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

Junwei Lu,* Mladen Kolar,† Han Liu‡

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.

1 Introduction

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

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

*Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115; Email: junweilu@hsph.harvard.edu

†Booth School of Business, The University of Chicago, Chicago, IL 60637; Email: mkolar@chicagobooth.edu

‡Department of Computer Science, Northwestern University, Evanston, IL 60208; Email: hanliu@northwestern.edu

where $\boldsymbol{X} \in \mathcal{X}^d \subseteq \mathbb{R}^d$ is a d-dimensional random vector in $\mathcal{X}^d$, ε is random error satisfying $\mathbb{E}[\varepsilon \,|\, \boldsymbol{X}] = 0$, 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

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

where μ 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; Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Koltchinskii and Yuan, 2010; Kato, 2012; Petersen et al., 2014; Lou et al., 2014). These papers 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 (Koltchinskii, 2011; van der Vaart and Wellner, 1996; Bousquet, 2002). The de-biasing procedure for the kernel-sieve hybrid regression estimator extends the approach used in the ℓ_1 penalized high dimensional linear regression (Zhang and Zhang, 2013; van de Geer et al., 2014; Javanmard and Montanari, 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 et al. (2014a) and Chernozhukov et al. (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 high dimensional inference. Initial work on high dimensional statistics has focused on estimation and prediction (see, for example, 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 towards the latter problems. Initial work on construction of p-values in high dimensional models relied on correct inclusion of the relevant variables (Wasserman and Roeder, 2009; Meinshausen et al., 2009). Meinshausen and Bühlmann (2010) and Shah and Samworth (2013) study stability selection procedure, which provides the family-wise error rate for any selection procedure. Hypothesis testing and confidence intervals for low dimensional parameters in high dimensional linear and generalized linear models are studied in Belloni et al. (2013a), Belloni et al. (2013c), 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 ℓ_1 penalized estimator in the first stage. Similar results are obtained in the context of ℓ_1 penalized least absolute deviation and quantile regression (Belloni et al., 2015, 2013b). Kozbur (2015) extends the approach developed in Belloni et al. (2013a) to a nonparametric

regression setting, where a pointwise confidence interval is obtained based on the penalized series estimator. Meinshausen (2013) studies construction of one-sided confidence intervals for groups of variables under weak assumptions on the design matrix. Lockhart et al. (2014) studies significance of the input variables that enter the model along the lasso path. Lee et al. (2013) and Taylor et al. (2014) perform post-selection inference conditional on the selected model. Chatterjee and Lahiri (2013), Liu and Yu (2013), Chernozhukov et al. (2013) and Lopes (2014) study 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 high dimensional 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) study 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 papers have studied the sparse additive model in (1.2) (Sardy and Tseng, 2004; Lin and Zhang, 2006; Avalos et al., 2007; Ravikumar et al., 2009; Meier et al., 2009; Huang et al., 2010; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Kato, 2012; Petersen et al., 2014; Rosasco et al., 2013; Lou et al., 2014; Wahl, 2014). In addition, Xu et al. (2014) study a high dimensional convex nonparametric regression. Dalalyan et al. (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 et al. (2009), Huang et al. (2010), Koltchinskii and Yuan (2010), Raskutti et al. (2012), and Kato (2012) develop 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 ℓ_1/ℓ_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 Paper

The rest of the paper 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 $\boldsymbol{a} \in \mathbb{R}^d$, we let $\text{supp}(\boldsymbol{a}) = \{j \,|\, a_j \neq 0\}$ be the support set (with an analogous definition for matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times n_2}$), $\|\boldsymbol{a}\|_q$, for $q \in [1, \infty)$, the ℓ_q-norm defined as $\|\boldsymbol{a}\|_q = (\sum_{i\in[n]} |a_i|^q)^{1/q}$ with the usual extensions for $q \in \{0, \infty\}$, that is, $\|\boldsymbol{a}\|_0 = |\text{supp}(\boldsymbol{a})|$ and $\|\boldsymbol{a}\|_\infty = \max_{i\in[n]} |a_i|$. If the vector $\boldsymbol{a} \in \mathbb{R}^d$ is decomposed into groups such that $\boldsymbol{a} = (\boldsymbol{a}_{\mathcal{G}_1}, \ldots, \boldsymbol{a}_{\mathcal{G}_g})^T$, where $\mathcal{G}_1, \ldots, \mathcal{G}_g \subset [d]$ are disjoint sets, we denote $\|\boldsymbol{a}\|_{p,q}^q = \sum_{k=1}^g \|\boldsymbol{a}_{\mathcal{G}_k}\|_p^q$ and $\|\boldsymbol{a}\|_{p,\infty} = \max_{k\in[g]} \|\boldsymbol{a}_{\mathcal{G}_k}\|_p$ 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 $\boldsymbol{a}_{\backslash j} = (a_1, \ldots, a_{j-1}, a_{j+1} \ldots, a_d)^T$. 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 $\text{vec}(\mathbf{A})$ to denote the vector in $\mathbb{R}^{n_1 n_2}$ formed by stacking the columns of $\mathbf{A}$. We denote the Frobenius norm of $\mathbf{A}$ 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 $\alpha_n \asymp \beta_n$. The notation $\alpha_n = o(\beta_n)$ is used to denote that $a_n \beta_n^{-1} \xrightarrow{n\to\infty} 0$. Throughout the paper, 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 γ-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 γ. The derivative $f^{(\ell)}$ satisfies

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

Let $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ 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

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

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

$$\mathbb{E}[f_j(X_j)] = 0, \text{ for all } j = 1, \ldots, d. \tag{2.2}$$

Define the sparse additive functions class

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

Let $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ be n independent random samples of $(\boldsymbol{X}, Y)$ 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, $\mathcal{B}_m = \text{Span}(\phi_1, \ldots, \phi_m)$. In particular, we define

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

where ψ_{jk}^*'s are the locally centered bases defined as

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

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 ψ_{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 $\psi_{jk}(x) = \phi_k(x) - \bar{\phi}_{jk}$.

