sparse additive model are sieve-type estimators based on basis expansion. To bridge the gap, we propose a novel sparse additive model estimator called kernel-sieve hybrid estimator, which combines advantages from both the sieve and kernel methods. On one side, we can uniformly control the supreme norm rate of our estimator as typical sieve estimators for sparse additive models, while on the other, we can utilize the extreme value theory of kernel-type estimator to construct the confidence band.

To establish the validity of the proposed confidence bands, we develop three new technical ingredients: (1) the analysis of the suprema of a high-dimensional empirical process that arises from kernel-sieve hybrid regression estimator, (2) a de-biasing method for the proposed estimator, and (3) the approximation analysis for the Gaussian multiplier bootstrap procedure. The supremum norm for our estimator is derived by applying results on the suprema of empirical processes (van der Vaart and Wellner 1996; Bousquet 2002; Koltchinskii 2011). The debiasing procedure for the kernel-sieve hybrid regression esti-

© 2020 American Statistical Association

**CONTACT** Junwei Lu Sinweilu@hsph.harvard.edu Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115. Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

B Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

mator extends the approach used in the  $\ell_1$  penalized highdimensional linear regression (Zhang and Zhang 2013; Javanmard and Montanari 2014; van de Geer et al. 2014). Compared to the existing literature, this is the first work considering the de-biasing procedure for a high-dimensional nonparametric model. To prove the validity of the confidence band constructed by the Gaussian multiplier bootstrap, we generalize the method proposed in Chernozhukov, Chetverikov, and Kato (2014a, 2014b) to the high-dimensional nonparametric models.

### 1.1. Related Literature

Our work contributes to two different areas, and make new methodological and technical contributions in both of them.

First, we contribute to a growing literature on highdimensional inference. Initial work on high-dimensional statistics has focused on estimation and prediction (see, e.g., Bühlmann and van de Geer 2011, for a recent overview) and much less work has been done on quantifying uncertainty, for example, hypothesis testing and confidence intervals. Recently, the focus has started to shift toward the latter problems. Initial work on construction of *p*-values in high-dimensional models relied on correct inclusion of the relevant variables (Meinshausen, Meier, and Bühlmann 2009; Wasserman and Roeder 2009). Meinshausen and Bühlmann (2010) and Shah and Samworth (2013) studied stability selection procedure, which provides the family-wise error rate for any selection procedure. Hypothesis testing and confidence intervals for lowdimensional parameters in high-dimensional linear and generalized linear models are studied in Belloni, Chernozhukov, and Hansen (2013), Belloni, Chernozhukov, and Wei (2013), van de Geer et al. (2014), Javanmard and Montanari (2014), Javanmard and Montanari (2013), and Farrell (2013). These methods construct honest, uniformly valid confidence intervals and hypothesis test based on the  $\ell_1$  penalized estimator in the first stage. Similar results are obtained in the context of  $\ell_1$  penalized least absolute deviation and quantile regression (Belloni, Chernozhukov, and Kato 2013, 2015). Kozbur (2015) extended the approach developed in Belloni, Chernozhukov, and Hansen (2013) to a nonparametric regression setting, where a pointwise confidence interval is obtained based on the penalized series estimator. Meinshausen (2013) studied construction of one-sided confidence intervals for groups of variables under weak assumptions on the design matrix. Lockhart et al. (2014) studied significance of the input variables that enter the model along the lasso path. Lee et al. (2013) and Taylor et al. (2014) performed post-selection inference conditional on the selected model. Chatterjee and Lahiri (2013), Liu and Yu (2013), Chernozhukov, Chetverikov, and Kato (2013), and Lopes (2014) studied properties of the bootstrap in high-dimensions. Our work is different to the existing literature as it enables statisticians to make global inference under a nonparametric highdimensional regression setting for the first time.

Second, we contribute to the literature on high-dimensional nonparametric estimation, which has recently seen a lot of activity. Lafferty and Wasserman (2008), Bertin and Lecué (2008), Comminges and Dalalyan (2012), and Yang and Tokdar (2014) studied variable selection in a high-dimensional nonparametric

regression setting without assuming structural assumptions on  $f(\cdot)$  beyond that it depends only on a subset of variables. A large number of articles have studied the sparse additive model in (1.2) (Sardy and Tseng 2004; Lin and Zhang 2006; Avalos, Grandvalet, and Ambroise 2007; Meier, van de Geer, and Bühlmann 2009; Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010; Koltchinskii and Yuan 2010; Kato 2012; Raskutti, Wainwright, and Yu 2012; Rosasco et al. 2013; Lou et al. 2014; Petersen, Witten, and Simon 2014; Wahl 2014). In addition, Xu, Chen, and Lafferty (2014) studied a high-dimensional convex nonparametric regression. Dalalyan, Ingster, and Tsybakov (2014) study the compound model, which includes the additive model as a special case. Our approach differs from the existing literature in that we consider the ATLAS model, in which the additive model is only used as an approximation to the unknown function  $f(\cdot)$  at a fixed point z and allow such approximation to change with z. Our approach only imposes a local sparsity structure and thus allows for more flexible modeling. We also develop a novel method for estimation and inference. Meier, van de Geer, and Bühlmann (2009), Huang, Horowitz, and Wei (2010), Koltchinskii and Yuan (2010), Raskutti, Wainwright, and Yu (2012), and Kato (2012) developed estimation schemes mainly based on the basis approximation and sparsity-smoothness regularization. Our estimator approximates the function locally using a loss function combining both basis expansion and kernel method with a hybrid  $\ell_1/\ell_2$ -penalty. Our theoretical analysis also provides novel technical tools that were not available before and are of independent interest.

### 1.2. Organization of the Article

The rest of the article is organized as follows. In Section 2, we introduce the penalized kernel-sieve hybrid regression estimator as a solution to an optimization program. We then construct a confidence band for a component of a sparse additive model based on the proposed estimator. Section 3 provides the theoretical results on the statistical rate of convergence for the estimator and show that the proposed confidence band is honest. In Section 4, we generalize our method to nonparametric functions beyond sparse additive model. The numerical experiments for synthetic and real data are collected in Section 5.

### 1.3. Notation

Let [n] denote the set  $\{1, \ldots, n\}$  and let  $\mathbb{1}\{\cdot\}$  denote the indicator function. For a vector  $\mathbf{a} \in \mathbb{R}^d$ , we let  $\operatorname{supp}(\mathbf{a}) = \{j \mid a_j \neq 0\}$  be the support set (with an analogous definition for matrices  $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ ),  $\|\mathbf{a}\|_q$ , for  $q \in [1, \infty)$ , the  $\ell_q$ -norm defined as  $\|\mathbf{a}\|_q = (\sum_{i \in [n]} |a_i|^q)^{1/q}$  with the usual extensions for  $q \in \{0, \infty\}$ , that is,  $\|\mathbf{a}\|_0 = |\operatorname{supp}(\mathbf{a})|$  and  $\|\mathbf{a}\|_{\infty} = \max_{i \in [n]} |a_i|$ . If the vector  $\mathbf{a} \in \mathbb{R}^d$  is decomposed into groups such that  $\mathbf{a} = (\mathbf{a}_{\mathcal{G}_1}, \ldots, \mathbf{a}_{\mathcal{G}_g})^T$ , where  $\mathcal{G}_1, \ldots, \mathcal{G}_g \subset [d]$  are disjoint sets, we denote  $\|\mathbf{a}\|_{p,q}^q = \sum_{k=1}^g \|\mathbf{a}_{\mathcal{G}_k}\|_p^q$  and  $\|\mathbf{a}\|_{p,\infty} = \max_{k \in [g]} \|\mathbf{a}_{\mathcal{G}_k}\|_p$  for any  $p, q \in [1, \infty)$ . We also denote the set  $\{1, \ldots, j-1, j+1, \ldots, d\}$  as  $\backslash j$  and the vector  $\mathbf{a}_{\backslash j} = (a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_d)^T$ . For the function  $f \in L^2(\mathbb{R})$ , we define the  $L^2$  norm  $\|f\|_2 = [\int f^2(x) dx]^{1/2}$ , the supremum norm  $\|f\|_{\infty} = \sup_{x \in \mathbb{R}} |f(x)|$  and the  $L^2(\mathbb{P})$  norm  $\|f\|_{L^2(\mathbb{P})} = [\int f^2(x) d\mathbb{P}]^{1/2}$ . For a matrix  $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$ , we use

the notation vec(**A**) to denote the vector in  $\mathbb{R}^{n_1 n_2}$  formed by stacking the columns of **A**. We denote the Frobenius norm of **A** by  $||\mathbf{A}||_F^2 = \sum_{i \in [n_1], j \in [n_2]} \mathbf{A}_{ij}^2$  and denote the operator norm as  $\|\mathbf{A}\|_2 = \sup_{\|\mathbf{v}\|_2=1} \|\mathbf{A}\mathbf{v}\|_2$ . For two sequences of numbers  $\{\alpha_n\}_{n=1}^{\infty}$  and  $\{\beta_n\}_{n=1}^{\infty}$ , we use  $a_n = O(\beta_n)$  to denote that  $\alpha_n \leq C\beta_n$  for some finite positive constant *C*, and for all *n* large enough. If  $\alpha_n = O(\beta_n)$  and  $\beta_n = O(\alpha_n)$ , we use the notation  $\alpha_n \approx \beta_n$ . The notation  $\alpha_n = o(\beta_n)$  is used to denote that  $a_n\beta_n^{-1} \xrightarrow{n \to \infty} 0$ . Throughout the article, we let *c*, *C* be two generic absolute constants, whose values may change from line to line.

# 2. Penalized Kernel-Sieve Hybrid Regression

In this section, we describe our new nonparametric estimator that combines the local kernel regression with the B-spline based sieve method. The goal is to estimate component functions in the additive model (1.2) and construct a confidence band for one component of the model. The kernel-sieve hybrid regression applies the local kernel regression over the component of interest and uses basis expansion for the rest of components. The group lasso penalty is used to shrink the coefficients in the expansion and select relevant variables locally.

We first introduce the Hölder class  $\mathcal{H}(\gamma, L)$  of functions.

*Definition 2.1.* The  $\gamma$ th Hölder class  $\mathcal{H}(\gamma, L)$  on  $\mathcal{X}$  is the set of  $\ell = \lfloor \gamma \rfloor$  times differentiable functions  $f : \mathcal{X} \mapsto \mathbb{R}$ , where  $\lfloor \gamma \rfloor$  represents the largest integer smaller than  $\gamma$ . The derivative  $f^{(\ell)}$  satisfies

$$|f^{(\ell)}(x) - f^{(\ell)}(y)| \le L|x - y|^{\gamma - \ell}, \quad \text{for any } x, y \in \mathcal{X}.$$

Let  $X = (X_1, \ldots, X_d)^T$  be a *d*-dimensional random vector in  $\mathcal{X}^d \subseteq \mathcal{X}^d$ . We will consider both the case where  $\mathcal{X}$  is compact and the case where  $\mathcal{X}$  is unbounded. The sparse additive model (SpAM) is of the form given in (1.2), with only a small number of additive components nonzero. Let  $\mathcal{S} \subseteq [d]$  be of size  $s = |\mathcal{S}| \ll d$ . Then the model in (1.2) can be written as

$$Y = \mu + \sum_{j \in S} f_j(X_j) + \varepsilon$$
(2.1)

with  $f_j \in \mathcal{H}(2, L)$  for any  $j \in S$ . Moreover, we assume the identifiability condition that

$$\mathbb{E}[f_i(X_j)] = 0$$
, for all  $j = 1, \dots, d$ . (2.2)

Define the sparse additive functions class

$$\mathcal{K}_d(s) = \left\{ f = \sum_{j \in \mathcal{S}} f_j(X_j) \ \Big| \ |\mathcal{S}| \le s, f_j \in \mathcal{H}(2, L) \text{ and} \\ \mathbb{E}[f_j(X_j)] = 0, \text{ for } j \in \mathcal{S} \right\}.$$
(2.3)

Let  $\{(X_i, Y_i)\}_{i=1}^n$  be *n* independent random samples of (X, Y) distributed according to (2.1). Before describing our estimator, we first introduce the centered basis functions that will be used in the estimation. Let  $\{\phi_1, \ldots, \phi_m\}$  be the normalized B-spline basis functions (Schumaker 2007). Given *m* basis functions, we

denote  $f_{jm}(x)$  as the projection of  $f_j$  onto the space spanned by the basis,  $\mathcal{B}_m = \text{Span}(\phi_1, \dots, \phi_m)$ . In particular, we define

$$f_{jm}(\cdot) := \underset{f \in \mathcal{B}_m}{\arg\min} \|f - f_j\|_2 = \sum_{k=1}^m \beta_{jk}^* \psi_{jk}^*(\cdot), \qquad (2.4)$$

where  $\psi_{ik}^*$ 's are the locally centered bases defined as

$$\psi_{jk}^*(x) = \phi_k(x) - \mathbb{E}[\phi_k(X_j)], \text{ for all } j \in [d], m \in [k].$$
 (2.5)

Notice that basis functions  $\{\psi_{jk}^*\}_{j \in [d], k \in [m]}$  satisfy  $\mathbb{E}[\psi_{jk}^*(X_j)] = 0$ . This property ensures that  $f_{jm}(\cdot)$  also satisfies the identifiability condition (2.2). To compute  $\psi_{jk}^*$  we need to estimate the unknown  $\mathbb{E}[\phi_k(X_j)]$  by  $\bar{\phi}_{jk} = n^{-1} \sum_{i=1}^n \phi_k(X_{ij})$ . The centered B-spline basis in (2.5) as then  $\psi_{ik}(x) = \phi_k(x) - \bar{\phi}_{ik}$ .

