

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 $\mathbf{X} = (X_1, \dots, X_d)^T$ without imposing strong assumptions. Consider a model

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

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

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

where μ is a constant and $f_j(\cdot)$, $j = 1, \dots, d$, 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 $\{1, \dots, n\}$ and let $\mathbb{1}\{\cdot\}$ denote the indicator function. For a vector $\mathbf{a} \in \mathbb{R}^d$, we let $\text{supp}(\mathbf{a}) = \{j \mid a_j \neq 0\}$ 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 ℓ_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, $\|\mathbf{a}\|_0 = |\text{supp}(\mathbf{a})|$ 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 $\mathbf{a} = (\mathbf{a}_{\mathcal{G}_1}, \dots, \mathbf{a}_{\mathcal{G}_g})^T$, where $\mathcal{G}_1, \dots, \mathcal{G}_g \subset [d]$ 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 $\{1, \dots, j-1, j+1, \dots, d\}$ as $\setminus j$ and the vector $\mathbf{a}_{\setminus j} = (a_1, \dots, a_{j-1}, a_{j+1}, \dots, a_d)^T$. 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 $\alpha_n = O(\beta_n)$ 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 $\alpha_n \beta_n^{-1} \xrightarrow{n \rightarrow \infty} 0$. 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 $\mathbf{X} = (X_1, \dots, X_d)^T$ 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

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

Define the sparse additive functions class

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

Let $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ be n independent random samples of (\mathbf{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 $\{\phi_1, \dots, \phi_m\}$ 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

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

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

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

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) is 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

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

where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$. Let $\beta = (\beta_2^T, \dots, \beta_d^T)^T \in \mathbb{R}^{(d-1)m}$ with $\beta_j = (\beta_{j1}, \dots, \beta_{jm})^T \in \mathbb{R}^m$ be the coefficients of B-spline basis functions. The penalized kernel-sieve hybrid estimator at $z \in \mathcal{X}$ is defined as

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

where the penalty function is

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

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

$\hat{f}_j(x) = \sum_{k=1}^m \psi_{jk}(x) \hat{\beta}_{jk;z}$ for $j \geq 2$. Based on $\hat{\alpha}_z, \hat{\beta}_z$, we also estimate the d -dimensional function $f(z, x_2, \dots, x_d) = f_1(z) + \sum_{j=2}^d f_j(x_j)$ by

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

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 2.2. The estimators $\hat{\alpha}_z$ and $\hat{\beta}_z$ are estimating different quantities. Notice that $\hat{\alpha}_z$ estimates the scalar $f_1(z)$, while $\hat{\beta}_z$ 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 $\hat{\alpha}_z$ and $\hat{\beta}_z$ 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

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

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

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

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 $m \asymp n^{1/6}$. 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 $\hat{\beta}_1^{\text{sieve}}$, 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 $\hat{\alpha}_z$ of $f_1(z)$. This one dimensional 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 $j \geq 2$. We also write $\Psi = (\Psi_{\bullet 1}, \dots, \Psi_{\bullet d})$,

Algorithm 1 Randomized coordinate descent for group Lasso

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

Let $\beta_+^{(t)} = (\beta_1^{(t)}, \beta_2^{(t)T}, \dots, \beta_j^{(t)T})^T$.

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

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

$$T(\beta_j^{(t)}) = \arg \min_{\theta \in \mathbb{R}^{\dim(\beta_j)}} \left\{ \frac{\mu}{2} \|\theta\|_2^2 + \langle \nabla_j \mathcal{L}_z(\beta_+^{(t)}), \theta \rangle + \lambda_j \|\theta + \beta_j^{(t)}\|_2 \right\}. \quad (2.14)$$

Update $\beta_j^{(t+1)} = \beta_j^{(t)} + T(\beta_j^{(t)})$.

end for

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 $j \geq 2$. We further denote

$$\begin{aligned} \mathbf{Y} &= (Y_1 - \bar{Y}, \dots, Y_n - \bar{Y})^T \in \mathbb{R}^n, \quad \beta_+ = (\alpha, \beta^T)^T \in \mathbb{R}^{1+(d-1)m}, \beta_+^* = (f_1^*(z), \beta^{*T})^T \in \mathbb{R}^{1+(d-1)m} \\ &\text{and } \mathbf{W}_z = \text{diag}(K_h(X_{11} - z), \dots, K_h(X_{n1} - z)) \in \mathbb{R}^{n \times n}. \end{aligned} \quad (2.12)$$

To unify the notation in our algorithm, we also write $\beta_+ = (\beta_1, \beta_2^T, \dots, \beta_d^T)^T$, where $\beta_1 = \alpha$ and $\beta = (\beta_2^T, \dots, \beta_d^T)^T$. 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

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

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 $\nabla_j \mathcal{L}_z(\beta_+) := \partial \mathcal{L}_z(\beta_+)/\partial \beta_j$ denotes the gradient. Suppose the result of the t -th iteration is $\beta_+^{(t)}$. In the next iteration, we randomly choose one coordinate j_{t+1} from $\{1, \dots, d\}$ and update the $\beta_{j_t}^{(t)}$. Each update in (2.14) can be obtained in a closed form as

$$T(\beta_j^{(t)}) = \mathcal{T}_{\lambda_j/\mu} \left(\beta_j^{(t)} - \frac{1}{L} \nabla_j \mathcal{L}_z(\beta_+^{(t)}) \right) - \beta_j^{(t)}, \quad (2.15)$$

where μ is certain regularized constant and \mathcal{T}_λ 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 $\hat{\alpha}_z$ 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 $\mathcal{C}_n = \{\mathcal{C}_n(z) = [c_L(z), c_U(z)] \mid z \in \mathcal{X}\}$. 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 $\hat{\alpha}_z$ that reduces the bias introduced by (2.7).

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

$$\hat{\boldsymbol{\theta}}_z = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^{(d-1)m+1}} \boldsymbol{\theta}^T \hat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta}, \quad \text{subject to} \quad \|\hat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta} - \mathbf{e}_1\|_{2,\infty} \leq \gamma, \quad (2.16)$$

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

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

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

empirical process by the Gaussian multiplier process

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

where ξ_1, \dots, ξ_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

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

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 $p(x_1, \dots, x_d)$ denote the joint density of $\mathbf{X} = (X_1, \dots, X_d)$ and let $p_j(x_j)$ denote the marginal density of X_j , for $j \in [d]$. Furthermore, let $p_{jkl}(x_j, x_k, x_\ell)$ 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 $p(x_1, \dots, x_d)$ 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 $j, k \in \{2, \dots, d\}$.

(A2) (Kernel function) The kernel $K(u)$ is a continuous function with a bounded support satisfying

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

(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

$$\mathbb{C}_\beta^{(\kappa)}(J) = \left\{ \beta_+ = (\alpha, \beta^T)^T \mid \sum_{j \notin J, j \neq 1} \|\beta_j\|_2 \leq \kappa \sum_{j \in J, j \neq 1} \|\beta_j\|_2 + \kappa \sqrt{m} |\alpha| \right\}. \quad (3.1)$$

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

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

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

(A5) The nonparametric function $f(x_1, \dots, x_d) \in \mathcal{K}_d(s)$ 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 $j, k \in \{2, \dots, d\}$. 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

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

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 $\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma})$ where $\Sigma_{jj} = 1$ and $\Sigma_{jk} = \rho$ for all $2 \leq j < k \leq d$, we can bound the densities as $p_1(x_1) \leq 1, p(x_j|x_1) \leq (1 - \rho^2)^{-1/2}$ and $p(x_j, x_k|x_1) \leq (1 - \rho)^{-1}(1 + 2\rho)^{-1/2}$ for all $2 \leq j < k \leq d$. 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 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 $\mathbf{\Sigma}_z$ is the expectation of the Hessian matrix of the loss function $\mathcal{L}(\beta_+)$. We require $\mathbf{\Sigma}_z$ 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 β_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

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

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

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

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

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

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

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, $p_1(z) \rightarrow 0$ as $|z| \rightarrow \infty$. In this case, (3.6) does not provide a valid ρ_{\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 $h = o(1)$, $m \rightarrow \infty$ as $n \rightarrow \infty$, and we set

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

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

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

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

$$\|\hat{f} - f\|_2 \leq \rho_{\min}^{-1} s \sqrt{m} \lambda \quad (3.9)$$

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 $\|\hat{f} - f\|_2^2 = O_P(n^{-2/3} \log(dn))$. This convergence rate is slower than the optimal rate $O_P(n^{-4/5} + \log d/n)$ for estimating the sparse additive model (Raskutti 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 $\|\hat{f} - f\|_2^2 = O_P(n^{-2/3} \log(dn))$ 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 $T_n = \sup_{z \in \mathcal{X}} \max_{j \in [d]} \frac{1}{n} \|\Psi_{\bullet,j}^T \mathbf{W}_z \varepsilon\|_2 = O_P\left(\sqrt{\log(dn)/(nh)}\right)$, where \mathbf{W}_z is defined in (2.2). In comparison, Huang et al. (2010) only need to bound $T'_n = \sup_{z \in \mathcal{X}} \max_{j \in [d]} \frac{1}{n} \|\Psi_{\bullet,j}^T \varepsilon\|_2 = O_P\left(\sqrt{\log(dn)/n}\right)$ (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 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 ρ_{\min} is defined in (3.2). We assume that the density functions of \mathbf{X} satisfies

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

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 Σ_z defined in Assumption (A3). Due to the complicated definition of Σ_z , there is no straightforward interpretation of Σ_z^{-1} and any assumption imposed on Σ_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 $p(\mathbf{x})$ with unbounded support such that

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

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

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

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

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

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

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 $\mathcal{C}_{n,\alpha}^b$ in (2.19) is

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

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

$$\liminf_{n \rightarrow \infty} \mathbb{P}(f_1(z) \in \mathcal{C}_{n,\alpha}^b(z), \text{ for all } z \in \mathcal{X}) \geq 1 - \alpha.$$

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

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

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

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

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 $\text{MSE}(\hat{m}(z))$ will diverge if $p(z) \rightarrow 0$. The intuition is we have a kernel density estimator of $p(z)$ in the denominator of $\hat{m}(z)$ in (3.13). Therefore, in order to control the uniform rate $\sup_{z \in \mathcal{X}} |\hat{m}(z) - m(z)|$, 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 \hat{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 $b_n = \inf_{|z| \leq D_n} p_1(z) > 0$, then $\sup_{|z| \leq D_n} |\hat{m}(z) - m(z)| = O_P(b_n^{-1}(h^2 + \sqrt{\log n/nh}))$. 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 $b_n = \inf_{|z| \leq D_n} p_1(z) > 0$, we have

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

The corollary can be proved by applying Proposition 3.1 to Theorem 3.2. Assumption (A3) is not required here because ρ_{\min} could be zero when $\mathcal{X} = \mathbb{R}$. Proposition 3.1 characterizes how ρ_{\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

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

We can also show the coverage probability of the confidence band $\mathcal{C}_{n,\alpha}^b$ 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 $\mathcal{C}_{n,\alpha}^b$ has

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

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

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

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 $f(x_1, \dots, x_d)$ has a local sparse additive approximation for x_1 if for any $z \in \mathcal{X}$, there exist functions $f_{1z}(\cdot), \dots, f_{dz}(\cdot) \in \mathcal{H}(2, L)$, 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 $\mathbf{x}_{-1} = (x_2, \dots, x_d)^T \in \mathcal{X}^{d-1}$, if $x_1 \in (z - \delta_0, z + \delta_0)$, we have the approximation

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

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

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

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

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

We call X_1 the longitude variable and the functions $f_2(\cdot, z), \dots, f_d(\cdot, z)$ 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)

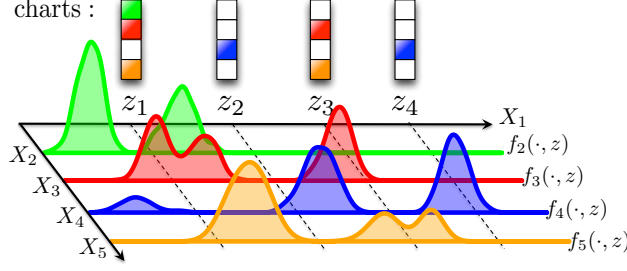


Figure 1: The illustration of ATLAS. As the longitude variable X_1 changes as $X_1 \in \{z_1, z_2, z_3, z_4\}$, the sparsity patterns of the charts are different. By fixing the latitude variable X_j for $j = 2, \dots, 5$, the values of charts $f_j(\cdot, z)$ change with z . Under the sparsity assumption,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

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

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 $j = 2, \dots, d$ and let $L(z, \mathbf{x}_{-1}) = \sum_{j \geq 2} a'_j(z) f_j(x_j)$. Then for any $x_1 \in (z - \delta_0, z + \delta_0)$ and $\mathbf{x}_{-1} \in \mathcal{X}^{d-1}$, 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 $j \geq 2$, 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

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

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 $(\hat{\alpha}_z, \hat{\beta}_z)$ be the output of [\(2.7\)](#), we estimate the true functions $f_1(z)$ and $f_j(x_j, z)$ by

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

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 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 $f(x_1, \dots, x_d) \in \mathcal{A}_d(s)$, 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 $K_{\text{quad}}(u) = (15/16) \cdot (1 - u^2)^2 \mathbb{1}(|u| < 1)$ as the kernel function in (2.7).