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

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

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

$$(\widehat{\alpha}_z, \widehat{\boldsymbol{\beta}}_z) = \underset{\alpha, \boldsymbol{\beta}}{\arg\min}\, \mathcal{L}_z(\alpha, \boldsymbol{\beta}) + \lambda \mathcal{R}(\alpha, \boldsymbol{\beta}), \tag{2.7}$$

where the penalty function is

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

with λ being a tuning parameter. We estimate the additive functions $\{f_j\}_{j\in[d]}$ by $\widehat{f}_1(z) = \widehat{\alpha}_z$ and

$\widehat{f}_j(x) = \sum_{k=1}^m \psi_{jk}(x)\widehat{\boldsymbol{\beta}}_{jk;z}$ for $j \geq 2$. Based on $\widehat{\alpha}_z, \widehat{\boldsymbol{\beta}}_z$, 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

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

where $\widehat{\boldsymbol{\beta}}_{jk;z}$ is the coordinate of $\widehat{\boldsymbol{\beta}}_z$ corresponding to the kth B-spline basis of the jth covariate.

Remark 2.2. The estimators $\widehat{\alpha}_z$ and $\widehat{\boldsymbol{\beta}}_z$ are estimating different quantities. Notice that $\widehat{\alpha}_z$ estimates the scalar $f_1(z)$, while $\widehat{\boldsymbol{\beta}}_z$ estimates the coefficients of B-splines. Given a function $g(x) = \sum_{k=1}^m \beta_k \phi_k(x)$, we have $\|g\|_2^2 \asymp m^{-1} \sum_{k=1}^m \beta_k^2$ (see, e.g., Corollary 15 in Chapter XI of de Boor (2001)). From this we see that the scales of $\widehat{\alpha}_z$ and $\widehat{\boldsymbol{\beta}}_z$ 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 papers consider the sieve-type estimator. For example, Huang et al. (2010) consider minimizing

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

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

Kozbur (2015) proposes 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 $\cup_{k=0}^m I_k$. 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 α. Therefore, in order to study the asymptotic properties of the extreme value

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

we need to analyze the m-dimensional estimator $\widehat{\boldsymbol{\beta}}_1^{\text{sieve}}$ whose dimension m is increasing with the sample size n at the rate $m \asymp n^{1/6}$. 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 $\widehat{\boldsymbol{\beta}}_1^{\text{sieve}}$, while the extreme value is more challenging as it is non-differentiable. We further note that most existing papers on 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 $\widehat{\alpha}_z$ of $f_1(z)$. This one dimensional estimator $\widehat{\alpha}_z$ 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 $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_{1\bullet}, \ldots, \boldsymbol{\Psi}_{n\bullet})^T \in \mathbb{R}^{n\times(1+(d-1)m)}$, where $\boldsymbol{\Psi}_{ij} = (\psi_{j1}(X_{ij}), \ldots, \psi_{jm}(X_{ij}))^T$ and $\boldsymbol{\Psi}_{i\bullet} = (1, \boldsymbol{\Psi}_{i2}^T, \ldots, \boldsymbol{\Psi}_{id}^T)^T \in \mathbb{R}^{1+(d-1)m}$ for $i \in [n]$ and $j \geq 2$. We also write $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_{\bullet 1}, \ldots, \boldsymbol{\Psi}_{\bullet d})$,

Algorithm 1 Randomized coordinate descent for group Lasso

for $t = 1, 2, \ldots$ **do**

Let $\boldsymbol{\beta}_+^{(t)} = (\boldsymbol{\beta}_1^{(t)}, \boldsymbol{\beta}_2^{(t)T}, \ldots, \boldsymbol{\beta}_j^{(t)T})^T$.

Choose $j_t = j \in [d]$ with probability $1/d$.

Compute $T(\boldsymbol{\beta}_j^{(t)})$ for the j-th block as

$$T(\boldsymbol{\beta}_j^{(t)}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^{\dim(\boldsymbol{\beta}_j)}}{\arg\min} \left\{ \frac{\mu}{2} \|\boldsymbol{\theta}\|_2^2 + \langle \nabla_j \mathcal{L}_z(\boldsymbol{\beta}_+^{(t)}), \boldsymbol{\theta} \rangle + \lambda_j \|\boldsymbol{\theta} + \boldsymbol{\beta}_j^{(t)}\|_2 \right\}. \tag{2.14}$$

Update $\boldsymbol{\beta}_j^{(t+1)} = \boldsymbol{\beta}_j^{(t)} + T(\boldsymbol{\beta}_j^{(t)})$.

end for

where $\boldsymbol{\Psi}_{\bullet 1} = (1, \ldots, 1)^T \in \mathbb{R}^n$ and $\boldsymbol{\Psi}_{\bullet j} = (\boldsymbol{\Psi}_{1j}, \ldots, \boldsymbol{\Psi}_{nj})^T \in \mathbb{R}^{n \times m}$ for $j \geq 2$. We further denote

$$\boldsymbol{Y} = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^T \in \mathbb{R}^n, \quad \boldsymbol{\beta}_+ = (\alpha, \boldsymbol{\beta}^T)^T \in \mathbb{R}^{1+(d-1)m}, \boldsymbol{\beta}_+^* = \big(f_1^*(z), \boldsymbol{\beta}^{*T}\big)^T \in \mathbb{R}^{1+(d-1)m}$$
$$\text{and } \mathbf{W}_z = \text{diag}\big(K_h(X_{11} - z), \ldots, K_h(X_{n1} - z)\big) \in \mathbb{R}^{n \times n}. \tag{2.12}$$

To unify the notation in our algorithm, we also write $\boldsymbol{\beta}_+ = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_d^T)^T$, where $\boldsymbol{\beta}_1 = \alpha$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_d^T)^T$. 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

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

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

$$T(\boldsymbol{\beta}_j^{(t)}) = \mathcal{T}_{\lambda_j/\mu}\Big(\boldsymbol{\beta}_j^{(t)} - \frac{1}{L}\nabla_j \mathcal{L}_z\big(\boldsymbol{\beta}_+^{(t)}\big)\Big) - \boldsymbol{\beta}_j^{(t)}, \tag{2.15}$$

where μ is certain regularized constant and $\mathcal{T}_\lambda$ is the soft-thresholding 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 material.

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 $\mathcal{C}_n$ is a set of confidence intervals $\mathcal{C}_n = \{\mathcal{C}_n(z) = [c_L(z), c_U(z)] \,|\, z \in \mathcal{X}\}$. 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 $f \in \mathcal{C}_n$ to denote that f lies in the confidence band, that is, $f(z) \in \mathcal{C}_n(z)$ 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 $\widehat{\alpha}_z$ that reduces the bias introduced by (2.7).