With this notation, we are ready to introduce the penalized kernel-sieve hybrid regression estimator. Let the kernel function  $K : \mathcal{X} \mapsto \mathbb{R}$  be a symmetric density function with bounded support and denote  $K_h(\cdot) = h^{-1}K(\cdot/h)$  where h > 0 is the bandwidth. The kernel-sieve hybrid loss function at a fixed point  $z \in \mathcal{X}$  is given as

$$\mathcal{L}_{z}(\alpha, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} K_{h}(X_{i1} - z) \left( Y_{i} - \bar{Y} - \alpha - \sum_{j=2}^{d} \sum_{k=1}^{m} \psi_{jk}(X_{ij}) \boldsymbol{\beta}_{jk} \right)^{2}, \qquad (2.6)$$

where  $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ . Let  $\boldsymbol{\beta} = (\boldsymbol{\beta}_2^T, \dots, \boldsymbol{\beta}_d^T)^T \in \mathbb{R}^{(d-1)m}$ with  $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j1}, \dots, \boldsymbol{\beta}_{jm})^T \in \mathbb{R}^m$  be the coefficients of B-spline basis functions. The penalized kernel-sieve hybrid estimator at  $z \in \mathcal{X}$  is defined as

$$(\widehat{\alpha}_z, \widehat{\beta}_z) = \arg\min_{\alpha, \beta} \mathcal{L}_z(\alpha, \beta) + \lambda \mathcal{R}(\alpha, \beta),$$
 (2.7)

where the penalty function is

$$\mathcal{R}(\alpha, \boldsymbol{\beta}) = \sqrt{m} \cdot |\alpha| + \sum_{j \ge 2} ||\boldsymbol{\beta}_j||_2$$
(2.8)

with  $\lambda$  being a tuning parameter. We estimate the additive functions  $\{f_j\}_{j\in[d]}$  by  $\widehat{f}_1(z) = \widehat{\alpha}_z$  and  $\widehat{f}_j(x) = \sum_{k=1}^m \psi_{jk}(x)\widehat{\beta}_{jk;z}$  for  $j \ge 2$ . Based on  $\widehat{\alpha}_z, \widehat{\beta}_z$ , we also estimate the *d*-dimensional function  $f(z, x_2, \ldots, x_d) = f_1(z) + \sum_{j=2}^d f_j(x_j)$  by

$$\widehat{f}(z, x_2, \dots, x_d) = \widehat{\alpha}_z + \sum_{j=2}^d \sum_{k=1}^m \psi_{jk}(x_j) \widehat{\beta}_{jk;z}, \qquad (2.9)$$

where  $\hat{\beta}_{jk;z}$  is the coordinate of  $\hat{\beta}_z$  corresponding to the *k*th B-spline basis of the *j*th covariate.

*Remark 1.* The estimators  $\widehat{\alpha}_z$  and  $\widehat{\beta}_z$  are estimating different quantities. Notice that  $\widehat{\alpha}_z$  estimates the scalar  $f_1(z)$ , while  $\widehat{\beta}_z$  estimates the coefficients of B-splines. Given a function  $g(x) = \sum_{k=1}^{m} \beta_k \phi_k(x)$ , we have  $\|g\|_2^2 \approx m^{-1} \sum_{k=1}^{m} \beta_k^2$  (see, e.g., Corollary 15 in Chapter XI of de Boor (2001)). From this we see that the scales of  $\widehat{\alpha}_z$  and  $\widehat{\beta}_z$  are different, which explains the additional  $\sqrt{m}$  term multiplying  $|\alpha|$  in the penalty function (2.8).

### 2.1. Comparison to the Sieve Estimator

In this section, we explain why we consider the kernel-sieve estimator as the first step of a confidence band construction instead of the sieve estimator. In the literature of sparse additive model estimation, most articles consider the sieve-type estimator. For example, Huang, Horowitz, and Wei (2010) considered minimizing

$$\widehat{\boldsymbol{\beta}}^{\text{sieve}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \bar{Y} - \sum_{j=1}^{d} \sum_{k=1}^{m} \psi_{jk}(X_{ij}) \boldsymbol{\beta}_{jk} \right)^2 + \lambda \sum_{j=1}^{d} ||\boldsymbol{\beta}_j||_2, \qquad (2.10)$$

while similar variations were considered in Ravikumar et al. (2009), Meier, van de Geer, and Bühlmann (2009), Koltchinskii and Yuan (2010), and Kato (2012). These articles show that estimators like (2.10) are good enough to achieve the estimation consistency under the sparse additive model.

Kozbur (2015) proposed a post-nonparametric double selection procedure to conduct the inference for a differentiable functional of the function of interest,  $f_1$ , where estimation is performed by a sieve-type estimator. This method selects variables in three steps: (i) run Lasso regression  $\psi_{1k}(X_1)$  on  $\{\psi_{st}(X_s)\}_{s\geq 2,t\geq 1}$  and select the support  $I_k$  for all  $1\leq k\leq m$ ; (ii) run Lasso regression Y on  $\{\psi_{st}(X_s)\}_{s\geq 2,t\geq 1}$  and select the support  $I_0$ ; and (iii) run least square regression Y on  $\{\psi_{st}(X_s)\}$ for s, t belong to the support  $\bigcup_{k=0}^{m} I_k$ . The confidence interval of a functional  $a(f_1)$  could then be derived through the least square estimator in the last step. However, this approach cannot be directly used to construct a confidence band. First, Assumption 14 in Kozbur (2015) assumes that the functional  $a(f_1)$  is differentiable, which does not hold for the supremum operator. Second, the validity of the method is based on two high level assumptions on the variable selection (see Assumptions 9 and 10 in Kozbur (2015)) that are hard to verify in practice. In particular, they are not satisfied for the data generating process used in the simulation study in Section 5.

To sum up, it is challenging to study the uniform confidence band through pure sieve-type approaches. Technically, if we compare the loss functions of two estimators in (2.10) and (2.6), the sieve estimator approximates the function of interest  $f_1$ through its global basis expansion, while the kernel-sieve hybrid estimator only approximates  $f_1$  at the local point z by a scalar  $\alpha$ . Therefore, to study the asymptotic properties of the extreme value

$$\sup_{z \in \mathcal{X}} \left| \widehat{f}_1^{\text{sieve}}(z) - f_1(z) \right| = \sup_{z \in \mathcal{X}} \Big| \sum_{k=1}^m \psi_{1k}(z) \widehat{\boldsymbol{\beta}}_{1k}^{\text{sieve}} - f_1(z) \Big|,$$
(2.11)

we need to analyze the *m*-dimensional estimator  $\hat{\beta}_1^{\text{sieve}}$  whose dimension *m* is increasing with the sample size *n* at the rate  $m \approx n^{1/6}$ . This makes it challenging to estimate the asymptotic distribution of the extreme value statistic in (2.11). Kozbur (2015) studies the limiting distribution of a differential functional of  $\hat{\beta}_1^{\text{sieve}}$ , while the extreme value is more challenging as it is non-differentiable. We further note that most existing articles on

Algorithm 1 Randomized coordinate descent for group Lasso

for t = 1, 2, ... do Let  $\boldsymbol{\beta}_{+}^{(t)} = (\boldsymbol{\beta}_{1}^{(t)}, \boldsymbol{\beta}_{2}^{(t)T}, ..., \boldsymbol{\beta}_{j}^{(t)T})^{T}$ . Choose  $j_{t} = j \in [d]$  with probability 1/d. Compute  $T(\boldsymbol{\beta}_{j}^{(t)})$  for the *j*th block as

$$T(\boldsymbol{\beta}_{j}^{(t)}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^{\dim(\boldsymbol{\beta}_{j})}}{\operatorname{argmin}} \left\{ \frac{\mu}{2} \|\boldsymbol{\theta}\|_{2}^{2} + \langle \nabla_{j} \mathcal{L}_{z}(\boldsymbol{\beta}_{+}^{(t)}), \boldsymbol{\theta} \rangle + \lambda_{j} \|\boldsymbol{\theta} + \boldsymbol{\beta}_{j}^{(t)}\|_{2} \right\}.$$
 (2.14)

Update  $\boldsymbol{\beta}_{j}^{(t+1)} = \boldsymbol{\beta}_{j}^{(t)} + T(\boldsymbol{\beta}_{j}^{(t)}).$ end for

confidence bands are based on kernel or local polynomial methods (Härdle 1989; Sun and Loader 1994; Fan and Zhang 2000; Claeskens and Van Keilegom 2003; Zhang and Peng 2010). In comparison, the advantage of the kernel-sieve hybrid estimator is that it directly outputs a scalar estimator  $\hat{\alpha}_z$  of  $f_1(z)$ . This onedimensional estimator  $\hat{\alpha}_z$  allows us to construct a confidence band as we explain below. Furthermore, as we discuss in Section 4, the idea behind the kernel-sieve hybrid estimator can be extended to a number of different classes of nonparametric models for which the estimator in (2.10) does not generalize.

# 2.2. Computational Algorithm

In this section, we describe an algorithm to minimize (2.7). We start by introducing some extra notation. Denote  $\Psi = (\Psi_{1\bullet}, \dots, \Psi_{n\bullet})^T \in \mathbb{R}^{n \times (1+(d-1)m)}$ , where  $\Psi_{ij} = (\psi_{j1}(X_{ij}), \dots, \psi_{jm}(X_{ij}))^T$  and  $\Psi_{i\bullet} = (1, \Psi_{i2}^T, \dots, \Psi_{id}^T)^T \in \mathbb{R}^{1+(d-1)m}$  for  $i \in [n]$  and  $j \ge 2$ . We also write  $\Psi = (\Psi_{\bullet 1}, \dots, \Psi_{\bullet d})$ , where  $\Psi_{\bullet 1} = (1, \dots, 1)^T \in \mathbb{R}^n$  and  $\Psi_{\bullet j} = (\Psi_{1j}, \dots, \Psi_{nj})^T \in \mathbb{R}^{n \times m}$  for  $j \ge 2$ . We further denote

$$Y = (Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})^T \in \mathbb{R}^n,$$
  

$$\boldsymbol{\beta}_+ = (\boldsymbol{\alpha}, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{1+(d-1)m},$$
  

$$\boldsymbol{\beta}_+^* = (f_1^*(z), \boldsymbol{\beta}^{*T})^T \in \mathbb{R}^{1+(d-1)m} \text{ and}$$
  

$$W_z = \operatorname{diag}(K_h(X_{11} - z), \dots, K_h(X_{n1} - z)) \in \mathbb{R}^{n \times n}.$$
 (2.12)