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

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

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

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

The data sample X_{1j}, \dots, 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

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

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

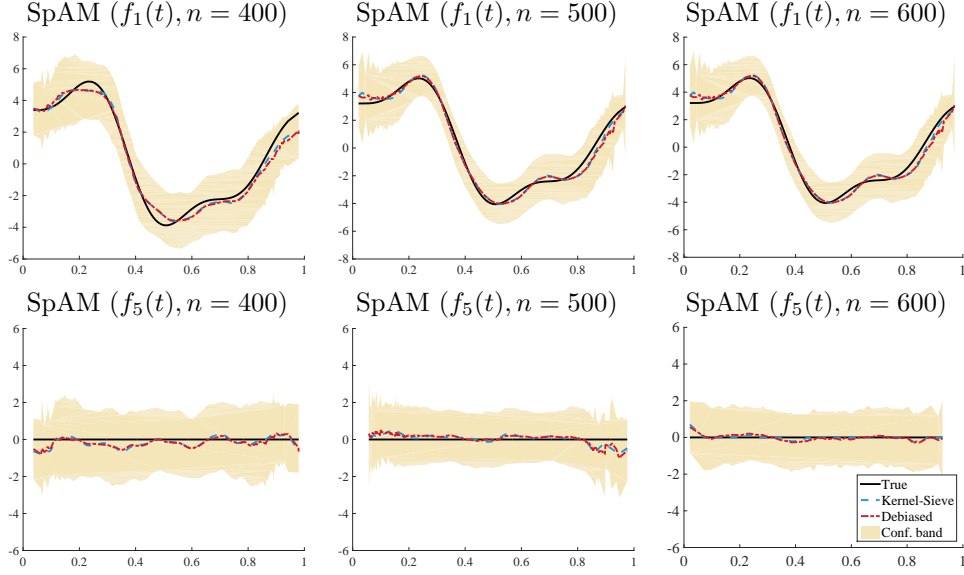


Figure 2: Kernel-sieve hybrid estimators for the $d = 600$ dimensional SpAM model $Y = \sum_{j=1}^4 f_j(X_j) + \varepsilon$, for $n = 400, 500, 600$ and the noise $\varepsilon \sim N(0, 1.5^2)$. The confidence bands at significant level 95% cover $f_1(t)$ on the first row and $f_5(t) = 0$ on the second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

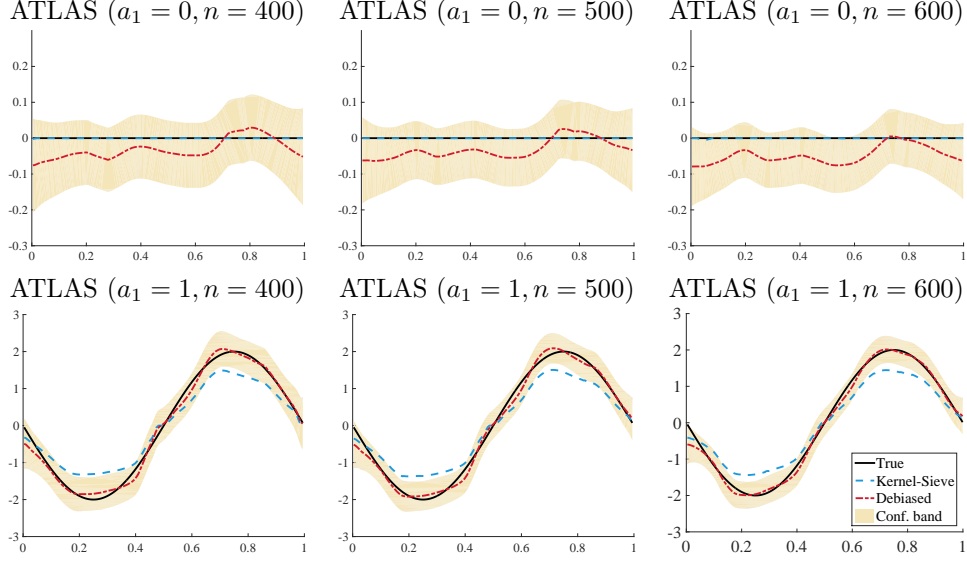


Figure 3: Kernel-sieve hybrid estimators for the $d = 600$ dimensional ATLAS model $Y = a_1 f_1(X_1) + \sum_{j=2}^4 a_j(X_1) f_j(X_j) + \varepsilon$, for $n = 400, 500, 600$ and the noise $\varepsilon \sim N(0, 1.5^2)$. The confidence bands at significant level 95% cover $f_1^* = a_1 f_1$ for $a_1 \in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

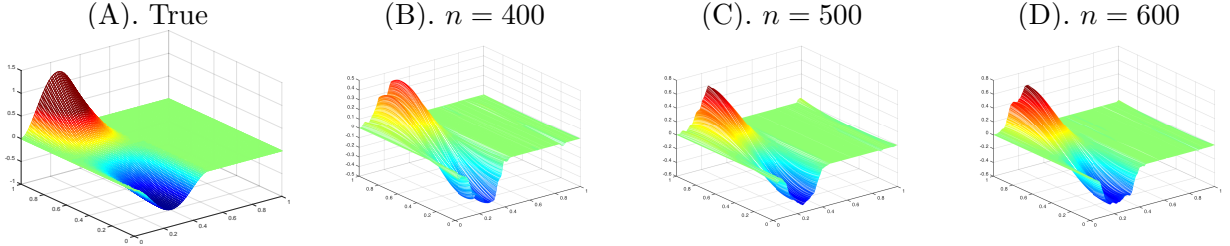


Figure 4: Kernel-sieve hybrid estimators for the two dimensional surface $a_2(x_1)f_2(x_2)$.

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

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

$$a_1 \in \{0, 1\}, \quad a_2(t) = 2K_{\text{quad}}(4t - 1), \quad a_3(t) = 3 \cos(2\pi t), \quad a_4(t) = 4.$$

Here two values of $a_1 \in \{0, 1\}$ correspond to two scenarios that the true function is zero and nonzero. The noise $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, \dots, n$ 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 $i = 1, \dots, n$ and $j = 1, \dots, d$. 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 $\hat{c}_n(\alpha)$ 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.

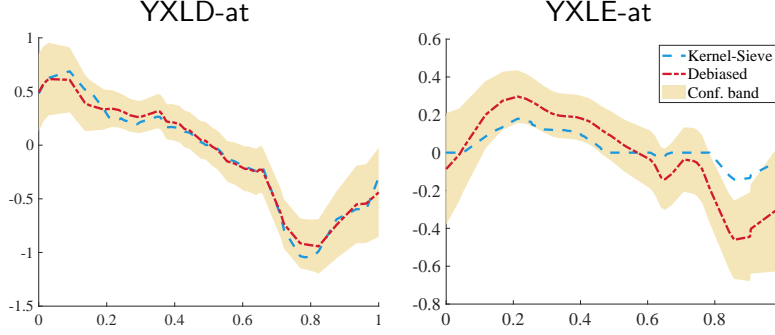


Figure 5: Kernel-sieve hybrid estimators for the riboflavin dataset using ATLAS model.

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.

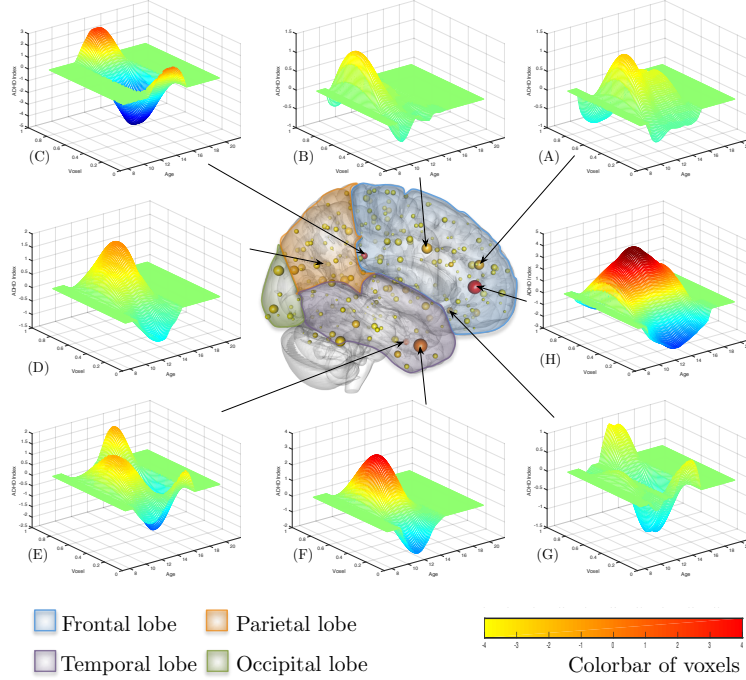


Figure 6: The estimated surfaces of first eight voxels with largest maximum norms. The radii of the balls in the brain represent the duration the voxels being active and the colors represent the maximum norms of the surfaces, whose corresponding values are indicated by the colorbar on the right bottom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

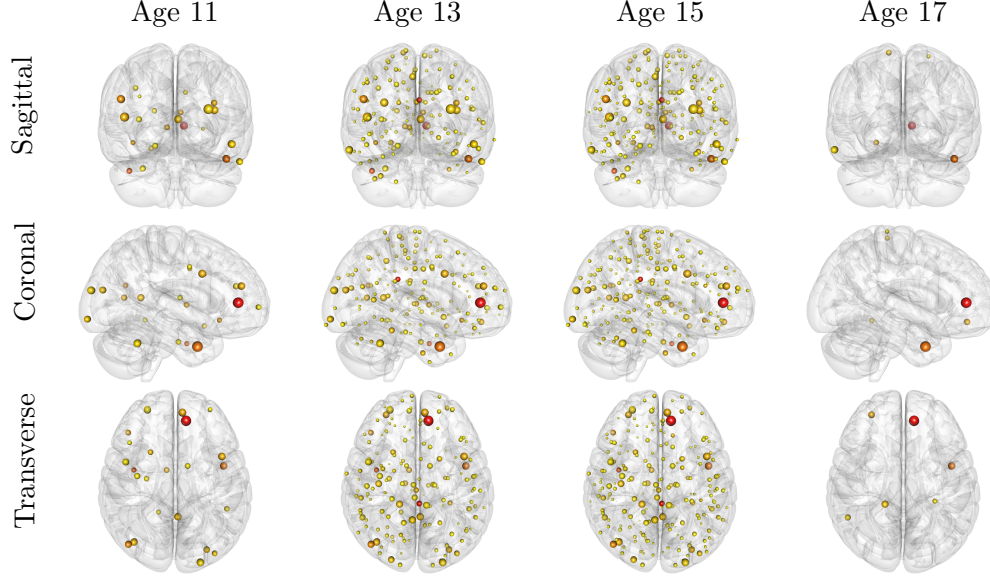


Figure 7: Active voxels varying with age. Each column shows the active voxels at each age. The radii and colors of the balls in a brain represent the duration and maximum norms of the active voxels as in Figure 6.

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 $\{\hat{f}_j(x_j, x_1)\}_{j=1}^d$ 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

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

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.

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 Mathématique* **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). *Connors 3rd Edition (Connors 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.

Supplementary material to

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

Junwei Lu^{*} Mladen Kolar[†] Han Liu[‡]

Abstract

This document contains the supplementary material to the paper “Kernel Meets Sieve: Post-Regularization Confidence Bands for Sparse Additive Model”. All the proofs in the supplementary material assume that true nonparametric function $f(x_1, \dots, x_d)$ belongs to the ATLAS model $\mathcal{A}_d(s)$. In Appendix A, we outline the proof of Theorem 3.2. In Appendix B, we introduce an accelerated method to derive our estimator. Appendix C proves the validity of bootstrap confidence bands. In Appendix D, we prove Propositions 3.1. Appendix E collects the technical lemmas on the estimation rate. Appendix F states some auxiliary results on the bootstrap confidence bands. In Appendix H, we list several useful results on empirical processes.

A Proof of the Statistical Rate of Kernel-Sieve Hybrid Estimator

For all the proofs in the following of the paper (including the supplementary material), we consider the most general case that true nonparametric function $f(x_1, \dots, x_d)$ belongs to the ATLAS model $\mathcal{A}_d(s)$. Since SpAM is a strictly smaller family of $\mathcal{A}_d(s)$, all the proofs apply to $\mathcal{K}_d(s)$ as well.

This section outlines the proof of Theorem 3.2 on the statistical estimation rate of the kernel-sieve hybrid estimator in (2.7). Before presenting the main proof, we list several technical lemmas whose proofs are deferred to Appendix E in the supplementary material.

The following lemma provides the restricted eigenvalue condition on the empirical Hessian matrix of the kernel-sieve hybrid loss in (2.6), which is $\hat{\Sigma}_z = n^{-1} \Psi \mathbf{W}_z \Psi^T$.

^{*}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, USA; Email: mkolar@chicagobooth.edu, [‡]Department of Computer Science, Northwestern University, Evanston, IL 60208; Email: hanliu@northwestern.edu

Lemma A.1. Under Assumptions (A1)-(A5), suppose $\beta \in \mathbb{R}^{(d-1)m}$ and $\alpha \in \mathbb{R}$ satisfy the cone restriction

$$\sum_{j \in S^c} \|\beta_j\|_2 \leq 3 \sum_{j \in S} \|\beta_j\|_2 + 3\sqrt{m}|\alpha|$$

for some index set $S \subset [d]$ with cardinality s . Denote $\theta = (\alpha, \beta^T)^T$. If $s\sqrt{m^3 \log(dm)/(nh)} + sm^2/(nh) = o(1)$, there exists a constant ρ_{\min} such that with high probability,

$$\inf_{z \in \mathcal{X}} \theta^T \widehat{\Sigma}_z \theta \geq \frac{\rho_{\min}}{2m} \|\beta\|_2^2 + \frac{\rho_{\min}}{2} |\alpha|_2^2.$$

The estimation error for the kernel-sieve hybrid estimator comes from three sources: (1) noise ε , (2) approximation error by finite B-spline bases, and (3) approximation error by s local additive functions to the true function. The following lemma provides the rate for the B-spline approximation error, which further illustrates how the number of B-spline basis functions m influences the rate.

Lemma A.2. Recall that $\{f_{jz}\}_{j=1}^d$ are defined in Definition 4.1. Let $\delta_z = (\delta_1(z), \dots, \delta_n(z))^T$ where $\delta_i(z) = \sum_{j=2}^d f_{jz}(X_{ji}) - f_{mj}(X_{ji})$ for $i = 1, \dots, n$, where $f_{mj}(\cdot)$ is defined in (2.4). Under Assumptions (A1)-(A5) there exists a constant $C > 0$ such that the following three inequalities hold with probability at least $1 - 1/n$,

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \delta_z\|_2 \leq C\sqrt{s} \cdot m^{-5/2}, \quad (\text{A.1})$$

$$\sup_{z \in \mathcal{X}} \frac{1}{n} |\Psi_{\bullet 1}^T \mathbf{W}_z \delta_z| \leq C\sqrt{s} \cdot m^{-2}, \quad (\text{A.2})$$

$$\sup_{z \in \mathcal{X}} \frac{1}{n} \|\mathbf{W}_z^{1/2} \delta_z\|_2^2 \leq Csm^{-4}. \quad (\text{A.3})$$

Our next lemma bounds the approximation error of charts under the ATLAS model (4.1). We can see that both the number of bases m and the bandwidth h play a role in the estimation.

Lemma A.3. Let $\xi_z = (\xi_1(z), \dots, \xi_n(z))^T$ and $\zeta_z = (\zeta_1(z), \dots, \zeta_n(z))^T$, where $\xi_i(z) = f_1(X_{1i}) - f_1(z)$ and $\zeta_i(z) = f(X_{1i}, \dots, X_{di}) - \sum_{j=1}^d f_{jz}(X_{ji})$ for $i \in [n]$. Under Assumptions (A1)-(A5) there exists a constant $C > 0$ such that the following three inequalities hold with probability at least

$1 - 1/n$,

$$\begin{aligned} \sup_{z \in \mathcal{X}} \max_{j \in [d]} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z (\boldsymbol{\xi}_z + \boldsymbol{\zeta}_z)\|_2 &\leq C \left(\sqrt{\frac{h \log(dh^{-1})}{n}} + \frac{m^{3/2} \log(dh^{-1})}{n} + \frac{h^2}{\sqrt{m}} \right), \\ \sup_{z \in \mathcal{X}} \frac{1}{n} |\Psi_{\bullet 1}^T \mathbf{W}_z (\boldsymbol{\xi}_z + \boldsymbol{\zeta}_z)| &\leq C \left(h^2 + \sqrt{h/n} \right), \\ \sup_{z \in \mathcal{X}} \frac{1}{n} \|\mathbf{W}_z^{1/2} (\boldsymbol{\xi}_z + \boldsymbol{\zeta}_z)\|_2^2 &\leq Ch^2. \end{aligned}$$

The following lemma quantifies the statistical error arising from the noise ε .

Lemma A.4. Let $T_n = \sup_{z \in \mathcal{X}} \max_{j \in [d]} n^{-1} \|\Psi_{\bullet j}^T \mathbf{W}_z \varepsilon\|_2$, where $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$. Under Assumptions (A1)-(A5) and if $m(nh)^{-1} = o(1)$, there exists a constant $C > 0$ such that with probability at least $1 - 1/n$,

$$T_n \leq C \sqrt{\log(dm^2 h^{-2})/(nh)}.$$

We are now ready to present the main proof of Theorem 3.2.