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

$$\widehat{\boldsymbol{\theta}}_z = \underset{\boldsymbol{\theta} \in \mathbb{R}^{(d-1)m+1}}{\arg\min} \; \boldsymbol{\theta}^T \widehat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta}, \qquad \text{subject to} \qquad \left\| \widehat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta} - \mathbf{e}_1 \right\|_{2,\infty} \leq \gamma, \tag{2.16}$$

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

$$\widehat{f}_1^u(z) = \widehat{\alpha}_z + \frac{1}{n} \widehat{\boldsymbol{\theta}}_z^T \boldsymbol{\Psi}^T \mathbf{W}_z (\boldsymbol{Y} - \boldsymbol{\Psi} \widehat{\boldsymbol{\beta}}_+). \tag{2.17}$$

We proceed to construct a confidence band based on this de-biased 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

$$\widehat{\mathbb{H}}_n(z) = \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^{n} \xi_i \cdot \frac{\widehat{\sigma} K_h(X_{i1} - z) \boldsymbol{\Psi}_i^T \widehat{\boldsymbol{\theta}}_z}{\widehat{\sigma}_n(z)}, \tag{2.18}$$

where $\xi_1, \ldots, \xi_n$ are independent $N(0,1)$ random variables, and the variance estimators are given as $\widehat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \widehat{\alpha}_{X_i} - \sum_{j=2}^d \sum_{k=1}^m \boldsymbol{\Psi}_{ij}^T \widehat{\boldsymbol{\beta}}_{jk;X_i})^2$ and $\widehat{\sigma}_n^2(z) = n^{-1} \widehat{\boldsymbol{\theta}}_z^T \boldsymbol{\Psi} \mathbf{W}_z^2 \boldsymbol{\Psi}^T \widehat{\boldsymbol{\theta}}_z$. 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 $100 \times (1-\alpha)\%$: $\mathcal{C}_{n,\alpha}^b = \{\mathcal{C}_{n,\alpha}^b(z) \mid z \in \mathcal{X}\}$, where

$$\mathcal{C}_{n,\alpha}^b(z) := [\widehat{f}_1^u(z) - \widehat{c}_n(\alpha)(nh)^{-1/2} \widehat{\sigma}_n(z), \widehat{f}_1^u(z) + \widehat{c}_n(\alpha)(nh)^{-1/2} \widehat{\sigma}_n(z)]. \tag{2.19}$$

We will show that the confidence band is asymptotically honest in Section 3.2 by building on the framework developed in Chernozhukov et al. (2014a) and Chernozhukov et al. (2014b), who study 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 $p(x_1, \ldots, x_d)$ denote the joint density of $\boldsymbol{X} = (X_1, \ldots, X_d)$ 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_ℓ), $p_{jk}(x_j, x_k)$ 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

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

(A3) (Design Matrix) Let $\mathbf{\Sigma}_z = \mathbb{E}[K_h(X_1 - z)\mathbf{\Psi}_{1\bullet}\mathbf{\Psi}_{1\bullet}^T]$, recalling that $\mathbf{\Psi}_{1\bullet} = (1, \mathbf{\Psi}_{12}^T, \ldots, \mathbf{\Psi}_{1d}^T)^T$. For any $J \subset [d]$, we define a cone

$$\mathbb{C}_{\beta}^{(\kappa)}(J) = \Big\{\boldsymbol{\beta}_+ = (\alpha, \boldsymbol{\beta}^T)^T \mid \textstyle\sum_{j\notin J, j\neq 1}\|\boldsymbol{\beta}_j\|_2 \leq \kappa \sum_{j\in J, j\neq 1}\|\boldsymbol{\beta}_j\|_2 + \kappa\sqrt{m}|\alpha|\Big\}. \tag{3.1}$$

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

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

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

(A5) The nonparametric function $f(x_1, \ldots, x_d) \in \mathcal{K}_d(s)$ 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 papers on additive model. For example, Huang et al. (2010) study 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

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

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 et al. (2010). As another example, consider $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}_{jj} = 1$ and $\boldsymbol{\Sigma}_{jk} = \rho$ 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 $\boldsymbol{\Sigma}$ 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 et al., 2009; Huang et al., 2010; Koltchinskii and Yuan, 2010; Raskutti et al., 2012; Kato, 2012).

Assumption (**A3**) is similar to the restricted strong convexity condition in Negahban et al. (2012). Note that $\boldsymbol{\Sigma}_z$ is the expectation of the Hessian matrix of the loss function $\mathcal{L}(\boldsymbol{\beta}_+)$. We require $\boldsymbol{\Sigma}_z$ 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 α and $\boldsymbol{\beta}_z$ are calibrated on the same scale (see Remark 2.2). Assumption (**A3**) can be derived from the assumption on the design in Koltchinskii and Yuan (2010). They consider the quantity

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

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

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

Proposition 3.1. We define a uniform quantity based on the constant (3.4) as

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

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

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

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

When the support $\mathcal{X}$ is compact and we assume there exists a fixed constant $b > 0$ such that $p_1(x_1) \geq b$ 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 et al., 2009; Huang et al., 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 3.2. Suppose that Assumptions (**A1**)-(**A5**) are satisfied. If $h = o(1), m \to \infty$ as $n \to \infty$, and we set

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

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

$$\sup_{z\in\mathcal{X}} \sum_{j=2}^{d} ||\widehat{\boldsymbol{\beta}}_{j;z} - \boldsymbol{\beta}_j^*||_2 \leq \frac{sm}{\rho_{\min}}\lambda \qquad \text{and} \qquad \sup_{z\in\mathcal{X}} |\widehat{a}_z - f_1(z)| \leq \frac{s\sqrt{m}}{\rho_{\min}}\lambda \tag{3.8}$$

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

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

with probability $1 - c/n$.

The estimation error comes from four sources. The noise ε contributes $O\left(\sqrt{\log(dmh^{-1})/nh}\right)$ in (3.7). The second term in (3.7), $O\left(\sqrt{s}m^{-5/2}\right)$, 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 3.2 is shown in Appendix A in the supplementary material.