To unify the notation in our algorithm, we also write  $\boldsymbol{\beta}_{+} = (\boldsymbol{\beta}_{1}, \boldsymbol{\beta}_{2}^{T}, \dots, \boldsymbol{\beta}_{d}^{T})^{T}$ , where  $\boldsymbol{\beta}_{1} = \alpha$  and  $\boldsymbol{\beta} = (\boldsymbol{\beta}_{2}^{T}, \dots, \boldsymbol{\beta}_{d}^{T})^{T}$ . The tuning parameters are set as  $\lambda_{j} = \lambda \sqrt{m}$  for j = 1 and  $\lambda_{j} = \lambda$  for  $j \geq 2$ . Using the above notation, the objective function in (2.7) can be written as

$$\mathcal{L}_{z}(\boldsymbol{\beta}_{+}) + \lambda \mathcal{R}(\boldsymbol{\beta}_{+}) = \frac{1}{n} (\boldsymbol{Y} - \boldsymbol{\Psi} \boldsymbol{\beta}_{+})^{T} \mathbf{W}_{z} (\boldsymbol{Y} - \boldsymbol{\Psi} \boldsymbol{\beta}_{+}) + \lambda \mathcal{R}(\boldsymbol{\beta}_{+}).$$
(2.13)

We minimize the objective function in (2.13) using the randomized coordinate descent for composite functions (RCDC) proposed in Richtárik and Takáč (2014). Details of the procedure are given in Algorithm 1, where  $\nabla_j \mathcal{L}_z(\boldsymbol{\beta}_+) :=$  $\partial \mathcal{L}_z(\boldsymbol{\beta}_+)/\partial \boldsymbol{\beta}_j$  denotes the gradient. Suppose the result of the *t*th iteration is  $\boldsymbol{\beta}_+^{(t)}$ . In the next iteration, we randomly choose one coordinate  $j_{t+1}$  from  $\{1, \ldots, d\}$  and update the  $\boldsymbol{\beta}_{j_t}^{(t)}$ . Each update in (2.14) can be obtained in a closed form as

$$T(\boldsymbol{\beta}_{j}^{(t)}) = \mathcal{T}_{\lambda_{j}/\mu} \left( \boldsymbol{\beta}_{j}^{(t)} - \frac{1}{L} \nabla_{j} \mathcal{L}_{z} \left( \boldsymbol{\beta}_{+}^{(t)} \right) \right) - \boldsymbol{\beta}_{j}^{(t)}, \qquad (2.15)$$

where  $\mu$  is certain regularized constant and  $\mathcal{T}_{\lambda}$  is the softthresholding operator, which is defined as  $\mathcal{T}_{\lambda}(\mathbf{v}) = (\mathbf{v}/\|\mathbf{v}\|_2) \cdot \max\{0, \|\mathbf{v}\|_2 - \lambda\}$ . If we evaluate the estimator  $\widehat{\alpha}_z$  for M different z's, a naïve approach is to run Algorithm 1 for M times. The computational complexity is  $O(dm^2nM)$ . However, we propose a method to accelerate Algorithm 1 by exploiting the special structure of kernel functions. The accelerated method improves the computational complexity to  $O(dm^2(n + M))$ . Therefore, the computational complexity of our method is comparable to applying RCDC to minimize the objective function in (2.10) for SpAM estimation. More details can be found in Appendix B in the supplementary materials.

### 2.3. Confidence Band

In this section, we present a procedure for constructing a confidence band for the additive component  $f_1$  based on a de-biased estimator. A confidence band  $C_n$  is a set of confidence intervals  $C_n = \{C_n(z) = [c_L(z), c_U(z)] | z \in \mathcal{X}\}$ . For simplicity, we define the interval  $c_0(z) \pm r_0(z) := [c_0(z) - r_0(z), c_0(z) + r_0(z)]$ . We use  $f \in C_n$  to denote that f lies in the confidence band, that is,  $f(z) \in C_n(z)$  for all  $z \in \mathcal{X}$ . Our idea for constructing the confidence band extends the results developed for de-biased estimators for high-dimensional linear regression in Zhang and Zhang (2013), van de Geer et al. (2014), and Javanmard and Montanari (2014). Our setting is much more challenging as it involves constructing a band for an infinite dimensional object and we need a novel correction for  $\hat{\alpha}_z$  that reduces the bias introduced by (2.7).

We define for any  $\boldsymbol{v} = (v_1, \boldsymbol{v}_2^T, \dots, \boldsymbol{v}_m^T)^T \in \mathbb{R}^{(d-1)m+1}$ with  $v_1 \in \mathbb{R}$  and  $\boldsymbol{v}_j \in \mathbb{R}^m$  for  $j \geq 2$ , the norm  $\|\boldsymbol{v}\|_{2,\infty} = \max(|v_1|, \|\boldsymbol{v}_2\|_2, \dots, \|\boldsymbol{v}_d\|_2)$ . Consider the following convex program

$$\widehat{\boldsymbol{\theta}}_{z} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{(d-1)m+1}}{\arg\min} \boldsymbol{\theta}^{T} \widehat{\boldsymbol{\Sigma}}_{z} \boldsymbol{\theta}, \qquad \text{subject to}$$
$$\left\| \widehat{\boldsymbol{\Sigma}}_{z} \boldsymbol{\theta} - \mathbf{e}_{1} \right\|_{2,\infty} \leq \gamma, \qquad (2.16)$$

where  $\widehat{\Sigma}_z = n^{-1} \Psi \mathbf{W}_z \Psi^T$  and  $\mathbf{e}_1$  is the first canonical basis in  $\mathbb{R}^{(d-1)m+1}$ . The de-biased estimator is given as

$$\widehat{f}_{1}^{u}(z) = \widehat{\alpha}_{z} + \frac{1}{n} \widehat{\boldsymbol{\theta}}_{z}^{T} \Psi^{T} \mathbf{W}_{z} (\boldsymbol{Y} - \Psi \widehat{\boldsymbol{\beta}}_{+}).$$
(2.17)

We proceed to construct a confidence band based on this debiased estimator by considering the distribution of the process  $\sup_{z \in \mathcal{X}} \sqrt{nh}(\widehat{f}_1^u(z) - f_1(z))$ . We can approximate the distribution of the empirical process by the Gaussian multiplier process

$$\widehat{\mathbb{H}}_{n}(z) = \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^{n} \xi_{i} \cdot \frac{\widehat{\sigma}K_{h}(X_{i1}-z)\Psi_{i}^{T}\widehat{\theta}_{z}}{\widehat{\sigma}_{n}(z)}, \qquad (2.18)$$

where  $\xi_1, \ldots, \xi_n$  are independent N(0, 1) random variables, and the variance estimators are given as  $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \widehat{\alpha}_{X_i} - \sum_{j=2}^d \sum_{k=1}^m \Psi_{ij}^T \widehat{\beta}_{jk;X_i})^2$  and  $\widehat{\sigma}_n^2(z) = n^{-1} \widehat{\theta}_z^T \Psi \mathbf{W}_z^2 \Psi^T \widehat{\theta}_z$ . Let  $\widehat{c}_n(\alpha)$  be the  $(1 - \alpha)$ th quantile of  $\sup_{z \in \mathcal{X}} \widehat{\mathbb{H}}_n(z)$ . We construct the confidence band at level  $100 \times (1 - \alpha)$ %:  $\mathcal{C}^b_{n,\alpha} = \{\mathcal{C}^b_{n,\alpha}(z) \mid z \in \mathcal{X}\}$ , where

$$\mathcal{C}_{n,\alpha}^{b}(z) := [\widehat{f}_{1}^{u}(z) - \widehat{c}_{n}(\alpha)(nh)^{-1/2}\widehat{\sigma}_{n}(z), \widehat{f}_{1}^{u}(z) + \widehat{c}_{n}(\alpha)(nh)^{-1/2}\widehat{\sigma}_{n}(z)].$$
(2.19)

We will show that the confidence band is asymptotically honest in Section 3.2 by building on the framework developed in Chernozhukov, Chetverikov, and Kato (2014a, 2014b), who studied Gaussian multiplier bootstrap for approximating the distribution of the suprema of an empirical process.

### 3. Theoretical Properties

We establish the rate of convergence for the proposed estimator in Section 3.1, while the confidence band for  $f_1$  is analyzed in Section 3.2.

### 3.1. Estimation Consistency

We start with stating the required assumptions. Let  $p(x_1, ..., x_d)$ denote the joint density of  $X = (X_1, ..., X_d)$  and let  $p_j(x_j)$ denote the marginal density of  $X_j$ , for  $j \in [d]$ . Furthermore, let  $p_{jk\ell}(x_j, x_k, x_\ell)$  be the joint density of  $(X_j, X_k, X_\ell)$ ,  $p_{jk}(x_j, x_k)$  be bivariate density and  $p(x_j|x_\ell) := p_{\ell j}(x_\ell, x_j)/p_\ell(x_\ell)$ ,  $p(x_j, x_k|x_\ell) := p_{\ell jk}(x_\ell, x_j, x_k)/p_\ell(x_\ell)$  be condition densities for any  $j, k, \ell \in [d]$ .

- (A1) (Density function). The density function  $p(x_1, \ldots, x_d)$ is continuous on  $\mathcal{X}^d$ . For all  $j, k \geq 2$  and  $x_j, x_k \in \mathcal{X}$ ,  $p_{1,j,k}(\cdot, x_j, x_k) \in \mathcal{H}(2, L)$ . There exist a fixed constant  $B < \infty$  such that  $p_1(x_1) \vee p(x_j|x_1) \vee p(x_j, x_k|x_1) \leq B$  for all  $(x_1, x_j, x_k) \in \mathcal{X}^3$  and  $j, k \in \{2, \ldots, d\}$ .
- (A2) (Kernel function). The kernel K(u) is a continuous function with a bounded support satisfying

$$\int_{\mathcal{X}} K(u) du = 1 \text{ and } \int_{\mathcal{X}} u K(u) du = 0.$$

(A3) (Design matrix). Let  $\Sigma_z = \mathbb{E}[K_h(X_1 - z)\Psi_{1\bullet}\Psi_{1\bullet}^T]$ , recalling that  $\Psi_{1\bullet} = (1, \Psi_{12}^T, \dots, \Psi_{1d}^T)^T$ . For any  $J \subset [d]$ , we define a cone

$$\mathbb{C}_{\beta}^{(\kappa)}(J) = \left\{ \boldsymbol{\beta}_{+} = (\alpha, \boldsymbol{\beta}^{T})^{T} \mid \sum_{j \notin J, j \neq 1} \|\boldsymbol{\beta}_{j}\|_{2} \\ \leq \kappa \sum_{j \in J, j \neq 1} \|\boldsymbol{\beta}_{j}\|_{2} + \kappa \sqrt{m} |\alpha| \right\}.$$
 (3.1)

There exists a universal constant  $\rho_{\min} > 0$  independent to n, d, z such that the restricted minimum eigenvalue on  $\mathbb{C}_{\beta}^{(\kappa)}(J)$  satisfies

$$\inf_{z \in \mathcal{X}} \inf_{|J| \le s} \inf_{\boldsymbol{\beta}_+ \in \mathbb{C}^{(\kappa)}_{\beta}(J)} \frac{\boldsymbol{\beta}_+^T \boldsymbol{\Sigma}_z \boldsymbol{\beta}_+}{\|\boldsymbol{\beta}\|_2^2 + m\alpha^2} \ge \frac{\rho_{\min}}{m}.$$
 (3.2)

(A4) (Noise term). The error term  $\varepsilon$  satisfies  $\mathbb{E}[\varepsilon] = 0$ , is independent to X, and is a subgaussian random variable such that  $\mathbb{E}[\exp(\lambda \varepsilon)] \le \exp(\lambda^2 \sigma_{\epsilon}^2/2)$  for any  $\lambda$ .