Proof of Theorem 3.2. We denote $\boldsymbol{\eta}_z = \varepsilon + \boldsymbol{\delta}_z + \boldsymbol{\xi}_z + \boldsymbol{\zeta}_z$ and define the event

$$\mathcal{E} = \left\{ \sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{4}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \boldsymbol{\eta}_z\|_2 \leq \lambda \right\} \cup \left\{ \sup_{z \in \mathcal{X}} \frac{4}{n} |\Psi_{\bullet 1}^T \mathbf{W}_z \boldsymbol{\eta}_z| \leq \lambda \sqrt{m} \right\}.$$

Using Lemma A.2, Lemma A.3 and Lemma A.4, there exist constants c, C such that $\mathbb{P}(\mathcal{E}) \geq 1 - c/n$ if the tuning parameter satisfies the following inequality

$$\lambda \geq C \left(\sqrt{\frac{\log(dm^2 h^{-2})}{nh}} + \sqrt{s} \cdot m^{-5/2} + \sqrt{\frac{h \log(dh^{-1})}{n}} + \frac{m^{3/2} \log(dh^{-1})}{n} + \frac{h^2}{\sqrt{m}} \right). \quad (\text{A.4})$$

In the rest of this proof, we are always conditioning on the event \mathcal{E} .

Denote $S_z := \{j \in \{2, \dots, d\} \mid f_{jz} \neq 0\}$ and $\boldsymbol{\Delta} = \widehat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}_+$, where $\boldsymbol{\Delta}_1 = \widehat{\alpha}_z - f_1(z)$ and $\boldsymbol{\Delta}_j = \widehat{\beta}_j - \beta_j$ for $j \geq 2$. We start by showing that $\boldsymbol{\Delta}$ falls into the cone

$$\mathcal{A}_z := \left\{ \boldsymbol{\Delta} : \sum_{j \in S_z^c} \|\boldsymbol{\Delta}_j\|_2 \leq 3 \sum_{j \in S_z} \|\boldsymbol{\Delta}_j\|_2 + 3\sqrt{m} |\boldsymbol{\Delta}_1| \right\}.$$

Since $\widehat{\beta}_+$ is a minimizer of the objective function,

$$\frac{1}{n} \|\mathbf{W}_z^{1/2}(\mathbf{Y} - \Psi \widehat{\beta}_+)\|_2^2 - \frac{1}{n} \|\mathbf{W}_z^{1/2}(\mathbf{Y} - \Psi \beta_+)\|_2^2 + \lambda \|\widehat{\beta}\|_{1,2} - \lambda \|\beta\|_{1,2} + \lambda \sqrt{m}(|\widehat{\alpha}_z| - |f_1(z)|) \leq 0.$$

On the event \mathcal{E} , we have the following inequality

$$\begin{aligned} \sup_{z \in \mathcal{X}} \frac{4}{n} \boldsymbol{\eta}_z^T \mathbf{W}_z \Psi \Delta &\leq \sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{4}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \boldsymbol{\eta}_z\|_2 \|\Delta_{2:d}\|_{1,2} + \sup_{z \in \mathcal{X}} \frac{4}{n} \|\Psi_{\bullet 1}^T \mathbf{W}_z \boldsymbol{\eta}_z\|_2 |\Delta_1| \\ &\leq \lambda \|\Delta_{2:d}\|_{1,2} + \lambda \sqrt{m} |\Delta_1|. \end{aligned}$$

The first inequality is due to the Hölder's inequality and the second one is by the definition of \mathcal{E} .

Furthermore, we derive the following inequality

$$\begin{aligned} \frac{1}{n} \|\mathbf{W}_z^{1/2} \Psi \Delta\|_2^2 &\leq \frac{2}{n} \boldsymbol{\eta}_z^T \mathbf{W}_z \Psi \Delta - \lambda \sum_{j=2}^d (\|\widehat{\beta}_j\| - \|\beta_j\|) - \lambda \sqrt{m}(|\widehat{\alpha}_z| - |f_1(z)|) \\ &\leq \frac{\lambda}{2} \|\widehat{\beta} - \beta\|_{1,2} + \lambda \sqrt{m} |\widehat{\alpha}_z - f_1(z)| - \lambda \sum_{j=2}^d (\|\widehat{\beta}_j\| - \|\beta_j\|) \\ &\leq \frac{3\lambda}{2} \sum_{j \in \mathcal{S}_z} \|\Delta_j\| + \frac{3\lambda}{2} \sqrt{m} |\Delta_1| - \frac{\lambda}{2} \sum_{j \in \mathcal{S}_z^c} \|\Delta_j\|. \end{aligned}$$

The last inequality shows that $\Delta \in \mathcal{A}_z$.

Next, we prove the rate of convergence by contradiction. Suppose that for some fixed t , which will be specified later, we have

$$\exists z \in \mathcal{X}, \quad \frac{1}{\sqrt{n}} \|\mathbf{W}_z^{1/2} \Psi \Delta\| > t. \quad (\text{A.5})$$

Equation (A.5) implies that there exists some $z \in \mathcal{X}$ such that

$$0 > \min_{\Delta \in \mathcal{A}_z, \|\widehat{\Sigma}_z^{1/2} \Delta\| \geq t} \frac{1}{n} \|\mathbf{W}_z^{1/2}(\mathbf{Y} - \Psi \widehat{\beta}_+)\|_2^2 - \frac{1}{n} \|\mathbf{W}_z^{1/2}(\mathbf{Y} - \Psi \beta_+)\|_2^2 + \lambda \|\widehat{\beta}_+\|_{1,2} - \lambda \|\beta_+\|_{1,2}.$$

Using the fact that \mathcal{A}_z is a cone, we can replace the constraint $\|\widehat{\Sigma}_z^{1/2} \Delta\| \geq t$ by $\|\widehat{\Sigma}_z^{1/2} \Delta\| = t$ and

the above inequality still preserves. Combining the event \mathcal{E} , we have

$$\begin{aligned}
0 &> \min_{\mathbf{\Delta} \in \mathcal{A}_z, \|\widehat{\mathbf{\Sigma}}_z^{1/2} \mathbf{\Delta}\| = t} \frac{1}{n} \|\mathbf{W}_z^{1/2} (\mathbf{Y} - \mathbf{\Psi} \widehat{\boldsymbol{\beta}}_+) \|_2^2 - \frac{1}{n} \|\mathbf{W}_z^{1/2} (\mathbf{Y} - \mathbf{\Psi} \boldsymbol{\beta}_+) \|_2^2 + \lambda \mathcal{R}(\widehat{\boldsymbol{\beta}}_+) - \lambda \mathcal{R}(\boldsymbol{\beta}_+) \\
&\geq \min_{\mathbf{\Delta} \in \mathcal{A}_z, \|\widehat{\mathbf{\Sigma}}_z^{1/2} \mathbf{\Delta}\| = t} \frac{1}{n} \|\mathbf{W}_z^{1/2} \mathbf{\Psi} \mathbf{\Delta}\|_2^2 - 2\lambda \mathcal{R}(\mathbf{\Delta}) + \lambda \mathcal{R}(\widehat{\boldsymbol{\beta}}_+) - \lambda \mathcal{R}(\boldsymbol{\beta}_+).
\end{aligned} \tag{A.6}$$

From Lemma A.1, we can bound the R.H.S. by

$$\begin{aligned}
2\mathcal{R}(\mathbf{\Delta}) - \mathcal{R}(\widehat{\boldsymbol{\beta}}_+) + \mathcal{R}(\boldsymbol{\beta}_+) &\leq 3 \sum_{j \in S} \|\mathbf{\Delta}_j\| + 3\sqrt{m} |\mathbf{\Delta}_1| \\
&\leq 3\sqrt{s} \|\mathbf{\Delta}_{S \cup \widehat{S}}\|_2 + 3\sqrt{m} |\mathbf{\Delta}_1| \\
&\leq 6\sqrt{2sm/\rho_{\min}} \|\widehat{\mathbf{\Sigma}}_z^{1/2} \mathbf{\Delta}\|_2.
\end{aligned} \tag{A.7}$$

Combining (A.6) and (A.7), we get a quadratic inequality

$$0 > t^2 - \frac{2\lambda\sqrt{sm}}{\rho_{\min}} t. \tag{A.8}$$

Setting $t = 2\sqrt{sm/\rho_{\min}} \cdot \lambda$, we obtain from (A.8) that $0 > t^2 - \left[2\lambda\sqrt{sm/\rho_{\min}}\right] t = 0$, which is a contradiction. Therefore, $\sup_{z \in \mathcal{X}} n^{-1/2} \|\mathbf{W}_z^{1/2} \mathbf{\Psi} \mathbf{\Delta}\|_2 \leq 2\lambda\sqrt{sm}/\rho_{\min}$. Using the rate for λ in (A.4) and $h = o(1)$, we have

$$\sup_{z \in \mathcal{X}} \frac{1}{\sqrt{n}} \|\mathbf{W}_z^{1/2} \mathbf{\Psi} \mathbf{\Delta}\|_2 \leq C\sqrt{sm} \left(\sqrt{\frac{\log(dmh^{-1})}{nh}} + \frac{\sqrt{s}}{m^{5/2}} + \frac{m^{3/2} \log(dh^{-1})}{n} + \frac{h^2}{\sqrt{m}} \right). \tag{A.9}$$

Now, using Lemma A.1, since $\mathbf{\Delta} \in \mathcal{A}_z$, we have that

$$\|\mathbf{\Delta}_{2:d}\|_{1,2} + \sqrt{m} |\mathbf{\Delta}_1| \leq \sqrt{s} \|\mathbf{\Delta}_{2:d}\|_2 + \sqrt{m} |\mathbf{\Delta}_1| \leq \sqrt{sm/(\rho_{\min} n)} \|\mathbf{W}_z^{1/2} \mathbf{\Psi} \mathbf{\Delta}\|_2$$

for any $z \in \mathcal{X}$, which leads to the following inequality

$$\sup_{z \in \mathcal{X}} \sqrt{m} |\widehat{\alpha}_z - f_1(z)| + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_{1,2} \leq Csm \left(\sqrt{\frac{\log(dmh^{-1})}{nh}} + \frac{\sqrt{s}}{m^{5/2}} + \frac{m^{3/2} \log(dh^{-1})}{n} + \frac{h^2}{\sqrt{m}} \right),$$

with probability at least $1 - c/n$. To obtain the best rate on the right hand side of the equation,

we choose $h \asymp n^{-1/6}$ and $m \asymp n^{1/6}$ to obtain

$$\sup_{z \in \mathcal{X}} \left\{ \sqrt{m} |\hat{a}_z - f_1(z)| + \sum_{i=2}^d \|\hat{\beta}_i - \beta_i\|_2 \right\} = O_P \left(\log(dn) n^{-1/4} \right).$$

According to Corollary 15 in Chapter XI of [de Boor \(2001\)](#), given a function $g(x) = \sum_{k=1}^m \beta_k \phi_k(x)$, we have

$$\|g\|_2^2 \asymp m^{-1} \sum_{k=1}^m \beta_k^2. \quad (\text{A.10})$$

Therefore, we have $\|\hat{f} - f\|_2 \leq \rho_{\min}^{-1} s \sqrt{m} \lambda$ and, when $h \asymp n^{-1/6}$ and $m \asymp n^{1/6}$, the rate becomes

$$\|\hat{f} - f\|_2^2 = O_P \left(n^{-2/3} \log(dn) \right).$$

This completes the proof. \square

B Accelerated Algorithm

This section presents details of our method to accelerate Algorithm 1. To estimate f_1 , we need to compute the estimator \hat{a}_z for a number of z values $z \in \{z_1, \dots, z_M\}$. A naïve approach is to run Algorithm 1 M times, once for each value of z 's. We provide a more efficient algorithm which significantly reduces the computational cost. From Algorithm 1 and (2.15), the most expensive operation is evaluation of the gradient

$$\nabla_j \mathcal{L}_z(\beta_+^{(t)}) = -\frac{1}{n} \Psi_{\bullet j}^T \mathbf{W}_z \left(\mathbf{Y} - \Psi \beta_+^{(t)} \right). \quad (\text{B.1})$$

Computing $\nabla_j \mathcal{L}_z(\beta_+^{(t)})$ for a single z requires $O(dm^2n)$ flops. If we trivially repeat the computation for M different z 's, the computational complexity is $O(dm^2nM)$ which is challenging when M is large. However, we can exploit the structure of $\nabla_j \mathcal{L}_z(\beta_+^{(t)})$ to reduce the computational complexity. According to (B.1) and the fact that $\psi_{jk}(X_{i1}) = \phi_k(X_{i1}) - \bar{\phi}_{jk}(z)$, the k -th coordinate of $\nabla_j \mathcal{L}_z(\beta_+^{(t)})$

has a formulation

$$\begin{aligned} \left(\nabla_j \mathcal{L}_z(\beta_+^{(t)}) \right)_k &= -\frac{1}{n} \sum_{i \in [n]} K_h(X_{i1} - z) \phi_k(X_{i1}) Y_i + \bar{\phi}_{jk}(z) \cdot \frac{1}{n} \sum_{i \in [n]} K_h(X_{i1} - z) Y_i \\ &\quad + \sum_{\ell \in [d], s \in [m]} \beta_{\ell s}^{(t)} \left\{ \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) [\phi_k(X_{i1}) - \bar{\phi}_{jk}(z)] [\phi_s(X_{i1}) - \bar{\phi}_{\ell s}(z)] \right\}. \end{aligned} \quad (\text{B.2})$$

The computation of $\nabla_j \mathcal{L}_z(\beta_+^{(t)})$ is mostly spent on evaluating the formulation

$$q(z) = \sum_{i=1}^n K_h(X_{i1} - z) u_i \quad (\text{B.3})$$

for $z \in \{z_1, \dots, z_M\}$ where u_1, \dots, u_n are fixed quantities (e.g., u_i could be Y_i , $\phi_k(X_{i1})$ or $Y_i \phi_k(X_{i1})$) when evaluating (B.2) independent of z . We introduce a fast method to calculate the general form $q(z)$ and apply it to the computation of (B.2). Without loss of generality, we assume that $z_1 < \dots < z_M$. The naïve method to evaluate $\{q(z_\ell)\}_{\ell \in [M]}$ separately for different z has the computational complexity $O(nM)$. However, if the kernel function has some special structure, we can reduce the complexity to $O(n + M)$. For example, for the uniform kernel $K(u) = \frac{1}{2} \mathbb{1}\{|u| \leq 1\}$, when we vary the value of z from z_ℓ to $z_{\ell+1}$, we just need to subtract u_i for those $i \in \{v : X_v \in [z_\ell - h, z_{\ell+1} - h]\}$ and add u_i for those $i \in \{v : X_v \in (z_\ell + h, z_{\ell+1} + h]\}$. For $M \gg h^{-1}$, the cardinality of $\{i : X_{i1} \in (z_\ell - h, z_{\ell+1} - h] \cup (z_\ell + h, z_{\ell+1} + h]\}$ does not increase with n or d . Therefore, the complexity to evaluate $\{q(z_\ell)\}_{\ell \in [M]}$ is reduced to $O(n + K)$. For the Epanechnikov kernel $K(u) = (3/4) \cdot (1 - u^2) \mathbb{1}\{|u| \leq 1\}$, suppose $q(z_\ell)$ is known and define $I_\ell = \{i : X_{i1} \in (z_\ell - h, z_{\ell+1} - h] \cup (z_\ell + h, z_{\ell+1} + h]\}$. We have $q(z_{\ell+1}) = q(z_\ell) + \Delta q(z_\ell)$, where

$$\Delta q(z_\ell) = q(z_{\ell+1}) - q(z_\ell) = \frac{3}{4} \sum_{i \in I_\ell} (1 - (X_{i1}/h)^2) u_i + \frac{3z}{2h^2} \sum_{i \in I_\ell} X_{i1} + \frac{3z^2}{4} \sum_{i \in I_\ell} u_i.$$