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 $\lambda \asymp n^{-5/12}\sqrt{\log(dn)}$. With these choices, we obtain $\|\widehat{f} - f\|_2^2 = O_P\left(n^{-2/3}\log(dn)\right)$. This convergence rate is slower than the optimal rate $O_P\left(n^{-4/5} + \log d/n\right)$ for estimating the sparse additive model (Raskutti et al., 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 $\|\widehat{f} - f\|_2^2 = O_P\left(n^{-2/3}\log(dn)\right)$ 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 $T_n = \sup_{z\in\mathcal{X}} \max_{j\in[d]} \frac{1}{n}\|\mathbf{\Psi}_{\bullet j}^T \mathbf{W}_z \boldsymbol{\varepsilon}\|_2 = O_P\left(\sqrt{\log(dn)/(nh)}\right)$, where $\mathbf{W}_z$ is defined in (2.2). In comparison, Huang et al. (2010) only need to bound $T_n' = \sup_{z\in\mathcal{X}} \max_{j\in[d]} \frac{1}{n}\|\mathbf{\Psi}_{\bullet j}^T \boldsymbol{\varepsilon}\|_2 = O_P\left(\sqrt{\log(dn)/n}\right)$ (see their Lemma 2). Note that $T_n = O_P\left(h^{-1/2}T_n'\right)$ 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 material.

3.2 Theoretical Results for Confidence Band

In order to establish valid theoretical results on the confidence band $\mathcal{C}_{n,\alpha}^b$, 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 $\boldsymbol{X}$ satisfies

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

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 et al. (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) relax the sparsity assumption by requiring the inverse of the Hessian matrix to have columns with bounded ℓ_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 $\mathbf{\Sigma}_z$ defined in Assumption (**A3**). Due to the complicated definition of $\mathbf{\Sigma}_z$, there is no straightforward interpretation of $\mathbf{\Sigma}_z^{-1}$ and any assumption imposed on $\mathbf{\Sigma}_z^{-1}$ 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 high dimensional analogue of the assumption in (3.3), which is considered by Opsomer and Ruppert (1997) and Fan and Jiang (2005) for fixed dimensional additive models. Since $\|p_{1,j,k} - p_1 p_j p_k\|_2$ measures the dependency among X_1, X_j and X_k, (3.10) requires the ℓ_1-norms of both $\{\|p_{1,j,k} - p_1 p_j p_k\|_2\}_{j\geq 2}$ and $\{\|p_{1,j} - p_1 p_j\|_2\}_{j\geq 2}$ 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 material.

Proposition 3.3. Given any $\rho \in (0, 1/2)$ satisfying $\rho \leq \rho_{\min}/(18B)$, there exists a d-dimensional density function $p(\boldsymbol{x})$ with unbounded support such that

$$\sum_{j=2}^{d} \mathrm{Cov}(X_1, X_j) \geq \rho/9 \text{ and } \mathrm{Cov}(X_j, X_k) \geq \rho/9 \text{ for all } j, k > 2 \text{ and } |j-k| \leq 1,$$

and Assumption (**A6**) is satisfied.

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

Lemma 3.4. Suppose that Assumptions (**A1**), (**A3**) and (**A6**) hold. Let

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

for sufficiently large constant C. Then the vector $\boldsymbol{\theta}_z = \mathbf{\Sigma}_z^{-1}\mathbf{e}_1$ is a feasible solution to the

optimization program in (2.16) with high probability. In particular, we have

$$\mathbb{P}\left(\sup_{x\in\mathcal{X}}\|\widehat{\mathbf{\Sigma}}_z\boldsymbol{\theta}_z-\mathbf{e}_1\|_{2,\infty}\le\gamma\right)\ge 1-c/d$$

for some constant c.

We defer the proof of this lemma to Appendix F.2 in the supplementary material. 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 3.5. 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 $\inf_z p_1(z)>0$. If $s=O(1)$, $m\asymp n^p$ for $p\in(1/5,3/13)$, $h\asymp n^{-\delta}$ for $\delta\in(5p-1,(1-3p)/2)$, λ 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 $\mathcal{C}^b_{n,\alpha}$ in (2.19) is

$$\mathbb{P}\big(f_1(z)\in\mathcal{C}^b_{n,\alpha}(z),\ \text{for all } z\in\mathcal{X}\big)\ge 1-\alpha-C_1n^{-c}. \tag{3.12}$$

In particular, the confidence band $\mathcal{C}^b_{n,\alpha}$ is asymptotically honest, that is,

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

Theorem 3.7 below provides a more general result for the setting where $\mathcal{X}$ is unbounded. Notice that we can no longer choose $h\asymp n^{-1/6}$ and $m\asymp n^{1/6}$ as in Theorem 3.2, since we need to under-regularize our estimator in order 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 papers on additive model (Opsomer and Ruppert, 1997; Fan and Jiang, 2005; Huang et al., 2010) impose bounded support condition, as well as many paper that study uniform convergence of kernel-type estimators (Peligrad, 1992; Masry, 1996; Fan and Yao,

2008; Nze and Doukhan, 2004). For example, consider the univariate nonparametric model that n i.i.d. 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

$$\widehat{m}(z) = \arg\min_a \sum_{i=1}^n K_h(X_i - z)(Y_i - a)^2 = \frac{\sum_{i=1}^n Y_i K_h(X_i - z)}{\sum_{i=1}^n K_h(X_i - z)}. \tag{3.13}$$

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

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

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 $\mathrm{MSE}(\widehat{m}(z))$ 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, in order 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) proves 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 3.6. Suppose Assumptions (**A1**), (**A2**), (**A4**), (**A5**) hold and $\bar{\beta}_{2,\kappa}$ in (3.5) is finite. We consider $s = O(1)$, $h \asymp n^{-1/6}$, $m \asymp n^{1/6}$ and $\lambda = Cn^{-5/12}\sqrt{\log(dn)}$ for sufficiently large constant C. Given any compact interval $[-D_n, D_n]$, if $b_n = \inf_{|z|\le D_n} p_1(z) > 0$, we have

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

The corollary can be proved by applying Proposition 3.1 to Theorem 3.2. 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 3.6 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 $p_1(z) \geq C||z| - a|^\beta$ for all $z \in [-a, a]$, under the assumptions of Corollary 3.6, 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 $\mathcal{C}^b_{n,\alpha}$ in (2.19) on a compact subset of $\mathbb{R}$ in the following theorem.