# (A5) The nonparametric function $f(x_1, ..., x_d) \in \mathcal{K}_d(s)$ defined in Definition 2.3.

When the support  $\mathcal{X}$  is compact, Assumption (A1) is satisfied if there exist fixed constants  $0 < c < C < \infty$  such that  $p_1(x_1) \geq c$  and  $p_{1jk}(x_1, x_j, x_k) \leq C$  for all  $(x_1, x_j, x_k) \in \mathcal{X}^3$  and  $j, k \in \{2, \ldots, d\}$ . The assumption that density functions are bounded away from infinity and zero is used in many articles on additive model. For example, Huang, Horowitz, and Wei (2010) studied estimation of sparse additive models under assumption that the univariate densities  $\{p_j(x_j)\}_{j \in [d]}$  are bounded away from infinity and zero. Opsomer and Ruppert (1997) and Fan and Jiang (2005) study the additive model with two covariates:  $Y = \mu + f_1(X_1) + f_2(X_2) + \epsilon$  and impose

$$\sup_{x_1, x_2 \in \mathcal{X}} \left| \frac{p_{12}(x_1, x_2)}{p_1(x_1) p_2(x_2)} - 1 \right| < 1,$$
(3.3)

which implies that  $p_{12}(x_1, x_2)$  is bounded from infinity and zero. Since the loss function in (2.6) involves interaction terms of multiple variables, Assumption (A1) imposes boundedness on the density related to three covariates. Moreover, Assumption (A1) can be satisfied even if the support  $\mathcal{X}$  is non-compact. In comparison, whenever a density function is bounded away from zero, it has to have bounded support. Therefore, our assumption is more general than those used in Opsomer and Ruppert (1997), Fan and Jiang (2005), and Huang, Horowitz, and Wei (2010). As another example, consider  $X \sim N(\mathbf{0}, \mathbf{\Sigma})$  where  $\mathbf{\Sigma}_{jj} =$ 1 and  $\mathbf{\Sigma}_{jk} = \rho$  for all  $2 \le j < k \le d$ , we can bound the densities as  $p_1(x_1) \le 1$ ,  $p(x_j|x_1) \le (1 - \rho^2)^{-1/2}$  and  $p(x_j, x_k|x_1) \le$  $(1 - \rho)^{-1}(1 + 2\rho)^{-1/2}$  for all  $2 \le j < k \le d$ . Therefore, Assumption (A1) is satisfied in this example as long as  $\mathbf{\Sigma}$  is positive definite.

Assumption (A2) is standard in the literature on local linear regression (Fan 1993), while Assumption (A4) is standard in the literature on sparse additive modeling (Meier, van de Geer, and Bühlmann 2009; Huang, Horowitz, and Wei 2010; Koltchinskii and Yuan 2010; Kato 2012; Raskutti, Wainwright, and Yu 2012).

Assumption (A3) is similar to the restricted strong convexity condition in Negahban et al. (2012). Note that  $\Sigma_z$  is the expectation of the Hessian matrix of the loss function  $\mathcal{L}(\boldsymbol{\beta}_+)$ . We require  $\Sigma_z$  to be positive definite when restricted to vectors in the cone  $\mathbb{C}_{\beta}^{(\kappa)}(J)$ . Again, the additional factor  $\sqrt{m}$  in front of  $|\alpha|$  makes sure that  $\alpha$  and  $\boldsymbol{\beta}_z$  are calibrated on the same scale (see Remark 2.1). Assumption (A3) can be derived from the assumption on the design in Koltchinskii and Yuan (2010). They considered the quantity

$$\beta_{2,\kappa}(J) = \inf \left\{ \beta > 0 \, \Big| \, \sum_{j \in J} \|h_j\|_{L^2(\mathbb{P})}^2 \le \beta^2 \, \|\sum_{j=1}^d h_j\|_{L^2(\mathbb{P})}^2, \\ (h_1, \dots, h_d) \in \mathbb{C}_h^{(\kappa)}(J, \mathbb{P}) \right\},$$
(3.4)

where  $\mathbb{C}_{h}^{(\kappa)}(J,\mathbb{P}) = \{(h_1,\ldots,h_d) \mid \sum_{j \notin J} \|h_j\|_{L^2(\mathbb{P})} \le \kappa \sum_{j \in J} \|h_j\|_{L^2(\mathbb{P})} \}$  for  $J \subset [d]$ .

Let  $\mu_z$  be the measure defined as  $\int g d\mu_z = \mathbb{E}[g(\mathbf{X})|X_1 = z]$  for any *g*. The following proposition describes the connection between  $\beta_{2,\kappa}(J)$  and Assumption (A3).

*Proposition 1.* We define a uniform quantity based on the constant (3.4) as

$$\bar{\beta}_{2,\kappa} = \sup_{|J| \le s} \inf \left\{ \beta > 0 \, \Big| \, \sum_{j \in J} \|h_j\|_{L^2(\mu_z)}^2 \le \beta^2 \, \Big\| \, \sum_{j=2}^d h_j \, \Big\|_{L^2(\mu_z)}^2, \\ (h_1, \dots, h_d) \in \mathbb{C}_h^{(\kappa)}(J, \mu_z), z \in \mathcal{X} \right\}.$$
(3.5)

Under Assumption (A1), there exist constants c, C > 0 such that for any subset  $\mathcal{X}' \subseteq \mathcal{X}$ ,

$$\inf_{z \in \mathcal{X}'} \inf_{|J| \le s} \inf_{\boldsymbol{\beta}_+ \in \mathbb{C}_{\beta}^{(\kappa)}(J)} \frac{\boldsymbol{\beta}_+^T \boldsymbol{\Sigma}_z \boldsymbol{\beta}_+}{\|\boldsymbol{\beta}\|_2^2 + m\alpha^2} \ge \inf_{z \in \mathcal{X}'} p_1(z) \cdot \frac{C\bar{\beta}_{2,c\kappa}^{-2}}{s(c\kappa+1)^2} \frac{1}{m}.$$
(3.6)

The proof of Proposition 3.1 is stated in Appendix D in the supplementary materials.

When the support  $\mathcal{X}$  is compact and we assume there exists a fixed constant b > 0 such that  $p_1(x_1) \ge b$  for all  $x_1 \in \mathcal{X}$ , Proposition 3.1 implies that if the number of active components *s* is finite, we can choose  $\rho_{\min} = Cb\bar{\beta}_{2,c\kappa}^{-2}/(s(c\kappa + 1)^2)$  and Assumption (A3) is satisfied if  $\bar{\beta}_{2,c\kappa} < \infty$ . The assumption that *s* is finite is required in the previous works (Meier, van de Geer, and Bühlmann 2009; Huang, Horowitz, and Wei 2010; Koltchinskii and Yuan 2010; Kato 2012). However, when the support  $\mathcal{X}$  is unbounded,  $p_1(z) \to 0$  as  $|z| \to \infty$ . In this case, (3.6) does not provide a valid  $\rho_{\min}$  satisfying Assumption (A2). We will discuss such a case in Section 3.3.

In the following, we present the rate of convergence of the kernel-sieve hybrid regression estimator.

*Theorem 3.1.* Suppose that Assumptions (A1)–(A5) are satisfied. If  $h = o(1), m \to \infty$  as  $n \to \infty$ , and we set

$$\lambda = C \left( \sqrt{\frac{\log(dmh^{-1})}{nh}} + \frac{\sqrt{s}}{m^{5/2}} + \frac{m^{3/2}\log(dh^{-1})}{n} + \frac{h^2}{\sqrt{m}} \right),$$
(3.7)

for a sufficiently large constant *C*, the estimator  $(\widehat{\alpha}_z, \widehat{\beta}_z^T)^T$  defined in (2.7) satisfies

$$\sup_{z \in \mathcal{X}} \sum_{j=2}^{d} ||\widehat{\boldsymbol{\beta}}_{j;z} - \boldsymbol{\beta}_{j}^{*}||_{2} \le \frac{sm}{\rho_{\min}}\lambda \quad \text{and}$$
$$\sup_{z \in \mathcal{X}} |\widehat{a}_{z} - f_{1}(z)| \le \frac{s\sqrt{m}}{\rho_{\min}}\lambda \quad (3.8)$$

with probability 1 - c/n for some constant c > 0, where  $\widehat{\beta}_{j;z}$  is a subvector of  $\widehat{\beta}_z$  corresponding to the coefficients of B-spline basis of the *j*th covariate and same for  $\beta_j^*$  to  $\beta^*$  defined in (2.4). Furthermore, the estimator  $\widehat{f}$  in (2.9) satisfies

$$\|\widehat{f} - f\|_2 \le \rho_{\min}^{-1} s \sqrt{m\lambda} \tag{3.9}$$

with probability 1 - c/n.

The estimation error comes from four sources. The noise  $\varepsilon$  contributes  $O\left(\sqrt{\log(dmh^{-1})/nh}\right)$  in (3.7). The second-term in (3.7),  $O\left(\sqrt{sm^{-5/2}}\right)$ , comes from the approximation error

introduced by using *m* B-spline basis functions to estimate the true functions  $\{f_j\}_{j=2}^d$ . The third source of error comes from the kernel method, which uses a constant to estimate  $f_{1z}$  locally. The fourth source of error comes from searching for correct local approximation by *s* additive functions due to (4.1). Both the third and fourth sources contribute  $O\left(n^{-1}m^{3/2}\log(dh^{-1}) + h^2/\sqrt{m}\right)$  to the estimation error. The detailed proof of Theorem 3.1 is shown in Appendix A in the supplementary materials.

When  $\rho_{\min}^{-1} = O(1)$  and s = O(1), the statistical rate in (3.9) is minimized when we choose  $h \asymp n^{-1/6}$ ,  $m \asymp$  $n^{1/6}$  and  $\lambda \simeq n^{-5/12} \sqrt{\log(dn)}$ . With these choices, we obtain  $\|\widehat{f} - f\|_2^2 = O_P(n^{-2/3}\log(dn))$ . This convergence rate is slower than the optimal rate  $O_P(n^{-4/5} + \log d/n)$  for estimating the sparse additive model (Raskutti, Wainwright, and Yu 2012). However, we will show that this rate is enough to construct an honest confidence band for  $f_1$  in Section 3.2. Besides, our kernel-sieve hybrid estimator can be applied to functions beyond the sparse additive model. It can actually estimate the functions in the form  $f_1(x_1) + \sum_{j=2}^d f_j(x_j, x_1)$ , which has two-dimensional additive functions. We refer Section 4 for the details of the generalization. In fact, the rate  $\|\widehat{f} - f\|_2^2 = O_P(n^{-2/3}\log(dn))$  we achieve is nearly optimal up to logarithmic factors for the two-dimensional Hölder class (Stone 1980). Technically, the slower rate comes from the error term  $T_n = \sup_{z \in \mathcal{X}} \max_{j \in [d]} \frac{1}{n} \| \Psi_{oj}^T \mathbf{W}_z \boldsymbol{\varepsilon} \|_2$  $O_P\left(\sqrt{\log(dn)/(nh)}\right)$ , where  $\mathbf{W}_z$  is defined in (2.2). In comparison, Huang, Horowitz, and Wei (2010) only need to bound  $T'_n = \sup_{z \in \mathcal{X}} \max_{j \in [d]} \frac{1}{n} \| \Psi^T_{\bullet j} \varepsilon \|_2 = O_P\left(\sqrt{\log(dn)/n}\right)$  (see their Lemma 2). Note that  $T_n = O_P(h^{-1/2}T'_n)$  because the kernel matrix  $\mathbf{W}_z$  increases its variance by  $O_P(h^{-1/2})$ . Detailed technical analysis of  $T_n$  is given in Lemma A.4 in the supplementary materials.

### 3.2. Theoretical Results for Confidence Band

To establish valid theoretical results on the confidence band  $C_{n,\alpha}^b$ , we need to strengthen the weak dependency assumption in Assumption (A3) as follows.

Assumption (A6) (Nonparametric weak dependency). Recall that the constant *B* is defined in Assumption (A1) and  $\rho_{\min}$  is defined in (3.2). We assume that the density functions of *X* satisfies

$$\sum_{j=2}^{d} \|p_{1,j} - p_1 p_j\|_2 \le \frac{\rho_{\min}}{2B} \quad \text{and}$$
$$\sup_{k\ge 2} \sum_{j< k} \|p_{1,j,k} - p_1 p_j p_k\|_2 \le \frac{\rho_{\min}}{2B}.$$
(3.10)

The nonparametric weak dependency assumption quantifies how strong the dependency between the covariates can be, while still allowing us to construct an honest confidence band. In particular, Assumption (A6) allows us to ensure validity of the orthogonality property proposed by Chernozhukov, Hansen, and Spindler (2015) and its equivalent characterization in Zhang and Zhang (2013) and van de Geer et al. (2014). Heuristically, for *M*-estimators, the orthogonality property essentially requires the inverse of the Hessian matrix of the population loss function to have sparse columns (Ning and Liu 2017). For linear models, Javanmard and Montanari (2014) relaxed the sparsity assumption by requiring the inverse of the Hessian matrix to have columns with bounded  $\ell_1$ -norms. We extend the approach of Javanmard and Montanari (2014) to our nonparametric setting here. For our loss function in (2.6), the population Hessian matrix is  $\Sigma_z$  defined in Assumption (A3). Due to the complicated definition of  $\Sigma_z$ , there is no straightforward interpretation of  $\Sigma_z^{-1}$  and any assumption imposed on  $\Sigma_z^{-1}$ would imply restrictions on the data generating process that are hard to verify. In comparison, the nonparametric weak dependency assumption in (3.10) is straightforward and easy to check in practice. Furthermore, Assumption (A6) is a highdimensional analogue of the assumption in (3.3), which was considered by Opsomer and Ruppert (1997) and Fan and Jiang (2005) for fixed dimensional additive models. Since  $||p_{1,i,k} - p_1 p_i p_k||_2$  measures the dependency among  $X_1, X_i$  and  $X_k$ , (3.10) requires the  $\ell_1$ -norms of both  $\{\|p_{1,j,k} - p_1 p_j p_k\|_2\}_{j \ge 2}$ and  $\{\|p_{1,j} - p_1 p_j\|_2\}_{j \ge 2}$  to be bounded.