Similar to the argument for the case of uniform kernel, we also have $|I_\ell| = O(1)$ if $K \gg h^{-1}$. The computational complexity of $\sum_{i \in I_z} (1 - X_{i1}^2) u_i$ and the other two summations above for $z = 1, \dots, z_K$ is $O(n + K)$ and hence the computational complexity of $\{q(z_\ell)\}_{\ell \in [K]}$ for Epanechnikov kernel is also $O(n + K)$. We can also apply a similar trick to many other kernels. Now we turn back to the

calculation of the gradient $\nabla_j \mathcal{L}_z(\beta_+^{(t)})$. Let $\hat{p}_1(z) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - z)$,

$$\begin{aligned} Y_k^{(1)}(z) &= \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) Y_i, & Y_k^{(2)}(z) &= \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \phi_k(X_{ij}) Y_i, \\ Y_k^{(3)}(z) &= \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \phi_k(X_{ij}) & \text{and } R_{ks}(z) &= \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \phi_k(X_{ij}) \phi_s(X_{iu}). \end{aligned}$$

For different values of z , we denote the components of β_+ corresponding to the k -th B-spline basis for the j -th covariate as $\beta_{jk;z}$. According to the expansion in (B.2), we can write the k -th coordinate of $\nabla_j \mathcal{L}_z(\beta_+^{(t)})$ as

$$\begin{aligned} \left(\nabla_j \mathcal{L}_z(\beta_+^{(t)}) \right)_k &= -Y_k^{(1)}(z) + \bar{\phi}_{jk}(z) Y_k^{(2)}(z) + \frac{1}{n} \sum_{\ell \in [d], s \in [m]} \beta_{\ell s;z} \left(R_{kv}(z) - \bar{\phi}_{jk}(z) Y_v^{(2)}(z) \right) \\ &\quad - \frac{1}{n} \sum_{\ell \in [d], s \in [m]} \beta_{\ell s;z} \left(\bar{\phi}_{\ell s}(z) Y_k^{(3)}(z) - \bar{\phi}_{jk}(z) \bar{\phi}_{\ell s}(z) \hat{p}_1(z) \right). \end{aligned}$$

Based on the previous discussion on the calculation of $q(z)$ in (B.3), we note that it takes $O(n + M)$ operations to evaluate $\hat{p}_1(z), Y_k^{(1)}(z), Y_k^{(2)}(z), Y_k^{(3)}(z)$ and $R_{ks}(z)$ for M different values of z . Therefore, the computational complexity of each iteration in Algorithm 1 can be reduced from $O(dm^2nM)$ to $O(dm^2(n + M))$. Therefore under the case $M = O(n)$, we can estimate f_1 using the introduced procedure with the same computational complexity as (2.10). Since most of existing algorithms for the group Lasso involve evaluating the gradient (Yuan and Lin, 2007; Friedman et al., 2010; Farrell, 2013; Qin et al., 2013), the above argument is applicable to other solvers as well.

C Covering Properties of the Bootstrap Confidence Bands

In this section, we prove the theorems on the coverage probabilities for the Gaussian multiplier bootstrap confidence bands $\mathcal{C}_{n,\alpha}^b$ in (2.19). We will first prove Theorem 3.7 and Theorem 3.5 can be proven by following the same steps.

C.1 Proof of Theorem 3.7

We first prove that $\mathcal{C}_{n,\alpha}^b$ in (2.19) is honest. To simplify the notation, we will use \mathcal{X} to represent the interval $[-D_n, D_n]$ when there is no confusion. The strategy to prove the result is to establish a

sequence of processes from $\widehat{\mathbb{H}}_n(z)$ that approximate $\widetilde{Z}_n(z)$. We consider the following four stochastic processes

$$\widehat{\mathbb{H}}_n(z) = \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^n \xi_i \cdot \frac{\widehat{\sigma} K_h(X_{i1} - z) \boldsymbol{\Psi}_{i\bullet}^T \widehat{\boldsymbol{\theta}}_z}{\widehat{\sigma}_n(z)}; \quad (\text{C.1})$$

$$\widehat{\mathbb{H}}_n^{(1)}(z) = \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^n \xi_i \cdot \frac{\sigma K_h(X_{i1} - z) \boldsymbol{\Psi}_{i\bullet}^T \widehat{\boldsymbol{\theta}}_z}{\widehat{\sigma}_n(z)}, \quad (\text{C.2})$$

$$\widetilde{\mathbb{H}}_n(z) = \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^n \varepsilon_i \frac{K_h(X_{i1} - z) \boldsymbol{\Psi}_{i\bullet}^T \widehat{\boldsymbol{\theta}}_z}{\widehat{\sigma}_n(z)}, \quad (\text{C.3})$$

$$\widetilde{Z}_n(z) = \sqrt{nh} \cdot \widehat{\sigma}_n^{-1}(z) \left(\widehat{f}_1^u(z) - f_1(z) \right). \quad (\text{C.4})$$

Corollary 3.1 of [Chernozhukov et al. \(2014a\)](#) provides sufficient conditions for the confidence band to be asymptotically honest. Specifically, we need to verify the following high-level conditions:

H1 There exists a Gaussian process $\mathbb{G}_n(z)$ and a sequence of random variables W_n^0 such that

$$W_n^0 \stackrel{d}{=} \sup_{z \in \mathcal{X}} \mathbb{G}_n(z). \text{ Furthermore, } \mathbb{E}[\sup_{z \in \mathcal{X}} \mathbb{G}_n] \leq C\sqrt{\log n} \text{ and}$$

$$\mathbb{P}(|W_n^Z - W_n^0| > \varepsilon_{1n}) < \delta_{1n}$$

for some ε_{1n} and δ_{1n} .

H2 For any $\epsilon > 0$, the anti-concentration inequality

$$\sup_{x \in \mathbb{R}} \mathbb{P} \left(\left| \sup_{z \in \mathcal{X}} |\mathbb{G}_n(z)| - x \right| \leq \epsilon \right) \leq C\epsilon\sqrt{\log n}.$$

holds.

H3 Let $c_n(\alpha)$ be the $(1 - \alpha)$ -quantile of W_n^Z and $\widehat{c}_n(\alpha)$ be the $1 - \alpha$ quantile of \widehat{W}_n . There exists

τ_n, ε_{2n} and δ_{2n} such that

$$\mathbb{P}(\widehat{c}_n(\alpha) < c_n(\alpha + \tau_n) - \varepsilon_{2n}) \leq \delta_{2n} \quad \text{and} \quad \mathbb{P}(\widehat{c}_n(\alpha) > c_n(\alpha - \tau_n) + \varepsilon_{2n}) \leq \delta_{2n}.$$

H4 There exists ε_{3n} and δ_{3n} such that

$$\mathbb{P}\left(\sup_{z \in \mathcal{X}} \left| \frac{\hat{\sigma} \sqrt{\hat{p}_1(z)}}{\sigma \sqrt{p_1(z)}} - 1 \right| > \varepsilon_{3n} \right) \leq \delta_{3n}.$$

If the high-level conditions **H1** - **H4** are verified, Corollary 3.1 in [Chernozhukov et al. \(2014a\)](#) implies that

$$\mathbb{P}(f_1 \in \mathcal{C}_{n,\alpha}^b) \geq 1 - \alpha - (\varepsilon_{1n} + \varepsilon_{2n} + \varepsilon_{3n} + \delta_{1n} + \delta_{2n} + \delta_{3n}).$$

In the remaining part of the proof, we show that the conditions are satisfied.

The roadmap is to establish that the process in (C.4) is close to the process in (C.1) following the chain $\tilde{Z}_n \rightarrow \tilde{\mathbb{H}}_n \rightarrow \hat{\mathbb{H}}_n^{(1)} \rightarrow \hat{\mathbb{H}}_n$. After that, we can check conditions **H1** - **H3**. Since we do not use the population $\sigma_n(z) = \mathbb{E}[\hat{\sigma}_n(z)]$ in the intermediate processes, we do not need to check the condition **H4**.

In order to verify the condition **H1**, we first bound the difference between $\sup_{z \in \mathcal{X}} \tilde{\mathbb{H}}_n(z)$ and $\sup_{z \in \mathcal{X}} Z_n(z)$. We begin by considering two auxiliary processes

$$\tilde{\mathbb{H}}'_n(z) = \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^n \varepsilon_i K_h(X_{i1} - z) \Psi_{i\bullet}^T \hat{\boldsymbol{\theta}}_z \quad \text{and} \quad \tilde{Z}'_n(z) = \sqrt{nh} \left(\hat{f}_1^u(z) - f_1(z) \right).$$

Notice that the above processes are un-normalized version of (C.2) and (C.4), that is, $\tilde{\mathbb{H}}'_n(z) = \hat{\sigma}_n(z) \tilde{\mathbb{H}}_n(z)$ and $\tilde{Z}'_n(z) = \hat{\sigma}_n(z) \tilde{Z}_n(z)$. The following lemma provides a direct bound for the difference between $\tilde{\mathbb{H}}'_n(z)$ and $\tilde{Z}'_n(z)$.

Lemma C.1. Under the same conditions of Theorem 3.7, there exists a constant $c_0 > \delta/2$ such that with probability $1 - c/n$,

$$\sup_{z \in \mathcal{X}} \left| \tilde{\mathbb{H}}'_n(z) - \tilde{Z}'_n(z) \right| \leq Cn^{-c_0}.$$

We defer the proof of the lemma to Section F.5 and proceed to prove Theorem 3.7. We also need to study $\hat{\sigma}$ and $\hat{\sigma}_n(z)$ in the following lemmas.

Lemma C.2. Let the estimator for $\text{Var}(\varepsilon) = \sigma^2$ be $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2$. Let

$$r_n := \sqrt{\frac{s^2 \log(dmh^{-1})}{nm^{-2}h}} + \frac{s^{3/2}}{m^{3/2}} + \frac{s \log(dh^{-1})}{nm^{-5/2}} + s\sqrt{mh^2}. \quad (\text{C.5})$$

Under the same conditions of Theorem 3.7, there exists constants c, C such that $\mathbb{P}(|\hat{\sigma}^2 - \sigma^2| \geq Cn^{-c}) \leq 6/n$.

Lemma C.3. Let $\Sigma'_z = n^{-1} \Psi \mathbf{W}_z^2 \Psi^T$. Under the same conditions of Theorem 3.7, there exist constants c, C such that for sufficiently large n , with probability $1 - c/n$, for any $z \in \mathcal{X}$,

$$ch^{-1} \mathbf{e}_1^T \boldsymbol{\theta}_z \leq \hat{\boldsymbol{\theta}}_z^T \Sigma'_z \hat{\boldsymbol{\theta}}_z \leq Ch^{-1} \mathbf{e}_1^T \boldsymbol{\theta}_z.$$

We defer the proof of this lemma to Section F.6. Notice that we can no longer choose $h \asymp n^{-1/6}$, $m \asymp n^{1/6}$ used for the estimation rate in Theorem 3.2. This is because we need to under-regularize our estimator to make the bias terms ignorable.

From Lemmas C.3 and F.3, we have an upper bound of the inverse of $\hat{\sigma}_n^2(z) = \hat{\boldsymbol{\theta}}_z^T \Sigma'_z \hat{\boldsymbol{\theta}}_z$ as

$$\sup_{z \in \mathcal{X}} \sqrt{h} \cdot \hat{\sigma}_n^{-1}(z) \leq C. \quad (\text{C.6})$$

With Lemma C.1 and Lemma C.3, we are ready to bound the difference between $\sup_{z \in \mathcal{X}} \tilde{\mathbb{H}}_n(z)$ and $\sup_{z \in \mathcal{X}} Z_n(z)$. Let c_0 be the constant in Lemma C.1. We choose h, m satisfying the scaling condition of Theorem 3.7. We denote $c = c_0 - \delta/2$ and observe that $c > 0$ by Lemma C.1. From Lemma C.1 and (C.6), we have

$$\begin{aligned} \mathbb{P}\left(\sup_{z \in \mathcal{X}} |\tilde{\mathbb{H}}_n(z) - \tilde{Z}_n(z)| \geq Cn^{-c}\right) &\leq \mathbb{P}\left(\sup_{z \in \mathcal{X}} |\tilde{\mathbb{H}}'_n(z) - \tilde{Z}'_n(z)| \geq C\hat{\sigma}_n(z)n^{-c_0}/\sqrt{h}\right) \\ &\leq \mathbb{P}\left(\sup_{z \in \mathcal{X}} |\tilde{\mathbb{H}}'_n(z) - \tilde{Z}'_n(z)| \geq C^2n^{-c_0}\right) \leq 1/n. \end{aligned}$$

Define $V_n^0 = \sup_{z \in \mathcal{X}} \tilde{\mathbb{H}}_n(z)$ and $\tilde{V}^Z = \sup_{z \in \mathcal{X}} \tilde{Z}_n(z)$. Since $\sup_{z \in \mathcal{X}} \tilde{\mathbb{H}}_n(z)$ is a Gaussian process conditional on $\{\mathbf{X}_{i1}\}_{i \in [n]}$, we verify H1 by

$$\mathbb{P}\left(|V_n^0 - \tilde{V}^Z| \geq Cn^{-c}\right) \leq \mathbb{P}\left(\sup_{z \in \mathcal{X}} |\tilde{\mathbb{H}}_n(z) - \tilde{Z}_n(z)| \geq Cn^{-c}\right) \leq \frac{1}{n}. \quad (\text{C.7})$$

The condition **H2** follows from **H1** and the anti-concentration inequality in Corollary 2.1 of Chernozhukov et al. (2014a).

Next, we check **H3** by bounding the difference between (C.1) and (C.2). We first approximate $\widehat{\mathbb{H}}_n(z)$ by $\widehat{\mathbb{H}}_n^{(1)}(z)$. By Lemma C.2, if we choose h, m satisfying the scaling condition of Theorem 3.7., with probability $1 - 6/n$, $|\widehat{\sigma} - \sigma| < C\sqrt{r_n m^{1/4}} = o(n^{-c})$, where r_n is defined in (C.5).

We denote $\widehat{V}_n = \sup_{z \in \mathcal{X}} \widehat{\mathbb{H}}_n(z)$, $\widehat{V}_n^{(1)} = \sup_{z \in \mathcal{X}} \widehat{\mathbb{H}}_n^{(1)}(z)$ and the difference between $\widehat{V}_n - \widehat{V}_n^{(1)}$. Let $\Delta \mathbb{H}^{(1)}(z) = \widehat{\mathbb{H}}_n^{(1)}(z) - \widehat{\mathbb{H}}_n(z)$. We have

$$\sup_{z \in \mathcal{X}} |\Delta \mathbb{H}^{(1)}(z)| \leq |\widehat{\sigma} - \sigma| \sup_{z \in \mathcal{X}} \sqrt{h} \cdot \widehat{\sigma}_n^{-1}(z) \left(\sup_{z \in \mathcal{X}} I_1(z) + \sup_{z \in \mathcal{X}} I_2(z) \right),$$

where $I_1(z) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - z) |\Psi_{i\bullet}^T(\widehat{\theta}_z - \theta_z)|$ and $I_2(z) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - z) |\Psi_{i\bullet}^T \theta_z|$.