Theorem 3.7. 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 $n^{-\alpha}(\inf_{|z| \leq n} p_1(z))^{-1} = O(1)$. Given any $D_n = O(n^\beta)$ for $\beta < 1/(10\alpha \vee 5)$, if $s = O(1)$, $m \asymp n^p$ for $p \in (1/5, (3-2\beta)/13)$, $h \asymp n^{-\delta}$ for $\delta \in (5p-1, (1-3p)/2-\beta)$, λ 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

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

In particular, the confidence band $\mathcal{C}^b_{n,\alpha}$ is asymptotically honest on any interval $[-D_n, D_n]$ with $D_n = O(n^\beta)$ for $\beta < 1/(10\alpha \vee 5)$, i.e.,

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

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

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 $\mathcal{S}$ 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 $\boldsymbol{x}_{-1} = (x_2, \ldots, x_d)^T \in \mathcal{X}^{d-1}$, if $x_1 \in (z - \delta_0, z + \delta_0)$, we have the approximation

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

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 $\mathcal{S}_z = \{j \in [d] : f_{jz}(\cdot) \not\equiv 0\}$. We call the function class containing functions satisfying Definition 4.1 the ATLAS model and denote it as $\mathcal{A}_d(s)$.

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

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

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

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

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 dimension is fixed. It has also been considered as compound functional model proposed in Dalalyan et al. (2014)

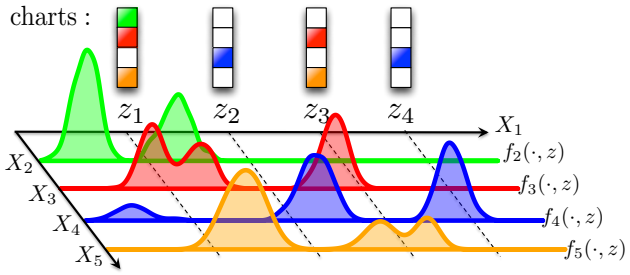

Figure 1: The illustration of ATLAS. As the longitude variable X_1 changes as $X_1 \in \{z_1, z_2, z_3, z_4\}$, the sparsity patterns of the charts are different. By fixing the lattitude variable X_j for $j = 2, \ldots, 5$, the values of charts $f_j(\cdot, z)$ change with z. Under the sparsity assumption,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

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 et al. (2014) must have fixed support. The following example gives another subset of ATLAS model.

Example 4.2. Consider a d-dimensional function with the structure

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

where $a_j(\cdot), f_k(\cdot) \in \mathcal{H}(2, L)$ for all $k \in [d]$ and $j \geq 2$. Moreover, for any fixed $z \in \mathcal{X}$, at most s of $\{a_j(z)\}_{j \geq 2}$ 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 $j = 2, \ldots, d$ and let $L(z, \boldsymbol{x}_{-1}) = \sum_{j \geq 2} a_j'(z) f_j(x_j)$. Then for any $x_1 \in (z - \delta_0, z + \delta_0)$ and $\boldsymbol{x}_{-1} \in \mathcal{X}^{d-1}$, we have

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

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 $j \geq 2$, 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 4.2 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 $j \geq 2$, we can write (4.4) as

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

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 et al. (2012), while Wei et al. (2011) study high dimensional 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 $(\widehat{\alpha}_z, \widehat{\boldsymbol{\beta}}_z)$ be the output of (2.7), we estimate the true functions $f_1(z)$ and $f_j(x_j, z)$ by

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

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 material. 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 3.2 and the property of confidence band in Theorem 3.7 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 $K_{\mathrm{quad}}(u) = (15/16) \cdot (1-u^2)^2 \mathbb{1}(|u| < 1)$ as the kernel function in (2.7).

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

$$f_1(t) = 6\big(0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3(\sin(2\pi t))^2 + 0.4(\cos(2\pi t))^3 + 0.5(\sin(2\pi t))^3\big),$$
$$f_2(t) = 3(2t-1)^2, \quad f_3(t) = 5t, \quad f_4(t) = 4\sin(2\pi t)/(2-\sin(2\pi t)).$$

The model is considered by Zhang and Lin (2006), Meier et al. (2009), and Huang et al. (2010). Let $W_1, \ldots, W_d$ and U follow i.i.d. Uniform$[0,1]$ and

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

The data sample $X_{1j}, \ldots, X_{nj}$ are i.i.d. 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 i.i.d. $N(0, 1.5^2)$. Let the dimension $d = 600$ and 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 γ in (2.16) is set to be $\gamma = 0.05 \log d\sqrt{m/nh}$. The tuning parameter λ and bandwidth h are chosen by cross validation according to the BIC criterion defined as

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

where RSS is the residual sums of squares and the degrees of freedom is defined as $\mathrm{df} = \widehat{s} \cdot m$ with $\widehat{s}$ 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

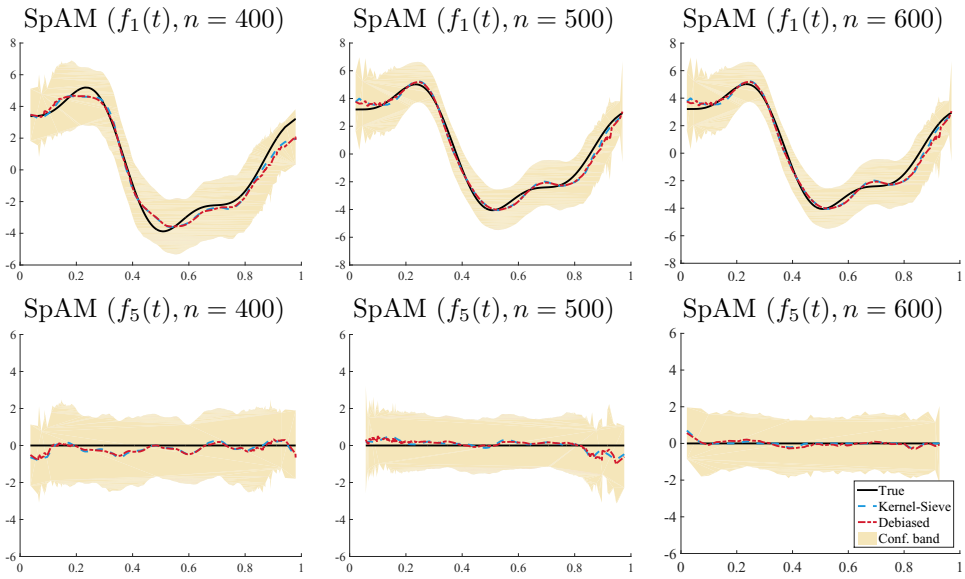

Figure 2: Kernel-sieve hybrid estimators for the $d = 600$ dimensional SpAM model $Y = \sum_{j=1}^{4} f_j(X_j) + \varepsilon$, for $n = 400, 500, 600$ and the noise $\varepsilon \sim N(0, 1.5^2)$. The confidence bands at significant level 95% cover $f_1(t)$ on the first row and $f_5(t) = 0$ on the second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

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) does 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) in order 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 $\mathcal{C}^b_{n,\alpha}$ is defined as $\int_{z\in\mathcal{X}} 2\widehat{c}_n(\alpha)(nh)^{-1/2}\widehat{\sigma}_n(z)dz$. 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.