The following proposition shows Assumption (A6) can be satisfied even if  $\mathcal{X}$  is unbounded and  $X_j$ 's are dependent. We refer the detailed construction of such an example to Appendix G in the supplementary materials.

**Proposition 2.** Given any  $\rho \in (0, 1/2)$  satisfying  $\rho \leq \rho_{\min}/(18B)$ , there exists a *d*-dimensional density function  $p(\mathbf{x})$  with unbounded support such that

$$\sum_{j=2}^{d} \operatorname{cov}(X_1, X_j) \ge \rho/9 \quad \text{and} \quad \operatorname{cov}(X_j, X_k) \ge \rho/9$$
  
for all  $j, k > 2$  and  $|j - k| < 1$ ,

and Assumption (A6) is satisfied.

The next lemma provides guidance to the selection of the tuning parameter  $\gamma$  in (2.16).

*Lemma 3.1.* Suppose that Assumptions (A1), (A3), and (A6) hold. Let

$$\gamma = C \log(|\mathcal{X}|/\rho_{\min}) \cdot \log d\sqrt{m/nh}, \qquad (3.11)$$

for sufficiently large constant *C*. Then the vector  $\boldsymbol{\theta}_z = \boldsymbol{\Sigma}_z^{-1} \mathbf{e}_1$  is a feasible solution to the optimization program in (2.16) with high probability. In particular, we have

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}\|\widehat{\boldsymbol{\Sigma}}_{z}\boldsymbol{\theta}_{z}-\mathbf{e}_{1}\|_{2,\infty}\leq\gamma\right)\geq1-c/d$$

for some constant c.

We defer the proof of this lemma to Appendix F.2 in the supplementary materials. We are now ready to present the main theorem of this section which establishes a valid confidence band for a component in the sparse additive model under the identifiability condition (2.2).

*Theorem 3.2.* We consider the SpAM model in (2.1) with identifiability condition (2.2). Suppose  $\varepsilon \sim N(0, \sigma^2)$  and Assumptions (A1)–(A6) hold. Suppose the support  $\mathcal{X}$  is bounded and inf<sub>z</sub>  $p_1(z) > 0$ . If s = O(1),  $m \simeq n^p$  for  $p \in (1/5, 3/13)$ ,  $h \simeq n^{-\delta}$  for  $\delta \in (5p - 1, (1 - 3p)/2)$ ,  $\lambda$  satisfies (3.7) and  $\gamma = C \log(dn) \sqrt{m/nh}$  for sufficiently large *C*, there exist constants  $c, C_1 > 0$  such that for any  $\alpha \in (0, 1)$ , the covering probability of  $C_{n,\alpha}^b$  in (2.19) is

$$\mathbb{P}(f_1(z) \in \mathcal{C}^b_{n,\alpha}(z), \text{ for all } z \in \mathcal{X}) \ge 1 - \alpha - C_1 n^{-c}.$$
(3.12)

In particular, the confidence band  $C_{n,\alpha}^b$  is asymptotically honest, that is,

$$\liminf_{n\to\infty} \mathbb{P}(f_1(z) \in \mathcal{C}^b_{n,\alpha}(z), \text{ for all } z \in \mathcal{X}) \ge 1 - \alpha.$$

Theorem 3.7 provides a more general result for the setting where  $\mathcal{X}$  is unbounded. Notice that we can no longer choose  $h \approx n^{-1/6}$  and  $m \approx n^{1/6}$  as in Theorem 3.1, since we need to underregularize our estimator to make the bias terms ignorable.

### 3.3. Results for Unbounded Support

Here we discuss the theoretical results of our method when the support  $\mathcal{X}$  is unbounded. Before discussing the technical details, we first provide some heuristic intuition why the case of unbounded support is challenging and many articles on additive model (Opsomer and Ruppert 1997; Fan and Jiang 2005; Huang, Horowitz, and Wei 2010) impose bounded support condition, as well as many article that study uniform convergence of kerneltype estimators (Peligrad 1992; Masry 1996; Nze and Doukhan 2004; Fan and Yao 2008). For example, consider the univariate nonparametric model that *n* iid samples  $\{X_i, Y_i\}_{i=1}^n$  are generated from  $Y_i = m(X_i) + \epsilon_i$  where  $m \in \mathcal{H}(2, L)$  and  $\mathbb{E}[\epsilon_i] = 0$ . The Nadaraya–Watson estimator for m(z) is

$$\widehat{m}(z) = \arg\min_{a} \sum_{i=1}^{n} K_{h}(X_{i} - z)(Y_{i} - a)^{2}$$
$$= \frac{\sum_{i=1}^{n} Y_{i}K_{h}(X_{i} - z)}{\sum_{i=1}^{n} K_{h}(X_{i} - z)}.$$
(3.13)

The pointwise mean square error of  $\widehat{m}(z)$  is given by Fan and Gijbels (1996) as

$$MSE(\widehat{m}(z)) := \mathbb{E}(\widehat{m}(z) - m(z))^2 \qquad (3.14)$$
$$\approx \frac{h^4 \sigma_K^4}{4} \left( m''(z) + 2m'(z) \frac{p'(z)}{p(z)} \right)^2 + \frac{R_K \sigma_\epsilon^2}{nhp(z)},$$

where  $\sigma_K^2 = \int u^2 K(u) du$ ,  $R_K = \int K^2(u) du$ ,  $\sigma_{\epsilon}^2 = \mathbb{E}[\epsilon_i^2]$ , p(z) is the density of  $X_i$ , and " $\approx$ " means we neglect higher order terms. From (3.14), we could see that  $MSE(\widehat{m}(z))$  will diverge if  $p(z) \to 0$ . The intuition is we have a kernel density estimator of p(z) in the denominator of  $\widehat{m}(z)$  in (3.13). Therefore, to control the uniform rate  $\sup_{z \in \mathcal{X}} |\widehat{m}(z) - m(z)|$ , many analyses assume  $\inf_{z \in \mathcal{X}} p(z) > 0$ , which is impossible if  $\mathcal{X}$  is unbounded.

When  $\mathcal{X}$  is unbounded, because of the argument above, the uniform rate of  $\widehat{m}$  are typically established on the compact subset of  $\mathcal{X}$ . For example, for  $\mathcal{X} = \mathbb{R}$ , Hansen (2008) proved that under certain regularity conditions, if  $b_n = \inf_{|z| \le D_n} p_1(z) > 0$ , then  $\sup_{|z| \le D_n} |\widehat{m}(z) - m(z)| = O_P(b_n^{-1}(h^2 + \sqrt{\log n/nh}))$ . We also show the uniform rate of our estimator on compact subsets of  $\mathcal{X} = \mathbb{R}$  in the following corollary. Corollary 3.1. Suppose Assumptions (A1), (A2), (A4), (A5) hold and  $\bar{\beta}_{2,\kappa}$  in (3.5) is finite. We consider  $s = O(1), h \simeq n^{-1/6}$ ,  $m \simeq n^{1/6}$  and  $\lambda = Cn^{-5/12}\sqrt{\log(dn)}$  for sufficiently large constant C. Given any compact interval  $[-D_n, D_n]$ , if  $b_n = \inf_{|z| < D_n} p_1(z) > 0$ , we have

$$\sup_{|z| \le D_n} |\widehat{a}_z - f_1(z)| = O_P(b_n^{-1} \log(dn) / n^{2/3}).$$
(3.15)

The corollary can be proved by applying Proposition 3.1 to Theorem 3.1. Assumption (A3) is not required here because  $\rho_{\min}$  could be zero when  $\mathcal{X} = \mathbb{R}$ . Proposition 3.1 characterizes how  $\rho_{\min}$  depends on  $p_1(z)$  when we choose  $\mathcal{X}' = [-D_n, D_n]$ in (3.6) and helps us obtain an explicit rate in (3.15).

When the support is bounded but  $p_1(z)$  goes to zero, Corollary 3.1 can also give us the uniform rate. Without loss of generality, let  $\mathcal{X} = [-a, a]$  for a > 0. If there exist  $C, \beta > 0$ such that  $p_1(z) \ge C||z| - a|^{\beta}$  for all  $z \in [-a, a]$ , under the assumptions of Corollary 3.1, we have

$$\sup_{z\in[-a+\delta_n,a-\delta_n]}|\widehat{a}_z-f_1(z)|=O_P\big(\delta_n^{-\beta}\log(dn)/n^{2/3}\big).$$

We can also show the coverage probability of the confidence band  $C^b_{n,\alpha}$  in (2.19) on a compact subset of  $\mathbb{R}$  in the following theorem.

Theorem 3.3. We consider the SpAM model in (2.1) with identifiability condition (2.2). Suppose  $\varepsilon \sim N(0, \sigma^2)$  and Assumptions (A1), (A2), (A4)–(A6) hold and  $\bar{\beta}_{2,\kappa}$  in (3.5) is finite. We assume there exists some  $\alpha > 1$  such that  $n^{-\alpha}(\inf_{|z| \le n} p_1(z))^{-1} = O(1)$ . Given any  $D_n = O(n^{\beta})$  for  $\beta < 1/(10\alpha \lor 5)$ , if s = O(1),  $m \asymp n^p$  for  $p \in (1/5, (3-2\beta)/13)$ ,  $h \asymp n^{-\delta}$  for  $\delta \in (5p-1, (1-3p)/2 - \beta)$ ,  $\lambda$  satisfies (3.7) and  $\gamma = C\log(dn)\sqrt{m/nh}$  for sufficiently large *C*, there exist constants c, C > 0 such that for any  $\alpha \in (0, 1)$ , the covering probability of  $\mathcal{C}^b_{n,\alpha}$  has

$$\mathbb{P}(f_1(z) \in \mathcal{C}^b_{n,\alpha}(z), \text{ for all } |z| \le D_n) \ge 1 - \alpha - Cn^{-c}.$$
(3.16)

In particular, the confidence band  $C_{n,\alpha}^b$  is asymptotically honest on any interval  $[-D_n, D_n]$  with  $D_n = O(n^\beta)$  for  $\beta < 1/(10\alpha \lor$ 5), that is,

$$\liminf_{n\to\infty} \mathbb{P}(f_1(z) \in \mathcal{C}^b_{n,\alpha}(z), \text{ for all } |z| \le D_n) \ge 1 - \alpha.$$

For the detailed proof of this theorem, see Appendix C.1 in the supplementary materials.

### 4. Generalization to Larger Nonparametric Family

In this section, we will show that our kernel-sieve estimator defined in (2.7) can be applied to a family of functions larger than the sparse additive model. We call this new function family as the additive local approximation model with sparsity (ATLAS). Notice that under the SpAM model, there are no interaction terms between different covariates. In addition, the set of covariates in S affect the response *Y* globally. The ATLAS model relaxes these two structural constraints.