In order to bound $I_1(z)$, we first state a technical lemma that characterizes the estimation error between $\widehat{\theta}_z$ and θ_z .

Lemma C.4. Let $\widehat{\theta}_z$ be a minimizer of (2.16). Suppose that Assumptions (A1), (A2), (A4)-(A6) hold and $\bar{\beta}_{2,\kappa}$ in (3.5) is finite. If the parameter γ in the optimization program (2.16) is chosen as in (3.11), then with probability $1 - c/d$,

$$\sup_{z \in \mathcal{X}} (\widehat{\theta}_z - \theta_z)^T \widehat{\Sigma}_z (\widehat{\theta}_z - \theta_z) \leq C \inf_{z \in \mathcal{X}} \frac{\log(D_n/p_1(z))}{p_1^2(z)} \cdot m \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \sqrt{\frac{\log(1/h)}{nh}} \right). \quad (\text{C.8})$$

We defer the proof of this lemma to Section F.4. Using Lemma C.4 we bound $I_1(z)$. Applying Cauchy-Schwarz inequality, we have

$$\begin{aligned} \sup_{z \in \mathcal{X}} |I_1(z)| &\leq \sup_{z \in \mathcal{X}} \left(\frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \left(\Psi_{i\bullet}^T (\widehat{\theta}_z - \theta_z) \right)^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \right)^{1/2} \\ &\leq C \inf_{z \in \mathcal{X}} \left(\frac{\log(D_n/p_1(z))}{p_1^2(z)} \right)^{1/2} \cdot \sqrt{m} \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \sqrt{\frac{\log(1/h)}{nh}} \right)^{1/2}, \end{aligned} \quad (\text{C.9})$$

where the last inequality is due to Proposition 3.1, Lemma C.4 and

$$\sup_{z \in \mathcal{X}} n^{-1} \sum_{i=1}^n K_h(X_{i1} - z) = \sup_{z \in \mathcal{X}} p_1(z) + o(1).$$

For $I_2(z)$, we have the following inequality

$$\begin{aligned}
\sup_{z \in \mathcal{X}} |I_2(z)| &\leq \sup_{z \in \mathcal{X}} \frac{1}{n} \|\Psi^T \mathbf{W}_z \mathbf{1}\|_{2,\infty} \|\boldsymbol{\theta}_z\|_1 \\
&\leq \sup_{z \in \mathcal{X}} \frac{1}{n} \|\mathbf{W}_z^{1/2} \Psi_{\bullet j} \Psi_{\bullet j}^T \mathbf{W}_z^{1/2}\|_2 \|\mathbf{W}_z^{1/2} \mathbf{1}\|_2 \|\boldsymbol{\theta}_z\|_1 \\
&\leq \frac{C}{m} \cdot \sup_{z \in \mathcal{X}} \sqrt{p(z)} \cdot \sqrt{m} = O(1/\sqrt{m}).
\end{aligned} \tag{C.10}$$

Therefore, combining (C.9) and (C.10), we have

$$\mathbb{P} \left(\left| \widehat{V}_n - \widehat{V}_n^{(1)} \right| > C\sqrt{r_n} m^{1/4} \right) \leq n^{-1}.$$

Under the scaling conditions of Theorem 3.7, there exists a constant c such that $\sqrt{r_n} m^{1/4} = O(n^{-c})$.

Since $\sigma \xi_i \stackrel{d}{=} \varepsilon_i$, we also have $\sup_{z \in \mathcal{X}} \widehat{\mathbb{H}}_n^{(1)}(z) \stackrel{d}{=} \sup_{z \in \mathcal{X}} \widetilde{\mathbb{H}}_n(z)$. Combining with (C.7), we have

$$\mathbb{P} \left(\left| \widehat{V}_n - \widetilde{V}_n^Z \right| > 2Cn^{-c} \right) \leq 2n^{-1}.$$

Therefore, we can bound the probability

$$\begin{aligned}
\mathbb{P}(\widetilde{V}_n^Z \leq \widehat{c}_n(\alpha) + 2Cn^{-c}) &\geq \mathbb{P}(\widetilde{V}_n^Z \leq \widehat{c}_n(\alpha)) - \mathbb{P}(|\widehat{V}_n - \widetilde{V}_n^Z| > 2Cn^{-c}) \\
&\geq 1 - \alpha - 2c/n^c,
\end{aligned} \tag{C.11}$$

which implies that the estimated quantile has the following lower bound

$$\widehat{c}_n(\alpha) \geq c_n(\alpha + 2Cn^{-c}) - 2cn^{-c}. \tag{C.12}$$

Similarly, we also have $\widehat{c}_n(\alpha) \leq c_n(\alpha - 2Cn^{-c}) + 2cn^{-c}$. By setting $\tau_n = 2Cn^{-c}$, $\varepsilon_{2n} = 2cn^{-c}$ and $\delta_{2n} = 2cn^{-c}$, we have

$$\mathbb{P} \left(\widehat{c}_n(\alpha) \geq c_n(\alpha + 2Cn^{-c}) - 2cn^{-c} \text{ and } \widehat{c}_n(\alpha) \leq c_n(\alpha - 2Cn^{-c}) + 2cn^{-c} \right) \leq 2c/n^c,$$

which verifies the condition **H3**.

Now, since we have checked the high-level conditions **H1** – **H3**, since the high-level conditions

H1 – **H4** are verified, the result follows from Corollary 3.1 in [Chernozhukov et al. \(2014a\)](#) such that

$$\mathbb{P}(f_1(z) \in \mathcal{C}_{n,\alpha}^b(z), \forall z \in \mathcal{X}) \geq 1 - \alpha - Cn^{-c},$$

which completes the proof of the theorem.

D Proof of Propositions 3.1

Let J be arbitrary subset of $[d]$ and for any $(\alpha, \beta) = (\alpha, \beta_2^T, \dots, \beta_d^T)^T \in \mathbb{C}_\beta^{(\kappa)}(J)$, where $\mathbb{C}_\beta^{(\kappa)}(J)$ is defined (3.1), we consider the functions $h_1(x_1) \equiv \alpha$ and $h_j(x_j) = \sum_{k=1}^m \beta_{jk} \psi_{jk}(x_j)$, for $j = 2, \dots, d$. From the cone restriction in (3.1), we have

$$\sum_{j \notin J, j \neq 1} \|\beta_j\|_2 \leq \kappa \sum_{j \in J, j \neq 1} \|\beta_j\|_2 + \kappa \sqrt{m} |\alpha|.$$

By the B-spline property [de Boor \(2001\)](#), there exists a constant $c_1 > 0$ such that

$$\begin{aligned} \sqrt{\frac{m}{c_1}} \sum_{j \notin J} \|h_j\|_{L^2(\mu_z)} &\leq \sum_{j \notin J \setminus \{1\}} \|\beta_j\|_2 + \sqrt{\frac{m}{c_1}} |\alpha| \\ &\leq \kappa \sum_{j \in J \setminus \{1\}} \|\beta_j\|_2 + \left(\kappa + c_1^{-1/2} \right) \sqrt{m} |\alpha| \\ &\leq \left((\sqrt{c_1} + 1) \kappa + c_1^{-1/2} \right) \sqrt{m} \cdot \kappa \sum_{j \in J} \|h_j\|_{L^2(\mu_z)}. \end{aligned} \quad (\text{D.1})$$

Let c be the smallest constant satisfying $c\kappa \geq (\sqrt{c_1} + 1) \kappa + c_1^{-1/2}$. Combining the above with (D.1), we have $(h_1, \dots, h_d) \in \mathbb{C}_h^{(c\kappa)}(J)$.

We first aim to bound $\beta_+^T \Sigma_z \beta_+ = \mathbb{E} \left[K_h(X_1 - z) \left(\sum_{j=2}^d h_j(X_j) \right)^2 \right]$ as

$$\begin{aligned} \mathbb{E} \left[K_h(X_1 - z) \left(\sum_{j=1}^d h_j(X_j) \right)^2 \right] &= \int K(u) \sum_{j,k} h_j(x_j) h_k(x_k) p_{1,j,k}(z + uh, x_j, x_k) du dx_j dx_k \\ &= \int \sum_{j,k} h_j(x_j) h_k(x_k) p_{1,j,k}(z, x_j, x_k) dx_j dx_k + O(h^2) \\ &\geq \frac{p_1(z)}{2} \cdot \mathbb{E} \left[\left(\sum_{j=2}^d h_j(X_j) \right)^2 \middle| X_1 = z \right] = \frac{p_1(z)}{2} \left\| \sum_{j=1}^d h_j \right\|_{L^2(\mu_z)}^2. \end{aligned}$$

Therefore, for any $z \in \mathcal{X}$, by the definition in (3.4),

$$\beta_+^T \Sigma_z \beta_+ \geq \frac{p_1(z)}{2} \left\| \sum_{j=1}^d h_j \right\|_{L^2(\mu_z)}^2 \geq \frac{p_1(z)}{2\beta_{2,c\kappa}^2(J)} \sum_{j \in J} \|h_j\|_{L^2(\mu_z)}^2 \quad (\text{D.2})$$

Applying the cone restriction on (h_1, \dots, h_d) , we further have

$$\begin{aligned} \left(\sum_{j \in J} \|h_j\|_{L^2(\mu_z)} \right)^2 &\geq \frac{1}{(c\kappa + 1)^2} \left(\sum_{j=1}^d \|h_j\|_{L^2(\mu_z)} \right)^2 \\ &\geq \frac{1}{(c\kappa + 1)^2} \sum_{j=1}^d \|h_j\|_{L^2(\mu_z)}^2 \geq \frac{Cm^{-1}}{(c\kappa + 1)^2} \left(\sum_{j=2}^d \|\beta_j\|_2^2 + m\alpha^2 \right). \end{aligned}$$

Combining the above inequality with (D.2), we obtain that with probability $q(\delta)$ for any $z \in \mathcal{X}$,

$$\beta_+^T \Sigma_z \beta_+ \geq p_1(z) \cdot \frac{Cm^{-1}\bar{\beta}_{2,c\kappa}^{-2}}{s(c\kappa + 1)^2} (\|\beta\|_2^2 + m\alpha^2), \quad \text{for any } \beta_+ \in \mathbb{C}_\beta^{(\kappa)}(J).$$

This completes the proof.

E Auxiliary Lemmas for Estimation Rate

In this section, we give detailed proofs of technical lemmas stated in Section A. The principal technique used in the proofs of this section is the control of the suprema of empirical processes. In the entire section, we will abuse the notation σ_P^2 as the variance for certain empirical process.

E.1 Restricted eigenvalue condition

We provide a proof of Lemma A.1 in this section. Before stating the main part of the proof, we begin with a technical lemma.

Lemma E.1 (Restricted eigenvalue condition). With probability larger than $1 - c/(dm)$, for any $\theta = (\alpha, \beta^T)^T$, with $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^{(d-1)m}$, and any $z \in \mathcal{X}$ it holds that

$$\theta^T \widehat{\Sigma}_z \theta \geq \theta^T \Sigma_z \theta - C \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \frac{1}{\sqrt{nh}} + h^2 \right) \|\theta\|_{1,2}^2,$$

where $\|\boldsymbol{\theta}\|_{1,2} = |\alpha| + \|\boldsymbol{\beta}\|_{1,2}$. Moreover, for any $j \in [d]$ there exists a constant C such that

$$\sup_{x \in \mathcal{X}} \frac{1}{n} \|\Psi_{\bullet,j} \mathbf{W}_z \Psi_{\bullet,j}^T\|_2^2 \leq C m^{-1}.$$

Proof. The proof strategy is to study suprema of the entries of $\widehat{\boldsymbol{\Sigma}}_z - \boldsymbol{\Sigma}_z$. We denote the (u, v) entry of $\boldsymbol{\Sigma}_z$ as $\Sigma_z(u, v)$ and similarly for $\widehat{\boldsymbol{\Sigma}}_z$. Let \mathbb{E}_n denote the empirical expectation. We first study the random variable

$$Z_{kk'jj'} = \sup_{z \in \mathcal{X}} (\mathbb{E}_n - \mathbb{E})[K_h(X_{i1} - z) \psi_{jk}(X_{ij}) \psi_{j'k'}(X_{ij'})].$$

Notice that $Z_{kk'jj'} = \sup_{z \in \mathcal{X}} [\widehat{\boldsymbol{\Sigma}}_z - \boldsymbol{\Sigma}_z](1 + (j-2)m + k, 1 + (j'-2)m + k')$ for $j, j' \geq 2$ and $k \in [m]$. In order to bound $Z_{kk'jj'}$, we turn to study the covering number of the space

$$\mathcal{G}_h = \{g_z(x_1, x_2, x_3) = h^{-1}K(h^{-1}(x_1 - z))\psi_{jk}(x_2)\psi_{j'k'}(x_3) \mid z \in \mathcal{X}, x_1, x_2, x_3 \in \mathcal{X}\}.$$

Let $\mathcal{F}_h = \{h^{-1}K(h^{-1}(\cdot - z)) \mid z \in \mathcal{X}\}$ and let $\|K\|_{\text{TV}}$ be the total variation of $K(\cdot)$. From Lemma H.3, we have

$$\sup_Q N(\mathcal{F}_h, L^2(Q), \epsilon) \leq \left(\frac{2\|K\|_{\text{TV}}A}{h\epsilon} \right)^4, \quad 0 < \epsilon < 1,$$

where the supremum is taken over all probability measures Q on \mathbb{R} . Let $\widetilde{\mathcal{F}}_h$ be an ϵ/L -cover of \mathcal{F}_h with respect to Q , where $L \geq \|\psi_{jk}\|_\infty$ for any k . We construct an ϵ -cover for \mathcal{G}_h with respect to $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_{i1}, \dots, X_{id}}$ as

$$\widetilde{\mathcal{G}}_h = \{f_1(x_1)\psi_{jk}(x_2)\psi_{j'k'}(x_3) \mid f_1 \in \widetilde{\mathcal{F}}_h\}.$$

For a function $g_z = h^{-1}K(h^{-1}(x_1 - z))\psi_{jk}(x_2)\psi_{j'k'}(x_3) \in \mathcal{G}_h$, let

$$\widetilde{g}_z = h^{-1}K(h^{-1}(x_1 - \widetilde{z}))\psi_{jk}(x_2)\psi_{j'k'}(x_3) \in \widetilde{\mathcal{G}}_h$$

be the corresponding element in the cover. Here $h^{-1}K(h^{-1}(x_1 - \widetilde{z})) \in \widetilde{\mathcal{F}}_h$ is the corresponding element in the cover for $h^{-1}K(h^{-1}(x_1 - z)) \in \mathcal{F}_h$. If we are able to show that $\|g_z - \widetilde{g}_z\|_{L^2(Q)}^2 \leq \epsilon^2$, then $\widetilde{\mathcal{G}}_h$ is the ϵ -cover for \mathcal{G}_h . Combining with Lemma 3, Giné and Nickl (2009) (also see Lemma

H.3), the covering number of \mathcal{G}_h can be bounded as

$$N(\mathcal{G}_h, L^2(\mathbb{P}_n), \epsilon) \leq \left(\frac{2\|K\|_{\text{TV}} AL}{h\epsilon} \right)^4. \quad (\text{E.1})$$