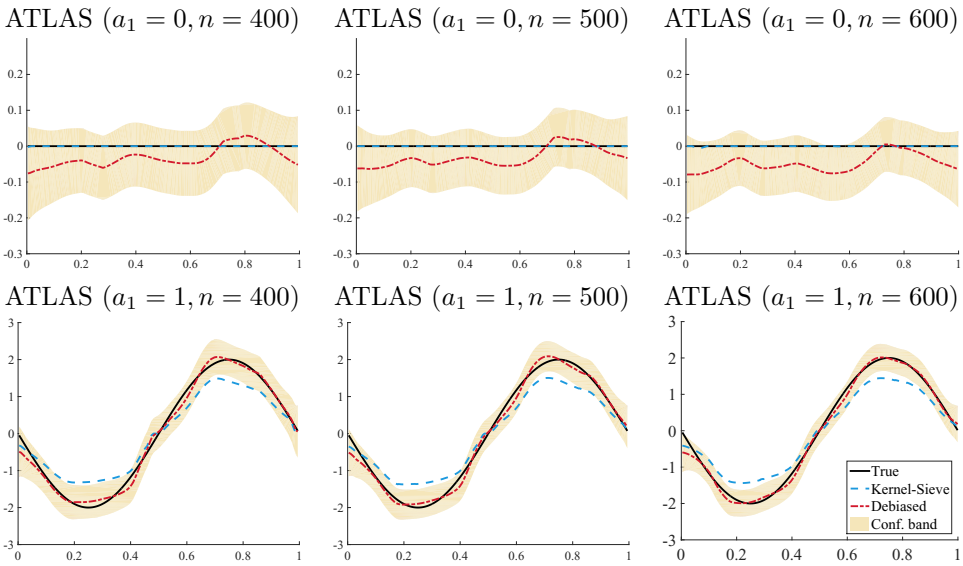

Figure 3: Kernel-sieve hybrid estimators for the $d = 600$ dimensional ATLAS model $Y = a_1 f_1(X_1) + \sum_{j=2}^{4} a_j(X_1) f_j(X_j) + \varepsilon$, for $n = 400, 500, 600$ and the noise $\varepsilon \sim N(0, 1.5^2)$. The confidence bands at significant level 95% cover $f_1^* = a_1 f_1$ for $a_1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

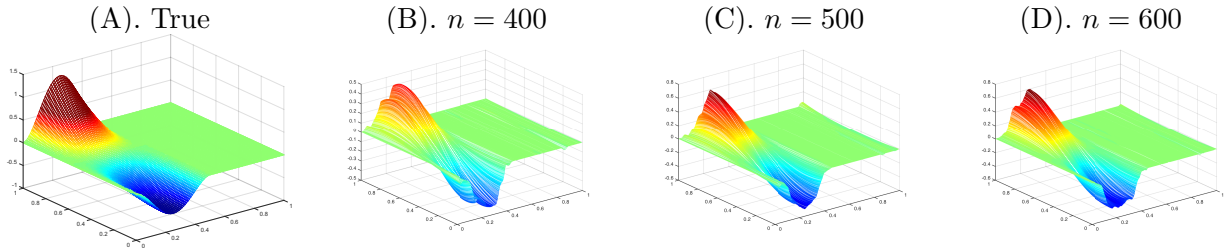

Figure 4: Kernel-sieve hybrid estimators for the two dimensional surface $a_2(x_1) f_2(x_2)$.

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

$$f_1(t) = -2\sin(2\pi t), \quad f_2(t) = t^2 - 1/3, \quad f_3(t) = t - 1/2, \quad f_4(t) = e^t + e^{-1} - 1;$$

$$a_1 \in \{0, 1\}, \quad a_2(t) = 2K_{\text{quad}}(4t - 1), \quad a_2(t) = 3\cos(2\pi t), \quad a_3(t) = 4.$$

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 $i = 1, \ldots, n$ with $\sigma = 1.5$. This ATLAS model is constructed based on the synthetic example in Ravikumar et al. (2009) by adding $a_j(t)$'s according to Example 4.2.

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

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

The covariates X_{ij} are independently and identically generated from Uniform$[0, 1]$ distributions for $i = 1, \ldots, n$ and $j = 1, \ldots, d$. It can be checked that this model follows the identifiability condition in (4.3). According to the argument in Example 4.2, 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 λ and h through cross validation by minimizing the BIC criterion. The confidence bands are constructed at the significance level 95% and the quantile estimator $\widehat{c}_n(\alpha)$ is computed by bootstrap with 500 repetitions. The coverage probability is computed via the same method as in the previous example. The numerical results are reported in Figure 3 and Table 1.

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

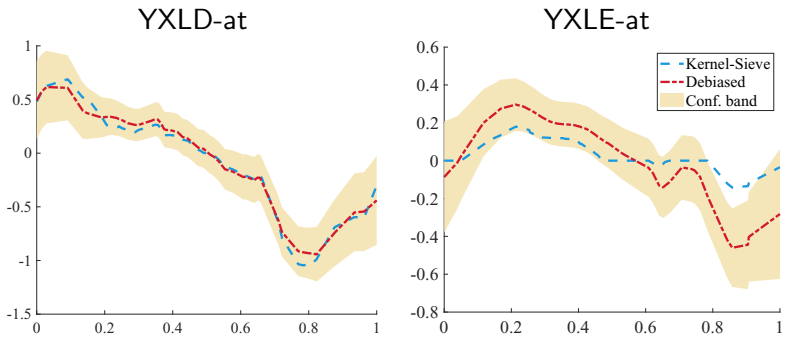

Figure 5: Kernel-sieve hybrid estimators for the riboflavin dataset using ATLAS model.