**Definition 4.1.** A *d*-dimensional function  $f(x_1, \ldots, x_d)$  has a local sparse additive approximation for  $x_1$  if for any  $z \in \mathcal{X}$ , there exist functions  $f_{1z}(\cdot), \ldots, f_{dz}(\cdot) \in \mathcal{H}(2, L)$ , two bounded functions  $L(\cdot) : \mathcal{X}^d \mapsto \mathbb{R}$ ,  $Q(\cdot) : \mathcal{X} \mapsto \mathbb{R}$  and a constant  $\delta_0 > 0$  such that for any  $\mathbf{x}_{-1} = (x_2, \ldots, x_d)^T \in \mathcal{X}^{d-1}$ , if  $x_1 \in (z - \delta_0, z + \delta_0)$ , we have the approximation

$$\left| f(x_1, \dots, x_d) - f_1(z) - \sum_{j=1}^d f_{jz}(x_j) - L(z, \mathbf{x}_{-1})(x_1 - z) \right| \le Q(z)(x_1 - z)^2.$$
(4.1)

Furthermore, we assume that the locally additive approximation functions are sparse in that at most *s* of the functions  $\{f_{jz}(\cdot)\}_{j=1}^d$  are not identical to zero. The sparsity pattern at each  $z \in \mathcal{X}$  is denoted as  $S_z = \{j \in [d] : f_{jz}(\cdot) \neq 0\}$ . We call the function class containing functions satisfying Definition 4.1 the ATLAS model and denote it as  $\mathcal{A}_d(s)$ .

By letting  $z \rightarrow x_1$  in (4.1), we observe that a function in the ATLAS model can be written as

$$f(x_1, \dots, x_d) = f_1(x_1) + \sum_{j=2}^d f_j(x_j, x_1),$$
(4.2)

where  $\{f_j(x_j, x_1)\}_{j=2}^d$  are *d* bivariate functions belonging to  $\mathcal{H}(2, L)$ . Similar to (2.2), we impose the identifiability condition

$$\mathbb{E}[f_1(X_1)] = 0 \quad \text{and} \quad \mathbb{E}[f_j(X_j, x_1)] = 0 \quad \text{for any}$$
  
$$x_1 \in \mathcal{X} \quad \text{and} \quad j = 2, \dots, d. \tag{4.3}$$

We call  $X_1$  the longitude variable and the functions  $f_2(\cdot, z), \ldots, f_d(\cdot, z)$  for each  $z \in \mathcal{X}$  as charts at longitude z. Notice that the sparsity patterns of charts may change with  $z \in \mathcal{X}$ , allowing for more flexible modeling compared to SpAM which assumes a fixed sparsity pattern. Therefore, ATLAS allows complex nonlinear interaction between  $X_1$  and other covariates. A visualization of a *d*-dimensional function in ATLAS is illustrated in Figure 1.

It is obvious that the sparse additive model is a subset of ATLAS with the fixed charts  $\{f_j\}_{j=1}^d$  which are invariant to any longitude variable. In fact, ATLAS model generalizes many existing nonparametric models in the literature. Functions like (4.2) are studied under the framework of time-varying additive models for longitudinal data (Zhang and Wang 2015) when the



**Figure 1.** The illustration of ATLAS. As the longitude variable  $X_1$  changes as  $X_1 \in \{z_1, z_2, z_3, z_4\}$ , the sparsity patterns of the charts are different. By fixing the lattitude variable  $X_j$  for j = 2, ..., 5, the values of charts  $f_j(\cdot, z)$  change with z. Under the sparsity assumption,  $f_j(\cdot, z)$  is zero for most of the range of z.

dimension is fixed. It has also been considered as compound functional model proposed in Dalalyan, Ingster, and Tsybakov (2014) under the high-dimensional setting. However, ATLAS allows the sparsity pattern to vary with the longitude covariate  $x_1$  while the compound functional model in Dalalyan, Ingster, and Tsybakov (2014) must have fixed support. The following example gives another subset of ATLAS model.

*Example 4.1.* Consider a *d*-dimensional function with the structure

$$f(x_1, \dots, x_d) = f_1(x_1) + \sum_{j=2}^d a_j(x_1) f_j(x_j), \qquad (4.4)$$

where  $a_j(\cdot), f_k(\cdot) \in \mathcal{H}(2, L)$  for all  $k \in [d]$  and  $j \ge 2$ . Moreover, for any fixed  $z \in \mathcal{X}$ , at most *s* of  $\{a_j(z)\}_{j\ge 2}$  are nonzero. The function in (4.4) satisfies Definition 4.1. We define  $f_j(x_j, x_1) =$  $a_j(x_1)f_j(x_j)$  for j = 2, ..., d and let  $L(z, \mathbf{x}_{-1}) = \sum_{j\ge 2} a'_j(z)f_j(x_j)$ . Then for any  $x_1 \in (z - \delta_0, z + \delta_0)$  and  $\mathbf{x}_{-1} \in \mathcal{X}^{d-1}$ , we have

$$\left| f(x_1, \dots, x_d) - \sum_{j=1}^d f_j(x_j, z) - L(z, \mathbf{x}_{-1})(x_1 - z) \right|$$
  
$$\leq s \max_{j \in [d]} \|f_j\|_{\infty} \|a_j''\|_{\infty} (x_1 - z)^2 := Q(z)(x_1 - z)^2,$$

which satisfies Definition 4.1 if *s* is finite. The nonparametric function in (4.4) allows nontrivial interactions between  $X_1$  and  $X_j$  for  $j \ge 2$ , which cannot be modeled with SpAM. The sparsity of the function in (4.4) originates from  $a_j(x_1)$  and there is no sparsity assumption on  $f_j(x_j)$ .

Example 4.1 shows that the ATLAS model is a generalization of the varying coefficient additive model for functional data (Zhang and Wang 2015). If  $f_j(x_j)$ 's are linear functions for all  $j \ge 2$ , we can write (4.4) as

$$f(x_1, \dots, x_d) = f_1(x_1) + \sum_{j=2}^d a_j(x_1)x_j,$$
 (4.5)

which is a high-dimensional varying coefficient linear model, where the support of the linear coefficients may vary with  $x_1$ . Varying coefficient linear models in fixed dimension have been extensively studied Hastie and Tibshirani (1993), Fan and Zhang (1999), Berhane and Tibshirani (1998), and Zhu, Li, and Kong (2012), while Wei, Huang, and Li (2011) study highdimensional varying coefficient linear models with fixed sparsity.

The locally additive assumption in (4.1) for the ATLAS model makes it possible for us to use the kernel-sieve hybrid estimator to estimate functions in  $\mathcal{A}_d(s)$ . The loss function for the kernel-sieve hybrid estimator in (2.6) has two parts: the kernel function makes the loss function only involve data points within the area  $(z-h, z+h) \times \mathcal{X}^{d-1}$  and the sieve approximation part is therefore good enough to approximate the true function according to (4.1). In particular, let  $(\widehat{\alpha}_z, \widehat{\beta}_z)$  be the output of (2.7), we estimate the true functions  $f_1(z)$  and  $f_i(x_i, z)$  by

$$\widehat{f}_1(z) = \widehat{\alpha}_z \text{ and } \widehat{f}_j(x_j, z) = \sum_{k=1}^m \psi_{jk}(x_j) \widehat{\beta}_{jk;z}, \text{ for } j = 2, \dots, d.$$

We can thus estimate the bivariate charts  $\{f_j(x_j, x_1)\}_{j=2}^d$  by "gluing" the local charts  $\{f_j(x_j, z)\}_{j=1}^d$  over different longitudes  $z \in \mathcal{X}$  through a fast algorithm proposed in Appendix B in the supplementary materials. Moreover, we can also construct a confidence band for  $f_1$  following the procedure in Section 3.2.

If we weaken Assumption (A5) and generalize it to the assumption that  $f(x_1, \ldots, x_d) \in \mathcal{A}_d(s)$ , the estimation rates in Theorem 3.1 and the property of confidence band in Theorem 3.3 remain true. In fact, we will prove these theorems under the ATLAS model and apply them to SpAM.

### 5. Numerical Experiments

In this section, we study the finite sample properties of confidence bands for the ATLAS model and sparse additive model. We apply the SpAM to a genomic dataset and the ATLAS model to a fMRI dataset.

# 5.1. Synthetic Data

We consider two kinds of synthetic models. In the first example, we evaluate the empirical properties of the bootstrap confidence band for sparse additive model. In the second example, we apply it to the ATLAS model.

In both examples, we use the quadratic kernel  $K_{\text{quad}}(u) = (15/16) \cdot (1 - u^2)^2 \mathbb{1}(|u| < 1)$  as the kernel function in (2.7).

*Example 5.1.* We consider the sparse additive model  $Y_i = \sum_{i=1}^{4} f_i(X_{ij}) + \varepsilon_i$ , where

$$f_1(t) = 6(0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3(\sin(2\pi t))^2 + 0.4(\cos(2\pi t))^3 + 0.5(\sin(2\pi t))^3),$$
  

$$f_2(t) = 3(2t - 1)^2, \quad f_3(t) = 5t,$$
  

$$f_4(t) = 4\sin(2\pi t)/(2 - \sin(2\pi t)).$$

The model is considered by Zhang and Lin (2006), Meier, van de Geer, and Bühlmann (2009), and Huang, Horowitz, and Wei (2010). Let  $W_1, \ldots, W_d$  and U follow iid Uniform[0, 1] and

$$X_j = \frac{W_j + tU}{1+t} \quad \text{for } j = 1, \dots, d.$$

The data sample  $X_{1j}, \ldots, X_{nj}$  are iid copies of  $X_j$ . The correlation between  $X_j, X_{j'}$  is therefore  $t^2/(1 + t^2)$  for  $j \neq j'$ . We set t = 0.3. The noise  $\{\varepsilon_i\}_{i=1}^n$  are iid  $N(0, 1.5^2)$ . Let the dimension d = 600and the sample sizes  $n \in \{400, 500, 600\}$ . In the kernel-sieve hybrid estimator (2.7), we use the cubic B-splines with nine evenly distributed knots and m = 5. The parameter  $\gamma$  in (2.16) is set to be  $\gamma = 0.05 \log d\sqrt{m/nh}$ . The tuning parameter  $\lambda$  and bandwidth *h* are chosen by cross-validation according to the BIC criterion defined as

$$BIC = \log\left(\frac{RSS}{nh}\right) + df \cdot \frac{\log nh}{nh},$$

where RSS is the residual sums of squares and the degrees of freedom is defined as df =  $\hat{s} \cdot m$  with  $\hat{s}$  being the number of variables selected by the estimator. We aim to construct the

confidence band for  $f_1^*(t) = f_1(t) - \mathbb{E}[f_1(X_1)]$ . In the simulation, we use the sample mean  $\mathbb{E}_n[f_1(X_1)] := n^{-1} \sum_{i=1}^n f(X_{i1})$  to center  $f_1(t)$ .

To test the coverage probability of confidence bands for inactive covariates, we also construct the confidence band for  $f_5(t) = 0$ . We set the significance level at 95%. We compute the empirical coverage probability via the percentage that the confidence band covers the truth on all the 500 grid points on [0, 1] in 500 repetitions. We compare the performance of our method on Example 5.1 with the oracle method in Kozbur (2015), which assumes that the nonzero functions are known beforehand. Since Kozbur (2015) did not provide a straightforward construction of the confidence band, we construct the confidence intervals for  $f_1(x)$  with all x's on the 500 grid points on [0, 1]. The significance levels of these confidence intervals are adjusted via Bonferroni correction (Efron 2012) to be fairly compared with our method.

The results are summarized in Figure 2 and Table 1. In Table 1, the "area" of the confidence band  $C_{n,\alpha}^b$  is defined as  $\int_{z \in \mathcal{X}} 2\hat{c}_n(\alpha) (nh)^{-1/2} \hat{\sigma}_n(z) dz$ . In the simulation, we calculate the integration via discretizing the interval into grids and averaging the results across 500 repetitions. We can see in Table 1 that the coverage probability of the oracle method in Kozbur (2015) is close to 1, however, the area of the confidence band is much larger comparing to our method. This is because Bonferroni correction is used for the confidence band of Kozbur (2015). This makes the confidence band too conservative and not nominal in coverage probability.