Now we show $\|g_z - \tilde{g}_z\|_{L^2(Q)}^2 \leq \epsilon^2$:

$$\begin{aligned} \|g_z - \tilde{g}_z\|_{L^2(Q)}^2 &= \mathbb{E}_Q [((K_h(X_1 - z) - K_h(X_1 - \tilde{z})) \psi_{jk}(X_{ji}) \psi_{j'k}(X_{j'i}))^2] \\ &\leq \mathbb{E}_Q [((K_h(X_1 - z) - K_h(X_1 - \tilde{z})))^2] \leq \epsilon^2. \end{aligned}$$

Observe that all functions in \mathcal{G}_h are bounded by $U = 4h^{-1}\|K\|_\infty$ and

$$\begin{aligned} \sigma_P^2 &:= \mathbb{E} \left[(K_h(X_1 - z) (\psi_{jk}(X_j) \psi_{j'k'}(X_{j'})))^2 \right] \\ &= h^{-2} \mathbb{E} \left[K^2(h^{-1}(X_1 - z)) \mathbb{E}[(\psi_{jk}^2(X_j) \psi_{j'k'}^2(X_{j'})) | X_1] \right] \\ &\leq Bm^{-2}h^{-2} \mathbb{E} [K^2(h^{-1}(X_1 - z))] \\ &= Bm^{-2}h^{-1} \int K^2(u) p_1(z + uh) du \leq B^2m^{-2}h^{-1}, \end{aligned}$$

where the first and last inequalities are due to Assumption **(A1)**. The bound above does not depend on the particular choice of z . If $m(nh)^{-1} = o(1)$, we have $n\sigma_P^2 \geq CU^2 \log(U\sigma^{-1})$, and from Lemma H.2, we have

$$\mathbb{E}[Z_{kk'jj'}] \leq C_1 \sqrt{\frac{\log(C_2 m)}{nm^2 h}}, \quad (\text{E.2})$$

where the constants C_1, C_2 are independent of k, k', j, j' . As $|Z_{kk'jj'}| \leq 4h^{-1}$ and $\sigma_P^2 \leq Cm^{-2}h^{-2}$, we can apply Lemma H.4 to obtain

$$\mathbb{P}\left(Z_{kk'jj'} \geq \mathbb{E}[Z_{kk'jj'}] + t\sqrt{Cm^{-2}h^{-1} + 4h^{-1}\mathbb{E}[Z_{kk'jj'}] + 4t^2h^{-1}/3}\right) \leq \exp(-nt^2). \quad (\text{E.3})$$

For $t = \log d/\sqrt{n}$, there exists a constant C such that

$$Z_{kk'jj'} \leq C \log dm/\sqrt{nm^2 h} + C/(nh)$$

with probability $1 - 1/d$. Combining (E.2) with (E.3), there exists a constant C such that

$$\begin{aligned} & \mathbb{P} \left(\max_{j,j' \geq 2, k, k' \in [m]} |Z_{kk'jj'}| > 2\mathbb{E}[Z_{kk'jj'}] + t\sqrt{Cm^{-2}h^{-1} + 4h^{-1}\mathbb{E}[Z_{kk'jj'}]} + 4t^2h^{-1}/3 \right) \\ & \leq \mathbb{P} \left(\max_{k, k' \in [m], j, j' \geq 2} |Z_{kk'jj'} - \mathbb{E}[Z_{kk'jj'}]| > t\sqrt{Cm^{-2}h^{-1} + 4h^{-1}\mathbb{E}[Z_{kk'jj'}]} + 4t^2h^{-1}/3 + \mathbb{E}[Z_{kk'jj'}] \right) \\ & \leq (dm)^2 \exp(-nt^2). \end{aligned}$$

Let $t = 3\sqrt{\log(dm)/n}$, $n_{jk} = 1 + (j - 2)m + k$ and $n_{j'k'} = 1 + (j' - 2)m + k'$ and we obtain that

$$\sup_{z \in \mathcal{X}} \max_{j, j' \geq 2, k, k' \in [m]} \left| [\widehat{\Sigma}_z - \Sigma_z](n_{jk}, n_{j'k'}) \right| = O_P \left(\frac{1}{nh} + \sqrt{\frac{\log(dm)}{nm^2h}} \right). \quad (\text{E.4})$$

Similarly, we define $\bar{Z}_{kj} = \sup_{z \in \mathcal{X}} (\mathbb{E}_n - \mathbb{E})[K_h(X_{i1} - z)\psi_{jk}(X_{ij})]$. Following the similar procedure as above, we apply Lemma H.2 to obtain that for some constant C ,

$$\sigma_P^2 := \mathbb{E} \left[(K_h(X_1 - z)\psi_{jk}(X_j))^2 \right] \leq Cm^{-1}h^{-1},$$

and $U \leq h^{-1}$, which implies the following inequality

$$\mathbb{E}[\bar{Z}_{kj}] \leq C_1 \sqrt{\frac{\log(C_2m)}{nmh}}. \quad (\text{E.5})$$

We now turn to study the remaining entries of $\widehat{\Sigma}_z - \Sigma_z$. Using the same arguments as in (E.4) and (E.5), we can derive an upper bound on the difference

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} \left| \widehat{\Sigma}_z(1 + (j - 2)m + k, 1) - \Sigma_z(1 + (j - 2)m + k, 1) \right| = O_P \left(\frac{1}{nh} + \sqrt{\frac{\log(dm)}{nmh}} \right). \quad (\text{E.6})$$

From Assumption (A1), the density function of X_1 , $p_1(x)$, is smooth. Recall that $\widehat{p}_1(z) = n^{-1} \sum_{i=1}^n K_h(X_{i1} - z)$. Applying the supreme norm rate for a kernel density estimator established in Theorem 2.3 of Giné and Guillou (2002), we have $\|\widehat{p}_1 - \mathbb{E}[\widehat{p}_1]\|_\infty = O_P(\sqrt{\log(1/h)/(nh)})$ and therefore we can get the rate

$$\sup_{z \in \mathcal{X}} \left| \widehat{\Sigma}_z(1, 1) - \Sigma_z(1, 1) \right| = \sup_{z \in \mathcal{X}} |\widehat{p}_1(z) - \mathbb{E}[\widehat{p}_1(z)]| = O_P \left(\sqrt{\frac{\log(1/h)}{nh}} \right). \quad (\text{E.7})$$

Combining (E.4), (E.6) and (E.7), according to Hölder inequality, we have for any $z \in \mathcal{X}$

$$\begin{aligned}
\left| \boldsymbol{\theta}^T (\widehat{\boldsymbol{\Sigma}}_z - \boldsymbol{\Sigma}_z) \boldsymbol{\theta} \right| &\leq \|\boldsymbol{\theta}\|_1^2 \|\widehat{\boldsymbol{\Sigma}}_z - \boldsymbol{\Sigma}_z\|_{\max} \\
&\leq \|\boldsymbol{\theta}\|_{1,2}^2 \left\{ \sup_{z \in \mathcal{X}} \max_{t, t' \neq 1} m \left| \widehat{\boldsymbol{\Sigma}}_z(t, t') - \boldsymbol{\Sigma}_z(t, t') \right| + \sup_{z \in \mathcal{X}} \left| \widehat{\boldsymbol{\Sigma}}_z(1, 1) - \boldsymbol{\Sigma}_z(1, 1) \right| \right\} \\
&\leq C \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \sqrt{\frac{\log(1/h)}{nh}} \right) \|\boldsymbol{\theta}\|_{1,2}^2,
\end{aligned} \tag{E.8}$$

which completes the proof of the first part of the lemma.

An upper bound on $\sup_{x \in \mathcal{X}} n^{-1} \|\boldsymbol{\Psi}_{\bullet j} \mathbf{W}_z \boldsymbol{\Psi}_{\bullet j}^T\|_2^2$ can be obtained in a way similar to the proof of Lemma 6.2 in Zhou et al. (1998). For any $\boldsymbol{\beta}_j = (\beta_1, \dots, \beta_m)^T$, let $u(x_j) = \sum_{k=1}^m \beta_k \psi_{jk}(x_j)$. Let the joint density function between X_1 and X_j be $p_{1,j}(x_1, x_j)$. From Assumption (A1), we have for any $z \in \mathcal{X}$,

$$\begin{aligned}
\frac{1}{n} \boldsymbol{\beta}_j^T \mathbb{E}[\boldsymbol{\Psi}_{\bullet j} \mathbf{W}_z \boldsymbol{\Psi}_{\bullet j}^T] \boldsymbol{\beta}_j &= \int \frac{1}{h} K\left(\frac{x_1 - z}{h}\right) u^2(x_j) p_{1,j}(x_1, x_j) dx_1 dx_j \\
&\leq C \int K(u) du \int u^2(x_j) dx_j \leq C m^{-1} \sum_{k=1}^m \beta_k^2.
\end{aligned} \tag{E.9}$$

Furthermore, we also have

$$\begin{aligned}
\sup_{x \in \mathcal{X}} \frac{1}{n} \boldsymbol{\beta}_j^T \mathbb{E}[\boldsymbol{\Psi}_{\bullet j} \boldsymbol{\Psi}_{\bullet j}^T | X_1 = x] \boldsymbol{\beta}_j &= \int u^2(x_j) \frac{p_{1,j}(x_1, x_j)}{p_1(x)} dx_1 dx_j \\
&\leq B \int K(u) du \int u^2(x_j) dx_j \leq C m^{-1} \sum_{k=1}^m \beta_k^2.
\end{aligned} \tag{E.10}$$

Let $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_{i1}, X_{ij}}$. We write the integration as

$$\sup_{z \in \mathcal{X}} \int K_h(x_1 - z) u^2(x_j) d\mathbb{P}_n = I_1 + I_2, \text{ where}$$

$$I_1 = \sup_{z \in \mathcal{X}} \int K_h(x_1 - z) u^2(x_j) d\mathbb{P}_{X_1, X_j} \text{ and } I_2 = \sup_{z \in \mathcal{X}} \left| \int K_h(x_1 - z) u^2(x_j) d(\mathbb{P}_n - \mathbb{P}_{X_1, X_j}) \right|.$$

Due to (E.9), we have $I_1 = O_P(m^{-1}) \|\boldsymbol{\beta}_j\|_2^2$ and a similar argument to one in Lemma 6.2 of Zhou et al. (1998) will derive $I_2 = o(h) \|\boldsymbol{\beta}_j\|_2^2$. This completes the proof. \square

Based on Lemma E.1, the remaining step is to prove Lemma A.1.

Proof of Lemma A.1. We can derive the restricted eigenvalue condition on the cone from Lemma E.1.

We apply Lemma E.1 in the last step. If the cone condition

$$\sum_{j \in S^c} \|\beta_j\|_2 \leq 3 \sum_{j \in S} \|\beta_j\|_2 + 3\sqrt{m}|\alpha|$$

is satisfied, by Hölder inequality, we have the upper bound

$$\|\beta\|_{1,2} \leq 4 \sum_{j \in S} \|\beta_j\|_2 + 3\sqrt{m}|\alpha| \leq 4\sqrt{s}\|\beta\|_2 + 3\sqrt{m}|\alpha|.$$

With large probability, we have the following inequality

$$\begin{aligned} \theta^T \widehat{\Sigma}_z \theta &\geq \theta^T \Sigma_z \theta - C \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \frac{1}{\sqrt{nh}} + h^2 \right) \|\theta\|_{1,2}^2 \\ &\geq \rho_{\min} |\alpha|^2 + \rho_{\min} \|\beta\|_2 / m - C \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \frac{1}{\sqrt{nh}} + h^2 \right) (4m|\alpha|^2 + 4s\|\beta\|_2^2) \\ &\geq \rho_{\min} |\alpha|^2 / 2 + \rho_{\min} m^{-1} \|\beta\|_2^2 / 2 \end{aligned}$$

for any $z \in \mathcal{X}$ and sufficiently large n if $s\sqrt{m^3 \log(dm)/(nh)} + sm^2/(nh) = o(1)$. \square

E.2 Proof of Lemma A.2

The proof can be separated into two cases: $j = 1$ and $j \geq 2$. For the simplicity of notation, we write $\delta_i(z)$ as δ_i in this proof. We first consider the situation when $j \geq 2$ and prove (A.1) and (A.3). From Lemma A.1,

$$\sup_{z \in \mathcal{X}} \|\mathbf{W}_z^{1/2} \Psi_{\bullet j} \Psi_{\bullet j}^T \mathbf{W}_z^{1/2}\|_2 / \sqrt{n} = \sup_{z \in \mathcal{X}} \|\Psi_{\bullet j} \mathbf{W}_z \Psi_{\bullet j}^T\|_2 / \sqrt{n} \leq \rho_{\max} m^{-1/2}$$

with high probability. Therefore

$$\sup_{z \in \mathcal{X}} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \delta\|_2 \leq \sup_{z \in \mathcal{X}} \frac{1}{n} \|\mathbf{W}_z^{1/2} \Psi_{\bullet j} \Psi_{\bullet j}^T \mathbf{W}_z^{1/2}\|_2 \|\mathbf{W}_z^{1/2} \delta\|_2 \leq \frac{C}{\sqrt{m}} \cdot \sup_{z \in \mathcal{X}} \frac{1}{\sqrt{n}} \|\mathbf{W}_z^{1/2} \delta\|_2. \quad (\text{E.11})$$

To complete the proof, we need a bound on

$$\sup_{z \in \mathcal{X}} \frac{1}{n} \|\mathbf{W}_z^{1/2} \boldsymbol{\delta}\|_2^2 = \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \delta_i^2. \quad (\text{E.12})$$

Using Equation (20) in [Zhou et al. \(1998\)](#) on B-spline, we have

$$\delta_i^2 = \left(\sum_{j=2}^d f_{jz}(X_{ji}) - f_{nj;z}(X_{ji}) \right)^2 \leq s^2 m^{-4}.$$

Define the following empirical process

$$U_n(z) = \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \delta_i^2 - \mathbb{E}[K_h(X_{11} - z) \delta_1^2].$$

Applying Hoeffding's inequality ([Hoeffding, 1963](#)), we have

$$\mathbb{P} \left(\sup_{z \in \mathcal{X}} U_n(z) - \mathbb{E} \left[\sup_{z \in \mathcal{X}} U_n(z) \right] > t \right) \leq \exp \left(-C \frac{nh^2 t^2}{(sm^{-4})^2} \right). \quad (\text{E.13})$$

Let

$$\mathcal{G}_h'' = \left\{ g_z(x_1, x_2) = h^{-1} K(h^{-1}(x_1 - z)) \delta^2(x_2) \mid z \in \mathcal{X}, x_1 \in \mathcal{X}, x_2 \in \mathcal{X}^{d-1} \right\},$$

where $\delta(x_2) = \sum_{j=2}^d f_j(x_{2j}) - f_{nj}(x_{2j})$. Similar to the covering number of \mathcal{G}_h in [\(E.1\)](#), since $\delta^2(x_2) \leq sm^{-4}$ for any x_2 , we have for any measure Q ,

$$\sup_Q N(\mathcal{G}_h'', L^2(Q), \epsilon) \leq \left(\frac{2\sqrt{s} \|K\|_{\text{TV}} A}{m^2 h \epsilon} \right)^4.$$