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 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 et al. (2014). The response variable Y represents the logarithm of the riboflavin production rate. The covariates are the logarithm of gene expression levels with dimension $d = 4,088$ and sample size $n = 71$. van de Geer et al. (2014), Bühlmann et al. (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 et al. (2014) finds the gene YXLD-at and Javanmard and Montanari (2014) finds two genes YXLD-at and YXLE-at to be significant. In this paper, 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.

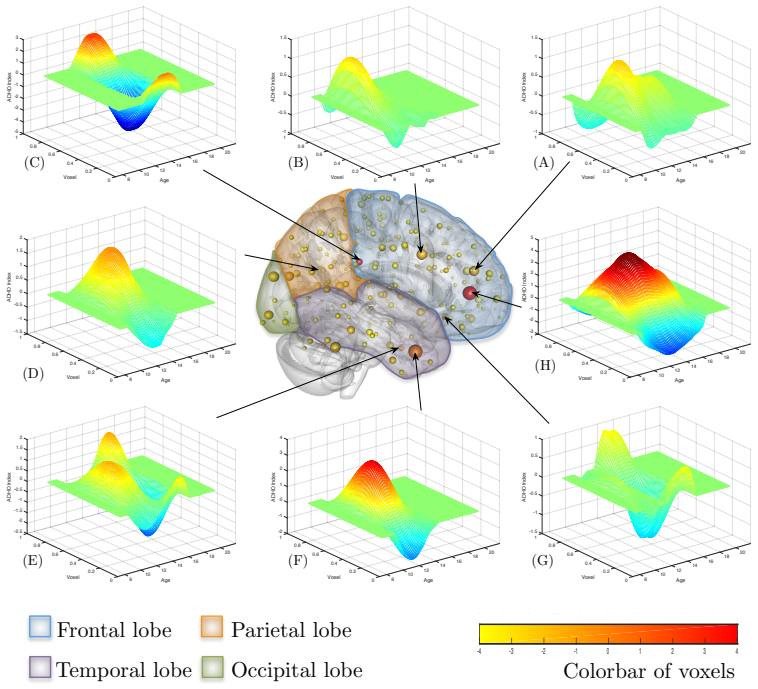

Figure 6: The estimated surfaces of first eight voxels with largest maximum norms. The radii of the balls in the brain represent the duration the voxels being active and the colors represent the maximum norms of the surfaces, whose corresponding values are indicated by the colorbar on the right bottom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

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 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. In order 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

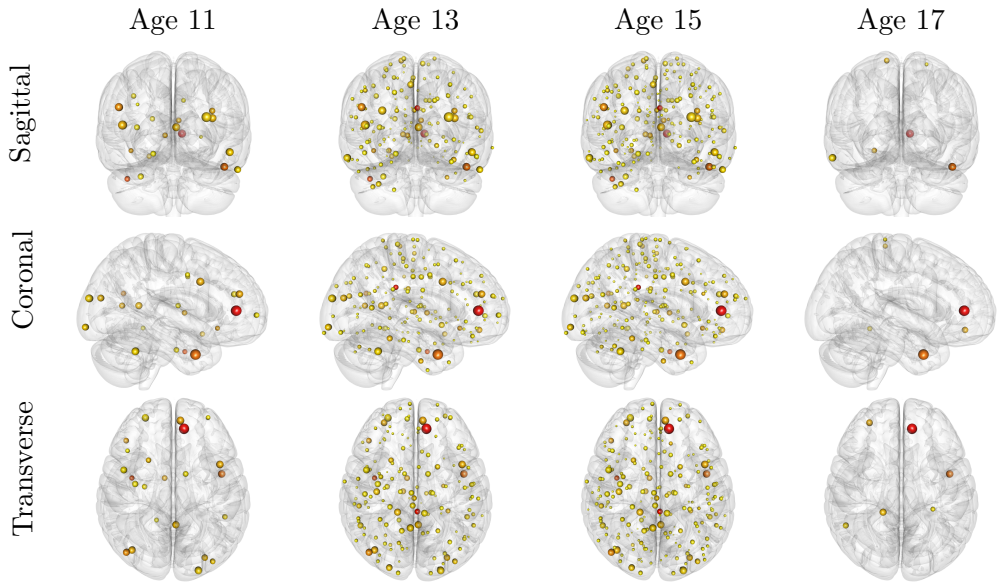

Figure 7: Active voxels varying with age. Each column shows the active voxels at each age. The radii and colors of the balls in a brain represent the duration and maximum norms of the active voxels as in Figure 6.

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 to 276 scans and all the scans are treated as independent observations.

The results of the regression are illustrated in Figure 6 and Figure 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 paper, we consider a novel nonparametric model, ATLAS, which is a generalization of the sparse additive 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

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

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.

References

Avalos, M., Grandvalet, Y. and Ambroise, C. (2007). Parsimonious additive models. *Comput. Stat. Data Anal.* **51** 2851–2870.

Belloni, A., Chernozhukov, V. and Hansen, C. B. (2013a). Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650.

Belloni, A., Chernozhukov, V. and Kato, K. (2013b). Robust inference in high-dimensional approximately sparse quantile regression models. *arXiv preprint arXiv:1312.7186* .

Belloni, A., Chernozhukov, V. and Kato, K. (2015). Uniform post selection inference for LAD regression and other Z-estimation problems. *Biometrika* **102** 77–94.

Belloni, A., Chernozhukov, V. and Wei, Y. (2013c). Honest confidence regions for logistic regression with a large number of controls. *arXiv preprint arXiv:1304.3969* .

Berhane, K. and Tibshirani, R. J. (1998). Generalized additive models for longitudinal data. *Canadian Journal of Statistics* **26** 517–535.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.* **2** 1224–1241.

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. et al. (2010). Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences* **107** 4734–4739.

Bousquet, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique* **334** 495–500.

Bühlmann, P., Kalisch, M. and Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Ann. Rev. Stat. & Appl.* **1** 255–278.

Bühlmann, P. and van de Geer, S. A. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics, Springer, Heidelberg. Methods, Theory and Applications.

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. *Ann. Stat.* **41** 1232–1259.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Stat.* **41** 2786–2819.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014a). Anti-concentration and honest, adaptive confidence bands. *Ann. Stat.* **42** 1787–1818.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014b). Gaussian approximation of suprema of empirical processes. *Ann. Stat.* **42** 1564–1597.