*Example 5.2.* We generate data from the following ATLAS model

$$Y_i = a_1 f_1(X_{i1}) + \sum_{j=2}^4 a_j(X_{i1}) f_j(X_{ij}) + \varepsilon_i,$$

where the additive functions are designed as follows

$$f_1(t) = -2\sin(2\pi t), \quad f_2(t) = t^2 - 1/3,$$
  

$$f_3(t) = t - 1/2, \quad f_4(t) = e^t + e^{-1} - 1;$$
  

$$a_1 \in \{0, 1\}, \quad a_2(t) = 2K_{\text{quad}}(4t - 1),$$
  

$$a_2(t) = 3\cos(2\pi t), \quad a_3(t) = 4.$$

Here, two values of  $a_1 \in \{0, 1\}$  correspond to two scenarios that the true function is zero and nonzero. The noise  $\epsilon_i \sim$  $N(0,\sigma^2)$  for i = 1, ..., n with  $\sigma = 1.5$ . This ATLAS model is constructed based on the synthetic example in Ravikumar et al. (2009) by adding  $a_i(t)$ 's according to Example 4.1. The covariates  $X_{ij}$  are independently and identically generated from Uniform[0, 1] distributions for i = 1, ..., n and j = 1, ..., d. It can be checked that this model follows the identifiability condition in (4.3). According to the argument in Example 4.1, the true function  $f_1^*(t) = a_1 f_1(t)$ . We set the dimension of covariates to be d = 600 and consider three sample sizes  $n \in$ {400, 500, 600}. We again use the cubic B-spline basis with nine evenly distributed knots and m = 5. We again tune  $\lambda$  and hthrough cross-validation by minimizing the BIC criterion. The confidence bands are constructed at the significance level 95% and the quantile estimator  $\hat{c}_n(\alpha)$  is computed by bootstrap with



**Figure 2.** Kernel-sieve hybrid estimators for the d = 600 dimensional SpAM model  $Y = \sum_{j=1}^{4} f_j(X_j) + \varepsilon$ , for n = 400, 500, 600 and the noise  $\varepsilon \sim N(0, 1.5^2)$ . The confidence bands at significant level 95 cover  $f_1(t)$  on the first row and  $f_5(t) = 0$  on the second row.



**Figure 3.** Kernel-sieve hybrid estimators for the d = 600 dimensional ATLAS model  $Y = a_1f_1(X_1) + \sum_{j=2}^4 a_j(X_1)f_j(X_j) + \varepsilon$ , for n = 400, 500, 600 and the noise  $\varepsilon \sim N(0, 1.5^2)$ . The confidence bands at significant level 95 cover  $f_1^* = a_1f_1$  for  $a_1 \in \{0, 1\}$ , respectively.

500 repetitions. The coverage probability is computed via the same method as in the previous example. The numerical results are reported in Figure 3, Figure 4, and Table 1.

our model's performance in variable selection and inferential analysis under real applications.

# 5.2.1. Genomic Data

We first consider the genomic dataset on the relation between gene and riboflavin (vitamin  $B_2$ ) production with bacillus subtilis. Instead of evaluating the performance of variable selection in the previous neural imaging application, we aim to

# 5.2. Real Data

We apply the kernel-sieve estimator to two types of real datasets: a genomic dataset and a neural imaging dataset. We aim to test



Figure 4. Kernel-sieve hybrid estimators for the two-dimensional surface  $a_2(x_1)f_2(x_2)$ .

**Table 1.** Comparison of coverage probability for confidence bands at significant level 95 for the zero function  $f_5$  and nonzero function  $f_1$  in SpAM model  $Y = \sum_{j=1}^{4} f_j(X_j) + \varepsilon$  as long as the zero function  $a_1f_1$  for  $a_1 = 1$  and nonzero function  $a_1f_1$  for  $a_1 = 0$  in the ATLAS model  $Y = a_1f_1(X_1) + \sum_{j=2}^{4} a_j(X_1)f_j(X_j) + \varepsilon$ .

n	Method	Zero function		Nonzero function	
		Coverage probability	Area	Coverage probability	Area
	SpAM	0.824	0.398	0.932	0.145
400	ATLAS	0.924	0.402	0.912	0.210
	Kozbur (2015)	0.984	3.583	0.992	3.543
	SpAM	0.836	0.377	0.928	0.137
500	ATLAS	0.922	0.346	0.924	0.158
	Kozbur (2015)	0.984	1.089	0.994	0.827
	SpAM	0.874	0.390	0.932	0.102
600	ATLAS	0.948	0.441	0.944	0.127
	Kozbur (2015)	0.988	1.791	0.984	1.550

NOTE: We also compare the numerical performance of the oracle method in Kozbur (2015) on the SpAM model. Here we set dimension d = 600, sample size n = 400, 500, 600 and  $\varepsilon \sim N(0, 1.5^2)$ . The covering probability and area are averaged based on the 500 repetitions.

demonstrate the inference analysis of our method. The dataset is provided by DSM (Kaiseraugst, Switzerland) and it is publicly available in Supplementary Section A.1 of Bühlmann, Kalisch, and Meier (2014). The response variable Y represents the logarithm of the riboflavin production rate. The covariates are the logarithm of gene expression levels with dimension d =4088 and sample size n = 71. van de Geer et al. (2014), Bühlmann, Kalisch, and Meier (2014), and Javanmard and Montanari (2014) use the linear model to find potentially significant genes. van de Geer et al. (2014) finds no significant genes, Bühlmann, Kalisch, and Meier (2014) finds the gene YXLD-at, and Javanmard and Montanari (2014) finds two genes YXLDat and YXLE-at to be significant. In this article, we use the sparse additive model to find whether the two genes YXLD-at and YXLE-at are significant. We first normalize the covariates onto [0, 1] and use (2.19) to construct confidence bands for the two genes YXLD-at and YXLE-at at significance level 95%. The results are illustrated in Figure 5. We can see that both genes have significantly nonzero effects. However, the gene YXLE-at has a larger part of the domain where zero is located within the confidence band compared to YXLD-at. Moreover, the magnitude of regression function on YXLE-at is smaller than YXLD-at. These explain the reason why YXLE-at is less significant than YXLD-at in the previous analysis.

# 5.2.2. Neural Imaging Data

The second application we consider is the ADHD-200 dataset (Biswal et al. 2010) on the resting-state fMRI of 195 children and adolescents diagnosed with attention deficit hyperactive



Figure 5. Kernel-sieve hybrid estimators for the riboflavin dataset using ATLAS model.

disorder (ADHD) along with 491 typically developing controls. Among them, 246 individuals are measured by the ADHD index (Conners 2008) which assesses the level of disorder. To explore the connection between ADHD and the brain activities, we aim to regress the ADHD index by the fMRI data of 264 voxels selected by Power et al. (2011) as the representative functional cerebral areas. Phenotypic information including age, gender, and intelligence quotient (IQ) is also provided.

Several studies have revealed that the maturation of the brains for the youth with ADHD is delayed in some cortical regions, compared to the ones without disorder (Mann et al. 1992; El-Sayed et al. 2003; Shaw et al. 2007). For example, Shaw et al. (2007) find that the cortical development for the individuals with ADHD is significantly slower in the frontal lobe and temporal lobe. Therefore, the functioning voxels related to ADHD vary with age and the ATLAS model can characterize such variation, while the sparse additive model cannot. We set the age as the longitude variable and the fMRI of 264 voxels as the other covariates. All the covariates are normalized to [0, 1]. Each of the 246 subjects with ADHD indices has 76–276 scans and all the scans are treated as independent observations.

The results of the regression are illustrated in Figures 6 and 7. We show the first eight estimated surfaces with largest maximum norms among  $\{\widehat{f}_j(x_j, x_1)\}_{j=1}^d$  in Figure 6. In the center of Figure 6, we demonstrate all voxels being activated (nonzero)

at certain times by small balls. The radius of a ball represents the length of time the corresponding voxel is activated and the maximum norm is represented by the ball's color where red means the largest values and yellow means the smallest (see the colorbar on the right bottom of Figure 6). We can see that most of the voxels with strongest signal strength are in the frontal and temporal lobes, which matches the results in Shaw et al. (2007). Moreover, the different flat zero areas of different surfaces in Figure 6 imply that the voxels are not activated simultaneously, which supports the necessity of the ATLAS model. In Figure 7, we show the activated voxels at different ages. The radii and colors of the balls are the same as Figure 6. We observe that, with the increasing age, the number of activated voxels first ascends and then reduces. This is similar to the results in Shaw et al. (2007) showing that 50 cortical points of ADHD groups attain peak thickness around the age of 10.5 years. The decreasing number of activated voxels after age 15 is also congruent with the discovery in Shaw et al. (2007).

# 6. Discussion

In this article, we consider a novel nonparametric model, ATLAS, which is a generalization of the sparse additive



Figure 6. The estimated surfaces of first eight voxels with largest maximum norms. The radii of the balls in the brain represent the duration the voxels being active and the colors represent the maximum norms of the surfaces, whose corresponding values are indicated by the colorbar on the right bottom of the figure.



Figure 7. Active voxels varying with age. Each column shows the active voxels at each age. The radii and colors of the balls in a brain represent the duration and maximum norms of the active voxels as in Figure 6.

model. ATLAS naturally models high-dimensional nonparametric functions having different sparsity in different local regions of the domain. We consider the kernel-sieve hybrid regression to estimate the unknown function. Since we consider functions in the 2nd order Hölder class, only Nadaraya– Watson-type kernel estimator is considered. However, it is not hard to generalize the loss function in (2.6) to local polynomial regression

$$\mathcal{L}_{z}(\alpha, \beta) = \frac{1}{n} \sum_{i=1}^{n} K_{h}(X_{i1} - z) \left( Y_{i} - \bar{Y} - \alpha - \sum_{\ell=1}^{p} \frac{(X_{i1} - z)^{\ell}}{\ell!} - \sum_{j=2}^{d} \sum_{k=1}^{m} \psi_{jk}(X_{ij}) \beta_{jk} \right)^{2}.$$

We can apply a similar proof technique to show the statistical rate of the estimator based on the generalized loss in higher order Hölder classes. Corresponding methods to construct confidence bands can also be applied.

# **Supplementary Materials**

In the supplementary materials, we provide proofs of the theoretical results in the paper and the technical lemmas supporting them. We also introduce an accelerated method to obtain our kernel-sieve estimator.

### Funding

The authors are grateful for the support of NSF DMS1916211, NIH R35CA220523, NSF CAREER Award DMS1454377, NSF IIS1408910, NSF IIS1332109, NIH R01MH102339, NIH R01GM083084, and NIH R01HG06841. This work is also supported by an IBM Corporation Faculty Research Fund at the University of Chicago Booth School of Business. This work was completed in part with resources provided by the University of Chicago Research Computing Center.

# References

- Avalos, M., Grandvalet, Y., and Ambroise, C. (2007), "Parsimonious Additive Models," *Computational Statistics & Data Analysis*, 51, 2851–2870. [2085]
- Belloni, A., Chernozhukov, V., and Hansen, C. B. (2013), "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650. [2085]
- Belloni, A., Chernozhukov, V., and Kato, K. (2013), "Robust Inference in High-Dimensional Approximately Sparse Quantile Regression Models," arXiv no. 1312.7186. [2085]
- (2015), "Uniform Post Selection Inference for LAD Regression and Other Z-Estimation Problems," *Biometrika*, 102, 77–94. [2085]
- Belloni, A., Chernozhukov, V., and Wei, Y. (2013), "Honest Confidence Regions for Logistic Regression With a Large Number of Controls," arXiv no. 1304.3969. [2085]
- Berhane, K., and Tibshirani, R. J. (1998), "Generalized Additive Models for Longitudinal Data," *Canadian Journal of Statistics*, 26, 517–535. [2092]
- Bertin, K., and Lecué, G. (2008), "Selection of Variables and Dimension Reduction in High-Dimensional Non-Parametric Regression," *Electronic Journal of Statistics*, 2, 1224–1241. [2085]
- Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., and Dogonowski, A. M. (2010), "Toward Discovery Science of Human Brain Function," *Proceedings of the National Academy of Sciences of the United States of America*, 107, 4734–4739. [2095]
- Bousquet, O. (2002), "A Bennett Concentration Inequality and Its Application to Suprema of Empirical Processes," *Comptes Rendus Mathematique*, 334, 495–500. [2084]
- Bühlmann, P., Kalisch, M., and Meier, L. (2014), "High-Dimensional Statistics With a View Toward Applications in Biology," *Annual Review of Statistics and Its Application*, 1, 255–278. [2095]
- Bühlmann, P., and van de Geer, S. A. (2011), Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer Series in Statistics, Heidelberg: Springer. [2085]
- Chatterjee, A., and Lahiri, S. N. (2013), "Rates of Convergence of the Adaptive LASSO Estimators to the Oracle Distribution and Higher Order Refinements by the Bootstrap," *The Annals of Statistics*, 41, 1232–1259. [2085]