Furthermore, $\sigma_P^2 := \mathbb{E}[K_h(X_{i1} - z) \delta_i^2]^2 = O((sm^{-4})^2 h^{-1})$. Since $g \leq U := Ch^{-1}(sm^{-4})$ for any $g \in \mathcal{G}_h''$ and $m^4(sn)^{-1} = o(1)$, we have $n\sigma_P^2 \geq C_1 U^2 \log(C_2 \sqrt{sm^{-4}} U / \sigma)$. By [Lemma H.2](#), we have

$$\mathbb{E} \left[\sup_{z \in \mathcal{X}} U_n(z) \right] \leq C \frac{sm^{-4}}{\sqrt{nh}} \sqrt{\log(m^2 / \sqrt{sh})}. \quad (\text{E.14})$$

We set $t = Cs(m^4 h)^{-1} \sqrt{\log n / n}$ in [\(E.13\)](#) and combine it with [\(E.14\)](#) to obtain that, with probability at least $1 - 1/n$,

$$\sup_{z \in \mathcal{X}} U_n(z) \leq C \frac{sm^{-4}}{\sqrt{nh}} \sqrt{\log(m^2 / \sqrt{sh})} + C \frac{s \sqrt{\log n / n}}{m^4 h}. \quad (\text{E.15})$$

Finally, we bound the maximal of the expectation by

$$\begin{aligned} \sup_{z \in \mathcal{X}} \mathbb{E}[K_h(X_{11} - z)\delta_1^2] &= \sup_{z \in \mathcal{X}} \int K_h(t - z) dP_{X_1}(t) \delta^2(u) dP_{X_{>2}|X_1=t}(u) \\ &\leq Csm^{-4} \sup_{z \in \mathcal{X}} \int K_h(t - z) dP_{X_1}(t) \leq Csm^{-4}. \end{aligned} \quad (\text{E.16})$$

Combining (E.15) and (E.16), with probability at least $1 - 1/n$, we have

$$\begin{aligned} \sup_{z \in \mathcal{X}} \frac{1}{n} \|\mathbf{W}_z^{1/2} \boldsymbol{\delta}\|_2^2 &= \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \delta_i^2 \\ &\leq \sup_{z \in \mathcal{X}} U_n(z) + \sup_{z \in \mathcal{X}} \mathbb{E}[K_h(X_{11} - z) \delta_1^2] \\ &\leq C \frac{sm^{-4}}{\sqrt{nh}} \sqrt{\log(m^2/\sqrt{sh})} + C \frac{s\sqrt{\log n/n}}{m^4 h} + Csm^{-4} \\ &= O(sm^{-4}), \end{aligned} \quad (\text{E.17})$$

where the last equality is due to $2/\sqrt{nh^2} = o(1)$. Therefore, we prove the upper bound in (A.3).

Combing (E.17) with (E.11), we have we can conclude that

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\boldsymbol{\Psi}_{\bullet j}^T \mathbf{W}_z \boldsymbol{\delta}\|_2 \leq \frac{C}{\sqrt{m}} \cdot \sup_{z \in \mathcal{X}} \frac{1}{\sqrt{n}} \|\mathbf{W}_z^{1/2} \boldsymbol{\delta}\|_2 \leq C \sqrt{\frac{s}{m^5}}.$$

This gives us the rate in (A.1).

The final step is to prove (A.2). Recall that $\boldsymbol{\Psi}_{\bullet 1} = (1, \dots, 1)^T$. For the case $j = 1$, following the proof for (A.3). According to (E.12), we have $|\delta_i| \leq sm^{-2}$ for any $i \in [n]$. Let

$$U'_n(z) = \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \delta_i - \mathbb{E}[K_h(X_{11} - z) \delta_1].$$

We use Hoeffding's inequality (Hoeffding, 1963) again and obtain

$$\mathbb{P} \left(\sup_{z \in \mathcal{X}} U'_n(z) - \mathbb{E} \left[\sup_{z \in \mathcal{X}} U'_n(z) \right] > t \right) \leq \exp \left(-C \frac{nh^2 t^2}{sm^{-4}} \right). \quad (\text{E.18})$$

Applying symmetrization inequality again, we have

$$\mathbb{E} \left[\sup_{z \in \mathcal{X}} U'_n(z) \right] \leq 2 \mathbb{E} \left[\sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n \xi_i K_h(X_{i1} - z) |\delta_i| \right],$$

where $\{\xi_i\}_{i=1}^n$ are i.i.d. Rademacher variables independent of data. Let

$$\tilde{\mathcal{G}}_h'' = \left\{ g_z(x_1, x_2) = h^{-1}K(h^{-1}(x_1 - z))\delta(x_2) \mid z \in \mathcal{X}, x_1 \in \mathcal{X}, x_2 \in \mathcal{X}^{d-1} \right\},$$

where $\delta(x_2) = \left| \sum_{j=2}^d f_j(x_{2j}) - f_{nj}(x_{2j}) \right|$. Just as the covering number of \mathcal{G}_h'' , we also have $\delta(x_2) \leq sm^{-2}$ for any x_2 , we have for any measure Q ,

$$\sup_Q N\left(\tilde{\mathcal{G}}_h'', L^2(Q), \epsilon\right) \leq \left(\frac{2s^{1/2}\|K\|_{\text{TV}}A}{m^2h\epsilon} \right)^4.$$

The variance of the process $\sigma_P^2 := \mathbb{E}[K_h(X_{i1} - z)\delta_i]^2 = O(sm^{-4}h^{-1})$. Since $g \leq U := Ch^{-1}(sm^{-4})^{1/2}$ for any $g \in \mathcal{G}_h''$ and $m^4(sn)^{-1} = o(1)$, we have $n\sigma_P^2 \geq C_1U^2 \log(C_2s^{1/4}m^{-1}U/\sigma)$. Applying Lemma H.2 again, we have

$$\mathbb{E} \left[\sup_{z \in \mathcal{X}} U'_n(z) \right] \leq C \frac{\sqrt{sm^{-2}}}{\sqrt{nh}} \sqrt{\log(m/\sqrt{sh})}. \quad (\text{E.19})$$

We let $t = C\sqrt{s}(m^2h)^{-1}\sqrt{\log n/n}$ in (E.18) and use it with (E.19). Therefore, we achieve with probability at least $1 - 1/n$,

$$\sup_{z \in \mathcal{X}} U'_n(z) \leq C \frac{\sqrt{sm^{-2}}}{\sqrt{nh}} \sqrt{\log(m^2/\sqrt{sh})} + C \frac{\sqrt{s \log n/n}}{m^2h}. \quad (\text{E.20})$$

We again bound the supreme of the expectation

$$\begin{aligned} \sup_{z \in \mathcal{X}} \mathbb{E}[K_h(X_{11} - z)\delta_1] &= \sup_{z \in \mathcal{X}} \int K_h(t - z) dP_{X_1}(t) \delta(u) dP_{X_{>2}|X_1=t}(u) \\ &\leq C\sqrt{sm^{-2}} \sup_{z \in \mathcal{X}} \int K_h(t - z) dP_{X_1}(t) \leq C\sqrt{sm^{-2}}. \end{aligned} \quad (\text{E.21})$$

Combining (E.20) and (E.21), with probability at least $1 - 1/n$, we have

$$\begin{aligned} \sup_{z \in \mathcal{X}} \frac{1}{n} |\Psi_{\bullet 1}^T \mathbf{W}_z \boldsymbol{\delta}_z| &\leq \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) |\delta_i| \\ &\leq \sup_{z \in \mathcal{X}} U'_n(z) + \sup_{z \in \mathcal{X}} \mathbb{E}[K_h(X_{11} - z)\delta_1] = O(\sqrt{sm^{-2}}). \end{aligned}$$

Therefore, we prove the upper bound in (A.2) which completes the proof of the lemma.

E.3 Proof of Lemma A.3

For $j \geq 2$, we bound the two terms $\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \xi_z\|_2$ and $\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \zeta_z\|_2$ separately. To bound the first term, let $\Delta f_z(x) = f_1(x) - f_1(z)$ and Ψ_{ij} be the i th row of $\Psi_{\bullet j}$. We can rewrite the suprema as

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \xi_z\|_2 = \max_{j \geq 2} \sup_{z \in \mathcal{X}} \sup_{\mathbf{v} \in \mathbb{B}^m} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}^T \Psi_{ij}. \quad (\text{E.22})$$

Let $N_v = \{\mathbf{v}_1, \dots, \mathbf{v}_M\}$ be a $1/2$ -covering of the sphere $\mathbb{B}^m = \{\mathbf{v} \in \mathbb{R}^m \mid \|\mathbf{v}\|_2 \leq 1\}$. Observe that for any $\mathbf{v} \in \mathbb{B}^m$, there exists $\pi(\mathbf{v}) \in N_v$ such that $\|\mathbf{v} - \pi(\mathbf{v})\|_2 \leq 1/2$. Therefore we have

$$\begin{aligned} & \sup_{\mathbf{v} \in \mathbb{B}^m} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}^T \Psi_{ij} \\ & \leq \sup_{k \in [M]} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}_k^T \Psi_{ij} + \sup_{\mathbf{v} \in \frac{1}{2}\mathbb{B}^m} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}^T \Psi_{ij} \\ & \leq \sup_{k \in [M]} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}_k^T \Psi_{ij} + \frac{1}{2} \sup_{\mathbf{v} \in \mathbb{B}^m} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}^T \Psi_{ij}. \end{aligned}$$

If we move the second term on the right hand side of the last inequality to the left hand side, we obtain the inequality that

$$\sup_{\mathbf{v} \in \mathbb{B}^m} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}^T \Psi_{ij} \leq 2 \sup_{k \in [M]} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \Delta f_z(X_{1i}) \mathbf{v}_k^T \Psi_{ij}.$$

Therefore, in order to bound (E.22), we need to study the following empirical process

$$V_n(z) = \max_{j \geq 2, k \in [M]} \sup_{z \in \mathcal{X}} \left\{ \frac{1}{n} \sum_{i=1}^n K_h(X_{1i} - z) \Delta f_z(X_{1i}) \mathbf{v}_k^T \Psi_{ij} - \mathbb{E}[K_h(X_{11} - z) \Delta f_z(X_{11}) \mathbf{v}_k^T \Psi_{ij}] \right\}.$$

We define the following function class

$$\mathcal{G}_h''' = \left\{ g_z(x_1, x_2) = h^{-1} K((x_1 - z)/h) \Delta f_z(x_1) \sum_{t=1}^m \mathbf{v}_{kt} \psi_t(x_j) \mid j \geq 2, k \in [M], z \in \mathcal{X} \right\}$$

and, similarly to argument in the covering number of \mathcal{G}_h in (E.1), we have

$$\sup_Q N(\mathcal{G}_h''', L^2(Q), \epsilon) \leq dM \left(\frac{2\sqrt{m} \|K\|_{\text{TV}} A}{h\epsilon} \right)^4.$$

From (E.10), we bound the maximal of the expectation by

$$\sup_{x \in \mathcal{X}} \mathbb{E}[(\mathbf{v}_k^T \boldsymbol{\Psi}_{ij})^2 | X_1 = x] \leq C \|\mathbf{v}_k\|^2 m^{-1}.$$

Furthermore, we can bound the variance by expanding the expectation as the integration and applying the Taylor expansion as follows

$$\begin{aligned} \sigma_P^2 &:= \mathbb{E}[K_h(X_1 - z) \Delta f_z(X_1) \mathbf{v}_k^T \boldsymbol{\Psi}_{ij}]^2 \\ &= h^{-2} \int K^2 \left(\frac{x - z}{h} \right) (f_1(x) - f_1(z))^2 p_{X_1}(x) dx \cdot \mathbb{E}[(\mathbf{v}_k^T \boldsymbol{\Psi}_{ij})^2 | X_1 = x] \\ &\leq C(mh)^{-1} \int K^2(u) (f_1(z + hu) - f_1(z))^2 p_{X_1}(z + hu) du \\ &= C(mh)^{-1} \int K^2(u) (f_1'(z)uh + o(uh))^2 (p_{X_1}(z) + p'_{X_1}(z)uh + o(uh)) du \\ &= Cm^{-1} [f_1'(z)]^2 p_{X_1}(z) \int u^2 K^2(u) du \cdot h + o(m^{-1}h) = Chm^{-1}. \end{aligned}$$

The uniform upper bound of $K_h(x - z) \Delta f_z(x)$ can be studied under two cases: (1) x is out of the support and (2) x is in the support. In particular, we have

- if $x \notin [z - h, z + h]$, then $K_h(x - z) \Delta f_z^2(x) = 0$;
- if $x \in [z - h, z + h]$, then, by mean value theorem,

$$K_h(x - z) \Delta f_z(x) \leq h^{-1} K(h^{-1}(x - z)) |f(x) - f(z)| \leq h^{-1} \|K\|_{\infty} \|f'\|_{\infty}^2 \cdot (2h) = 4 \|f_1'\|_{\infty}^2 \|K\|_{\infty}.$$

Combining with the fact that $|\mathbf{v}_k^T \boldsymbol{\Psi}_{ij}| \leq \sqrt{m}$ for any i, j, k , we conclude that $g \leq U := 4 \|f_1'\|_{\infty}^2 \|K\|_{\infty} \sqrt{m}$ for any $g \in \mathcal{G}_h'''$. Therefore by Lemma H.2 and $M = 6^m$, we have

$$\begin{aligned} \mathbb{E}V_n(z) &\leq C \sqrt{\frac{h \log(dMh^{-1})}{mn}} + C \frac{\sqrt{m} \cdot \log(dMh^{-1})}{n} \\ &= C \sqrt{\frac{h \log(dh^{-1})}{n}} + C \frac{m^{3/2} \log(dh^{-1})}{n}. \end{aligned} \tag{E.23}$$

Similar to the analysis of σ_P^2 , we also expand the expectation of the process as the integration and use the Taylor expansion to bound it as follows

$$\begin{aligned}
& \mathbb{E}[K_h(X_1 - z)\Delta f_z(X_1)\mathbf{v}_k^T \Psi_{ij}] \\
&= h^{-1} \int K\left(\frac{x-z}{h}\right) (f_1(x) - f_1(z))p_{X_1}(x)dx \cdot \mathbb{E}[\mathbf{v}_k^T \Psi_{1j} | X_1 = x] dx \\
&= \int K(u)[f_1'(z)uh + f_1''(z)(uh)^2/2 + o(uh)^2][p_{X_1}(z) + p'_{X_1}(z)uh + o(uh)] \\
&\quad \cdot \left(\mathbb{E}[\mathbf{v}_k^T \Psi_{1j} | X_1 = z] + uh \frac{d}{dz} \mathbb{E}[\mathbf{v}_k^T \Psi_{1j} | X_1 = z] + o(uh) \right) du \leq Ch^2/\sqrt{m}.
\end{aligned} \tag{E.24}$$

The last inequality is due to the fact that $K(\cdot)$ is an even function, $\|\psi_{jk}\|_\infty \leq 1$ and from (E.10). Moreover, the constant is independent to j, k and z . Using Lemma H.4, we have

$$\mathbb{P}\left(V_n(z) - \mathbb{E}V_n(z) > t\sqrt{2(\sigma_P^2 + 2U\mathbb{E}V_n(z))} + \frac{2Ut^2}{3}\right) \leq \exp(-nt^2). \tag{E.25}$$

Combining (E.23) and (E.24) with (E.25) for $t = \sqrt{\log n/n}$, with probability at least $1 - 1/n$, we have

$$\begin{aligned}
\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \boldsymbol{\xi}_z\|_2 &\leq 2 \max_{j \geq 2, k \in [M]} \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(X_{1i} - z) \Delta f_z(X_{1i}) \mathbf{v}_k^T \Psi_{ij} \\
&\leq V_n(z) + \max_{j,k} \sup_{z \in \mathcal{X}} \mathbb{E}[K_h(X_1 - z) \Delta f_z(X_1) \mathbf{v}_k^T \Psi_{ij}] \\
&\leq C \sqrt{\frac{h \log(dh^{-1})}{n}} + C \frac{m^{3/2} \log(dh^{-1})}{n} + C \frac{h^2}{\sqrt{m}},
\end{aligned} \tag{E.26}$$

where the last equality is because of $n^{-1}h = o(1)$.