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.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Stat.* **31** 1852–1884.

Comminges, L. and Dalalyan, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Stat.* **40** 2667–2696.

Conners, C. K. (2008). *Conners 3rd Edition (Conners 3),*. Multi-Health Systems.

Dalalyan, A. S., Ingster, Y. and Tsybakov, A. B. (2014). Statistical inference in compound functional models. *Probab. Theory Related Fields* **158** 513–532.

de Boor, C. (2001). *A practical guide to splines*, vol. 27 of *Applied Mathematical Sciences*. Revised ed. Springer-Verlag, New York.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1. Cambridge University Press.

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.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Stat.* **21** 196–216.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Fan, J. and Jiang, J. (2005). Nonparametric inferences for additive models. *J. Am. Stat. Assoc.* **100** 890–907.

Fan, J. and Yao, Q. (2008). *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics* 1491–1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Stat.* **27** 715–731.

Farrell, M. H. (2013). Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint arXiv:1309.4686* .

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *J. Am. Stat. Assoc.* **76** 817–823.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* **24** 726–748.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *J. Multivar. Anal.* **29** 163–179.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B. Methodological* **55** 757–796. With discussion and a reply by the authors.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, vol. 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Stat.* **38** 2282–2313.

Javanmard, A. and Montanari, A. (2013). Nearly optimal sample size in hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1311.0274* .

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15** 2869–2909.

Kato, K. (2012). Two-step estimation of high dimensional additive models. *ArXiv e-prints, arXiv:1207.5313* .

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*, vol. 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg.

Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660–3695.

Kozbur, D. (2015). Inference in additively separable models with a high-dimensional set of conditioning variables. *ArXiv e-prints, arXiv:1503.05436* .

Lafferty, J. D. and Wasserman, L. A. (2008). Rodeo: sparse, greedy nonparametric regression. *Ann. Stat.* **36** 28–63.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2013). Exact post-selection inference with the lasso. *ArXiv e-prints, arXiv:1311.6238* .

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Stat.* **34** 2272–2297.

Liu, H. and Yu, B. (2013). Asymptotic properties of Lasso+mLS and Lasso+Ridge in sparse high-dimensional linear regression. *Electron. J. Stat.* **7** 3124–3169.

Lockhart, R., Taylor, J. E., Tibshirani, R. J. and Tibshirani, R. J. (2014). A significance test for the lasso. *Ann. Stat.* **42** 413–468.

Lopes, M. (2014). A residual bootstrap for high-dimensional regression with near low-rank designs. In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 3239–3247.

Lou, Y., Bien, J., Caruana, R. and Gehrke, J. (2014). Sparse partially linear additive models. *ArXiv e-prints, arXiv:1407.4729* .

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.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis* **17** 571–599.

Meier, L., van de Geer, S. A. and Bühlmann, P. (2009). High-dimensional additive modeling. *Ann. Stat.* **37** 3779–3821.

Meinshausen, N. (2013). Group-bound: confidence intervals for groups of variables in sparse high-dimensional regression without assumptions on the design. *arXiv preprint arXiv:1309.3489* .

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. B* **72** 417–473.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *J. Am. Stat. Assoc.* **104**.

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. *Stat. Sci.* **27** 538–557.

Ning, Y. and Liu, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45** 158–195.

Nze, P. A. and Doukhan, P. (2004). Weak dependence: models and applications to econometrics. *Econometric Theory* **20** 995–1045.

Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Stat.* **25** 186–211.

Peligrad, M. (1992). Properties of uniform consistency of the kernel estimators of density and regression functions under dependence assumptions. *Stochastics: An International Journal of Probability and Stochastic Processes* **40** 147–168.

Petersen, A., Witten, D. M. and Simon, N. (2014). Fused lasso additive model. *ArXiv e-prints, arXiv:1409.5391* .

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. et al. (2011). Functional network organization of the human brain. *Neuron* **72** 665–678.

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427.

Ravikumar, P., Lafferty, J. D., Liu, H. and Wasserman, L. A. (2009). Sparse additive models. *J. R. Stat. Soc. B* **71** 1009–1030.

Richtárik, P. and Takáč, M. (2014). Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.* **144** 1–38.

Rosasco, L., Villa, S., Mosci, S., Santoro, M. and Verri, A. (2013). Nonparametric sparsity and regularization. *J. Mach. Learn. Res.* **14** 1665–1714.

Sardy, S. and Tseng, P. (2004). AMlet, RAMlet, and GAMlet: automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *J. Comp. Graph. Stat.* **13** 283–309.

Schumaker, L. L. (2007). *Spline functions: basic theory.* 3rd ed. Cambridge Mathematical Library, Cambridge University Press, Cambridge.

Shah, R. D. and Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. B* **75** 55–80.

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* **104** 19649–19654.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Stat.* **13** 689–705.

Sun, J. and Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics* 1328–1345.

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 preprint arXiv:1401.3889* .

van de Geer, S. A., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.* **42** 1166–1202.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.

Wahl, M. (2014). Variable selection in high-dimensional additive models based on norms of projections. *ArXiv e-prints, arXiv:1406.0052* .

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Wasserman, L. A. and Roeder, K. (2009). High-dimensional variable selection. *Ann. Stat.* **37** 2178–2201.

Wei, F., Huang, J. and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statist. Sinica* **21** 1515–1540.

Xu, M., Chen, M. and Lafferty, J. D. (2014). Faithful variable screening for high-dimensional convex regression. *ArXiv e-prints, arXiv:1411.1805* .

Yang, Y. and Tokdar, S. T. (2014). Minimax-optimal nonparametric regression in high dimensions. *ArXiv e-prints, arXiv:1401.7278* .

Zhang, C.-H. and Zhang, S. S. (2013). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. B* **76** 217–242.

Zhang, H. H. and Lin, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. *Stat. Sinica* **16** 1021–1041.

Zhang, W. and Peng, H. (2010). Simultaneous confidence band and hypothesis test in generalised varying-coefficient models. *J. Multivar. Anal.* **101** 1656–1680.

Zhang, X. and Wang, J.-L. (2015). Varying-coefficient additive models for functional data. *Biometrika* **102** 15–32.

Zhu, H., Li, R. and Kong, L. (2012). Multivariate varying coefficient model for functional responses. *Ann. Stat.* **40** 2634–2666.