- Chernozhukov, V., Chetverikov, D., and Kato, K. (2013), "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors," *The Annals of Statistics*, 41, 2786–2819. [2085]
- (2014a), "Anti-Concentration and Honest, Adaptive Confidence Bands," *The Annals of Statistics*, 42, 1787–1818. [2085,2088]
- (2014b), "Gaussian Approximation of Suprema of Empirical Processes," *The Annals of Statistics*, 42, 1564–1597. [2085,2088]
- Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688. [2090]
- Claeskens, G., and Van Keilegom, I. (2003), "Bootstrap Confidence Bands for Regression Curves and Their Derivatives," *The Annals of Statistics*, 31, 1852–1884. [2084,2087]
- Comminges, L., and Dalalyan, A. S. (2012), "Tight Conditions for Consistency of Variable Selection in the Context of High Dimensionality," *The Annals of Statistics*, 40, 2667–2696. [2085]
- Conners, C. K. (2008), *Conners (Conners 3)* (3rd ed.), Toronto, ON: Multi-Health Systems. [2096]
- Dalalyan, A. S., Ingster, Y., and Tsybakov, A. B. (2014), "Statistical Inference in Compound Functional Models," *Probability Theory and Related Fields*, 158, 513–532. [2085,2092]
- de Boor, C. (2001), *A Practical Guide to Splines*, Applied Mathematical Sciences (Vol. 27, revised ed.), New York: Springer-Verlag. [2086]
- Efron, B. (2012), Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction (Vol. 1), Cambridge, UK: Cambridge University Press. [2093]
- El-Sayed, E., Larsson, J.-O., Persson, H., Santosh, P., and Rydelius, P.-A. (2003), "Maturational Lag' Hypothesis of Attention Deficit Hyperactivity Disorder: An Update," *Acta Paediatrica*, 92, 776–784. [2096]
- Fan, J. (1993), "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216. [2089]
- Fan, J., and Gijbels, I. (1996), Local Polynomial Modelling and Its Applications, Monographs on Statistics and Applied Probability (Vol. 66), London: Chapman & Hall. [2091]
- Fan, J., and Jiang, J. (2005), "Nonparametric Inferences for Additive Models," *Journal of the American Statistical Association*, 100, 890–907. [2084, 2089,2090,2091]
- Fan, J., and Yao, Q. (2008), Nonlinear Time Series: Nonparametric and Parametric Methods, New York: Springer Science & Business Media. [2091]
- Fan, J., and Zhang, W. (1999), "Statistical Estimation in Varying Coefficient Models," *The Annals of Statistics*, 27, 1491–1518. [2092]
- (2000), "Simultaneous Confidence Bands and Hypothesis Testing in Varying-Coefficient Models," *Scandinavian Journal of Statistics*, 27, 715–731. [2084,2087]
- Farrell, M. H. (2013), "Robust Inference on Average Treatment Effects With Possibly More Covariates Than Observations," arXiv no. 1309.4686. [2085]
- Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," Journal of the American Statistical Association, 76, 817–823. [2084]
- Hansen, B. E. (2008), "Uniform Convergence Rates For Kernel Estimation With Dependent Data," *Econometric Theory*, 24, 726–748. [2091]
- Härdle, W. (1989), "Asymptotic Maximal Deviation of M-Smoothers," Journal of Multivariate Analysis, 29, 163–179. [2084,2087]
- Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, Monographs on Statistics and Applied Probability (Vol. 43), London: Chapman and Hall, Ltd. [2084]
- (1993), "Varying-Coefficient Models" (with discussion and a reply by the authors), *Journal of the Royal Statistical Society*, Series B, 55, 757– 796. [2092]
- Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [2084,2085,2087,2089,2090,2091,2093]
- Javanmard, A., and Montanari, A. (2013), "Nearly Optimal Sample Size in Hypothesis Testing for High-Dimensional Regression," arXiv no. 1311.0274. [2085]
  - (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [2085,2088,2090,2095]

- Kato, K. (2012), "Two-Step Estimation of High Dimensional Additive Models," arXiv no. 1207.5313. [2084,2085,2087,2089]
- Koltchinskii, V. (2011), Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems, Lecture Notes in Mathematics (Vol. 2033), Heidelberg: Springer. [2084]
- Koltchinskii, V., and Yuan, M. (2010), "Sparsity in Multiple Kernel Learning," *The Annals of Statistics*, 38, 3660–3695. [2084,2085,2087,2089]
- Kozbur, D. (2015), "Inference in Additively Separable Models With a High-Dimensional Set of Conditioning Variables," arXiv no. 1503.05436. [2085,2087,2093,2095]
- Lafferty, J. D., and Wasserman, L. A. (2008), "Rodeo: Sparse, Greedy Nonparametric Regression," *The Annals of Statistics*, 36, 28–63. [2085]
- Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2013), "Exact Post-Selection Inference With the Lasso," arXiv no. 1311.6238. [2085]
- Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 34, 2272–2297. [2084,2085]
- Liu, H., and Yu, B. (2013), "Asymptotic Properties of Lasso+mLS and Lasso+Ridge in Sparse High-Dimensional Linear Regression," *Electronic Journal of Statistics*, 7, 3124–3169. [2085]
- Lockhart, R., Taylor, J. E., Tibshirani, R. J., and Tibshirani, R. J. (2014), "A Significance Test for the Lasso," *The Annals of Statistics*, 42, 413–468. [2085]
- Lopes, M. (2014), "A Residual Bootstrap for High-Dimensional Regression With Near Low-Rank Designs," in Advances in Neural Information Processing Systems (Vol. 27), Curran Associates, Inc., pp. 3239–3247. [2085]
- Lou, Y., Bien, J., Caruana, R., and Gehrke, J. (2014), "Sparse Partially Linear Additive Models," arXiv no. 1407.4729. [2084,2085]
- Mann, C. A., Lubar, J. F., Zimmerman, A. W., Miller, C. A., and Muenchen, R. A. (1992), "Quantitative Analysis of EEG in Boys With Attention-Deficit-Hyperactivity Disorder: Controlled Study With Clinical Implications," *Pediatric Neurology*, 8, 30–36. [2096]
- Masry, E. (1996), "Multivariate Local Polynomial Regression for Time Series: Uniform Strong Consistency and Rates," *Journal of Time Series Analysis*, 17, 571–599. [2091]
- Meier, L., van de Geer, S. A., and Bühlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [2084,2085, 2087,2089,2093]
- Meinshausen, N. (2013), "Group-Bound: Confidence Intervals for Groups of Variables in Sparse High-Dimensional Regression Without Assumptions on the Design," arXiv no. 1309.3489. [2085]
- Meinshausen, N., and Bühlmann, P. (2010), "Stability Selection," Journal of the Royal Statistical Society, Series B, 72, 417–473. [2085]
- Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "p-Values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [2085]
- Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A Unified Framework for High-Dimensional Analysis of *M*-Estimators With Decomposable Regularizers," *Statistical Science*, 27, 538–557. [2089]
- Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals* of Statistics, 45, 158–195. [2090]
- Nze, P. A., and Doukhan, P. (2004), "Weak Dependence: Models and Applications to Econometrics," *Econometric Theory*, 20, 995–1045. [2091]
- Opsomer, J. D., and Ruppert, D. (1997), "Fitting a Bivariate Additive Model by Local Polynomial Regression," *The Annals of Statistics*, 25, 186–211. [2084,2089,2090,2091]
- Peligrad, M. (1992), "Properties of Uniform Consistency of the Kernel Estimators of Density and Regression Functions Under Dependence Assumptions," *Stochastics*, 40, 147–168. [2091]
- Petersen, A., Witten, D. M., and Simon, N. (2014), "Fused Lasso Additive Model," arXiv no. 1409.5391. [2084,2085]
- Power, J. D., Cohen, A. L., Nelson, S. M., Wig, G. S., Barnes, K. A., Church, J. A., Vogel, A. C., Laumann, T. O., Miezin, F. M., Schlaggar, B. L., and Petersen, S. E. (2011), "Functional Network Organization of the Human Brain," *Neuron*, 72, 665–678. [2096]
- Raskutti, G., Wainwright, M. J., and Yu, B. (2012), "Minimax-Optimal Rates for Sparse Additive Models Over Kernel Classes via Convex Programming," *Journal of Machine Learning Research*, 13, 389–427. [2085,2089,2090]

- Ravikumar, P., Lafferty, J. D., Liu, H., and Wasserman, L. A. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society*, Series B, 71, 1009–1030. [2084,2085,2087,2093]
- Richtárik, P., and Takáč, M. (2014), "Iteration Complexity of Randomized Block-Coordinate Descent Methods for Minimizing a Composite Function," *Mathematical Programming*, 144, 1–38. [2087]
- Rosasco, L., Villa, S., Mosci, S., Santoro, M., and Verri, A. (2013), "Nonparametric Sparsity and Regularization," *Journal of Machine Learning Research*, 14, 1665–1714. [2085]
- Sardy, S., and Tseng, P. (2004), "AMlet, RAMlet, and GAMlet: Automatic Nonlinear Fitting of Additive Models, Robust and Generalized, With Wavelets," *Journal of Computational and Graphical Statistics*, 13, 283– 309. [2084,2085]
- Schumaker, L. L. (2007), Spline Functions: Basic Theory, Cambridge Mathematical Library (3rd ed.), Cambridge: Cambridge University Press. [2086]
- Shah, R. D., and Samworth, R. J. (2013), "Variable Selection With Error Control: Another Look at Stability Selection," *Journal of the Royal Statistical Society*, Series B, 75, 55–80. [2085]
- Shaw, P., Eckstrand, K., Sharp, W., Blumenthal, J., Lerch, J., Greenstein, D. E. A., Clasen, L., Evans, A., Giedd, J., and Rapoport, J. (2007), "Attention-Deficit/Hyperactivity Disorder Is Characterized by a Delay in Cortical Maturation," *Proceedings of the National Academy of Sciences of the United States of America*, 104 19649–19654. [2096]
- Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360. [2090]
- (1985), "Additive Regression and Other Nonparametric Models," The Annals of Statistics, 13, 689–705. [2084]
- Sun, J., and Loader, C. R. (1994), "Simultaneous Confidence Bands for Linear Regression and Smoothing," *The Annals of Statistics*, 22, 1328– 1345. [2084,2087]
- Taylor, J. E., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. J. (2014), "Post-Selection Adaptive Inference for Least Angle Regression and the Lasso," arXiv no. 1401.3889. [2085]

- van de Geer, S. A., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [2085,2088,2090,2095]
- van der Vaart, A. W., and Wellner, J. A. (1996), Weak Convergence and Empirical Processes: With Applications to Statistics, New York: Springer. [2084]
- Wahl, M. (2014), "Variable Selection in High-Dimensional Additive Models Based on Norms of Projections," arXiv no. 1406.0052. [2085]
- Wasserman, L. (2006), All of Nonparametric Statistics, New York: Springer Science & Business Media. [2084]
- Wasserman, L. A., and Roeder, K. (2009), "High-Dimensional Variable Selection," *The Annals of Statistics*, 37, 2178–2201. [2085]
- Wei, F., Huang, J., and Li, H. (2011), "Variable Selection and Estimation in High-Dimensional Varying-Coefficient Models," *Statistica Sinica*, 21, 1515–1540. [2092]
- Xu, M., Chen, M., and Lafferty, J. D. (2014), "Faithful Variable Screening for High-Dimensional Convex Regression," arXiv no. 1411.1805. [2085]
- Yang, Y., and Tokdar, S. T. (2014), "Minimax-Optimal Nonparametric Regression in High Dimensions," arXiv no. 1401.7278. [2085]
- Zhang, C.-H., and Zhang, S. S. (2013), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal* of the Royal Statistical Society, Series B, 76, 217–242. [2085,2088,2090]
- Zhang, H. H., and Lin, Y. (2006), "Component Selection and Smoothing for Nonparametric Regression in Exponential Families," *Statistica Sinica*, 16, 1021–1041. [2093]
- Zhang, W., and Peng, H. (2010), "Simultaneous Confidence Band and Hypothesis Test in Generalised Varying-Coefficient Models," *Journal of Multivariate Analysis*, 101, 1656–1680. [2084,2087]
- Zhang, X., and Wang, J.-L. (2015), "Varying-Coefficient Additive Models for Functional Data," *Biometrika*, 102, 15–32. [2092]
- Zhu, H., Li, R., and Kong, L. (2012), "Multivariate Varying Coefficient Model for Functional Responses," *The Annals of Statistics*, 40, 2634– 2666. [2092]