Now we bound $\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \boldsymbol{\zeta}_z\|_2$. The procedure is similar to the first part of the proof. We again apply the 1/2-covering of \mathbb{B}^d so that

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \boldsymbol{\zeta}_z\|_2 \leq 2 \max_{j \geq 2, k \in [M]} \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \zeta_i(z) \mathbf{v}_k^T \Psi_{ij}.$$

Motivated by the above argument, we now turn to study the following empirical process

$$V'_n(z) = \max_{j \geq 2, k \in [M]} \sup_{z \in \mathcal{X}} \left\{ \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \zeta_i(z) \mathbf{v}_k^T \Psi_{ij} - \mathbb{E}[K_h(X_{11} - z) \zeta_i(z) \mathbf{v}_k^T \Psi_{ij}] \right\}$$

and the function class inspired from the above empirical process

$$\mathcal{G}_h'''' = \left\{ g_z(x_1, x_2) = h^{-1} K((x_1 - z)/h) \zeta_i(z) \sum_{t=1}^m \mathbf{v}_{kt} \psi_t(x_j) \mid j \geq 2, k \in [M], z, x_1, x_j \in \mathcal{X} \right\}.$$

Our method of bound the supreme of the process is same as the proof of previous lemmas. We need to study the covering number of the function space. Assembling the concentration inequality of the suprema with the upper bound of the expectation of suprema and suprema of the expectation, we will arrive at the final bound. Therefore, we first bound the covering number

$$\sup_Q N(\mathcal{G}_h''', L^2(Q), \epsilon) \leq dM \left(\frac{2\sqrt{m} \|K\|_{\text{TV}} A}{h\epsilon} \right)^4.$$

Using Definition 4.1, there exists $L_1(z, x_{\setminus 1})$ such that the approximation error is bounded by

$$|\zeta_i(z)| = \left| f(X_{i1}, \dots, X_{id}) - \sum_{j=1}^d f_{jz}(X_{ij}) \right| \leq |L_1(z, x_{\setminus 1})| \cdot |X_{i1} - z| + |U_j(z)|(X_{i1} - z)^2.$$

Therefore, the variance of the process V_n' can be bounded by computing the expectation

$$\sigma_P^2 := \mathbb{E}[K_h(X_1 - z) \zeta_i(z) \mathbf{v}_k^T \boldsymbol{\Psi}_{ij}]^2 \leq Chm^{-1} \quad (\text{E.27})$$

and $g \leq U := 4\|L_1(z, x_{\setminus 1})\|_\infty^2 \|K\|_\infty \sqrt{m}$ for all $g \in \mathcal{G}_h''''$. Lemma H.2 gives us

$$\mathbb{E}V_n'(z) \leq C \sqrt{\frac{h \log(dh^{-1})}{n}} + C \frac{m^{3/2} \log(dh^{-1})}{n}. \quad (\text{E.28})$$

Denote $\kappa(x) = \mathbb{E}[L_1(z, X_{\setminus 1}) \mid X_1 = x]$. Using Definition 4.1,

$$\begin{aligned} & \mathbb{E}[K_h(X_1 - z) \zeta_i(z) \mathbf{v}_k^T \boldsymbol{\Psi}_{ij}] \\ &= h^{-1} \int K\left(\frac{x - z}{h}\right) (\mathbb{E}[L_1(z, X_{\setminus 1}) \mid X_1 = x](x_1 - z) + U_j(z)(x - z)^2) \\ & \quad \cdot \mathbb{E}[\mathbf{v}_k^T \boldsymbol{\Psi}_{1j} \mid X_1 = x] p_{X_1}(x) dx \\ &= \int K(u) (\kappa(z)uh + (\kappa''(z) + \|U_j\|_\infty)(uh)^2/2 + o(uh)^2) (p_{X_1}(z) + p'_{X_1}(z)uh + o(uh)) \\ & \quad \cdot \left(\mathbb{E}[\mathbf{v}_k^T \boldsymbol{\Psi}_{1j} \mid X_1 = z] + uh \frac{d}{dz} \mathbb{E}[\mathbf{v}_k^T \boldsymbol{\Psi}_{1j} \mid X_1 = z] + o(uh) \right) du \leq Ch^2/\sqrt{m}. \end{aligned} \quad (\text{E.29})$$

The last inequality is due to Assumption **(A1)**. Since \mathcal{X} is compact, $\mathbb{E}[\mathbf{v}_k^T \boldsymbol{\Psi}_{1j} | X_1 = z]$ and $\kappa(z)$ are uniformly bounded on $z \in \mathcal{X}$. According to Lemma **H.2** and Lemma **H.4**, similar to the first part of the proof, (E.27), (E.28) and (E.29) can yield that for some constant C , with probability at least $1 - 1/d$, we can bound the suprema

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\boldsymbol{\Psi}_{\bullet j}^T \mathbf{W}_z \boldsymbol{\zeta}_z\|_2 \leq C \sqrt{\frac{h \log(dh^{-1})}{n}} + C \frac{m^{3/2} \log(dh^{-1})}{n} + C \frac{h^2}{\sqrt{m}}. \quad (\text{E.30})$$

Combining (E.26) and (E.30), we have the rate of $\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\boldsymbol{\Psi}_{\bullet j}^T \mathbf{W}_z (\boldsymbol{\xi}_z + \boldsymbol{\zeta}_z)\|_2$.

For the case when $j = 1$, $\boldsymbol{\Psi}_{\bullet 1} = (1, \dots, 1)^T \in \mathbb{R}^n$ and we can follow similar procedure to derive

$$\sup_{z \in \mathcal{X}} \frac{1}{n} \|\boldsymbol{\Psi}_{\bullet 1}^T \mathbf{W}_z (\boldsymbol{\xi}_z + \boldsymbol{\zeta}_z)\|_2 = \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(z - X_{i1}) (\xi_i(z) + \zeta_i(z)) = O_P(h^2 + \sqrt{h/n}).$$

The final step is to bound $\sup_{z \in \mathcal{X}} \frac{1}{n} \|W_z^{1/2} \boldsymbol{\xi}_z^2\|_2^2$ and $\sup_{z \in \mathcal{X}} \frac{1}{n} \|W_z^{1/2} \boldsymbol{\zeta}_z^2\|_2^2$. We just repeat the procedure again and consider $V_n'''(z) = \sup_{z \in \mathcal{X}} n^{-1} \sum_{i=1}^n K_h(X_{i1} - z) \xi_i(z) - \mathbb{E}[K_h(X_{11} - z) \xi_i(z)]$. First, we find that

$$\mathbb{E} V_n'''(z) \leq C \sqrt{\frac{h^3 \log(h^{-1/2})}{n}} + C \frac{h \log(h^{-1})}{n}. \quad (\text{E.31})$$

Next, we have the upper bound of the supreme of the expectation

$$\begin{aligned} \sup_{z \in \mathcal{X}} \mathbb{E}[K_h(X_1 - z) \xi_i^2(z)] &= h^{-1} \int K\left(\frac{x-z}{h}\right) (f_1(x) - f_1(z))^2 p_{X_1}(x) dx \\ &= \int K(u) (f_1'(z) u h + o(uh))^2 (p_{X_1}(z) + p'_{X_1}(z) u h + o(uh)) du \\ &= [f_1'(z)]^2 p_{X_1}(z) \int u^2 K(u) du \cdot h^2 + o(h^2) \leq C h^2. \end{aligned} \quad (\text{E.32})$$

Combining (E.31) and (E.32) with Lemma **H.4** with $t = \log n / \sqrt{n}$, with probability at least $1 - 1/n$,

$$\begin{aligned} \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \xi_i^2(z) &\leq V_n(z) + \sup_{z \in \mathcal{X}} \mathbb{E}[K_h(X_{11} - z) \xi_i^2(z)] \\ &\leq C \sqrt{\frac{h^3 \log(nh^{-1})}{n}} + C \frac{h^{5/4} \log(nh^{-1})}{n^{3/4}} + C \frac{h \log^2 n}{n} + C h^2 = O(h^2). \end{aligned}$$

Similarly, we also have $\sup_{z \in \mathcal{X}} n^{-1} \sum_{i=1}^n K_h(X_{i1} - z) \zeta_i^2(z) = o_P(h^2)$.

E.4 Proof of Lemma A.4

For $j = 2, \dots, n$, we define the process

$$G_n(z, k, j) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{h} K\left(\frac{X_{i1} - z}{h}\right) \psi_{jk}(X_{ji}) \varepsilon_i.$$

Since ε_i are subgaussian random variables, we have $\mathbb{P}(\max_i |\varepsilon_i| > C\sqrt{\log n}) \leq 1/n$. Conditioning on the event $\mathcal{A} = \{\max_i |\varepsilon_i| < C\sqrt{\log n}\}$, we can apply the Mc'Diarmid's inequality to obtain

$$\mathbb{P}\left(\max_{j,k} \sup_{z \in \mathcal{X}} G_n(z, k, j) - \mathbb{E}\left[\max_{j,k} \sup_{z \in \mathcal{X}} G_n(z, k, j) \mid \mathcal{A}\right] > t \mid \mathcal{A}\right) \leq \exp\left(-C \frac{nh^2 t^2}{\log^2 n}\right). \quad (\text{E.33})$$

Next, we bound $\mathbb{E}[\max_{j,k} \sup_{z \in \mathcal{X}} G_n(z, k, j) \mid \mathcal{A}]$. Using Dudley's entropy integral (see Corollary 2.2.5 in [van der Vaart and Wellner \(1996\)](#)), conditioning on $\{X_{ij}\}_{i \in [n], j \in [d]}$, we have with probability $1 - 1/n$, there exists a constant C such that

$$\mathbb{E}\left[\max_{1 \leq j \leq d} \max_{1 \leq k \leq m} \sup_{z \in \mathcal{X}} G_n(z, k, j) \mid \mathcal{A}\right] \leq \mathbb{E}\left[\int_0^{\sigma_n} \sqrt{\log N(\mathcal{G}'_h, L^2(\widehat{\mathbb{P}}_n), \epsilon)} d\epsilon \mid \mathcal{A}\right],$$

where $\widehat{\mathbb{P}}_n = n^{-1} \sum_{i=1}^n \delta_{X_{i1}, \dots, X_{id}}$, $\sigma_n = \max_{1 \leq j \leq d} \max_{1 \leq k \leq m} \sup_{z \in \mathcal{X}} \widehat{\mathbb{P}}_n[K_h(\cdot - z) \psi_{jk}(\cdot)]^2$ and

$$\mathcal{G}'_h = \{g_z(x_1, x_2) = h^{-1} K(h^{-1}(x_1 - z)) \psi_{jk}(x_2) \mid 1 \leq k \leq m, z \in \mathcal{X}, x_1, x_2 \in \mathcal{X}\}.$$

From Lemma H.3 and similar to the previous computation on the covering number, for any measure Q , we have the uniform upper bound of covering number as

$$\sup_Q N(\mathcal{G}'_h, L^2(Q), \epsilon) \leq dm \left(\frac{2\|K\|_{\text{TV}} A}{h\epsilon} \right)^4.$$

Following a similar argument as in the proof of Lemma A.1, we bound the variance of process by Cauchy-Schwarz inequality as

$$\begin{aligned} \sigma_P^2 &:= \mathbb{E}\left[\left(K_h(X_1 - z) \psi_{jk}(X_j) \mid \mathcal{A}\right)^2\right] \\ &\leq h^{-2} \mathbb{E}\left[K^2(h^{-1}(X_1 - z)) \mid \mathcal{A}\right] \mathbb{E}\left[\psi_{jk}^2(X_j) \mid X_1\right] \leq b\|K\|_\infty^2 (mh)^{-1}, \end{aligned}$$

and $g \leq \|K\|_\infty h^{-1}$ for any $g \in \mathcal{G}'_h$. Since $m(nh)^{-1} = o(1)$, we have $n\sigma_P^2 \geq h^{-2} \log(dm(2\|K\|_{\text{TV}}A/(h\sigma)))$.

Therefore, by Lemma H.2, we derive that

$$\mathbb{E} \left[\int_0^{\sigma_n} \sqrt{\log N(\mathcal{G}'_h, L^2(\widehat{\mathbb{P}}_n), \epsilon) d\epsilon} \mid \mathcal{A} \right] \leq C\sigma_P \sqrt{\log(dm(h\sigma_P)^{-4})} \leq C\sqrt{\frac{\log(dm^3h^{-2})}{mh}}$$

$$\text{and } \mathbb{E} \left[\max_{j,k} \sup_{z \in \mathcal{X}} G_n(z, k, j) \mid \mathcal{A} \right] \leq C\sqrt{\log(dm^3h^{-2})/(mh)}.$$

Choosing $t = C \log^2 n / (\sqrt{nh})$ in (E.33), we have

$$\mathbb{P} \left(\max_{j,k} \sup_{z \in \mathcal{X}} G_n(z, k, j) > C\sqrt{\frac{\log(dm^3h^{-2})}{mh}} + \frac{\log^2 n}{\sqrt{nh}} \right)$$

$$\leq \mathbb{P} \left(\max_{j,k} \sup_{z \in \mathcal{X}} G_n(z, k, j) > C\sqrt{\frac{\log(dm^3h^{-2})}{mh}} + \frac{\log^2 n}{\sqrt{nh}} \mid \mathcal{A} \right) + \mathbb{P}(\mathcal{A}^c) \leq 2/n.$$

Therefore, when $m(nh)^{-1} = o(1)$, with probability $1 - 2/n$,

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} \frac{1}{n} \|\Psi_{\bullet j}^T \mathbf{W}_z \boldsymbol{\varepsilon}\|_2 \leq \sqrt{\frac{m}{n}} \max_{1 \leq j \leq d} \max_{1 \leq k \leq m} \sup_{z \in \mathcal{X}} G_n(z, k, j) \leq 2C\sqrt{\frac{\log(dmh^{-1})}{nh}}.$$

When $j = 1$, recalling that $\Psi_{\bullet 1} = (1, \dots, 1)^T \in \mathbb{R}^n$, we have

$$\sup_{z \in \mathcal{X}} \frac{1}{n} \|\Psi_{\bullet 1}^T \mathbf{W}_z \boldsymbol{\varepsilon}\|_2 = \sup_{z \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^n K_h(z - X_{i1}) \varepsilon_i \right|,$$

and, similar to the case when $j \geq 2$, we can show that $\sup_{z \in \mathcal{X}} n^{-1} \|\Psi_{\bullet 1}^T \mathbf{W}_z \boldsymbol{\varepsilon}\|_2 \leq C\sqrt{\log(h^{-1})/(nh)}$

with probability $1 - 2/n$. This completes the proof.

F Auxiliary Lemmas for Bootstrap Confidence Bands

In this section, we describe the proof of these technical lemmas used in Section C. Section F.2 to Section F.6 provide the proofs of lemmas in Section C.1 supporting the proof of Theorem 3.7.

F.1 Proof of Lemma C.2

Recall the rate r_n of the estimated function shown in Theorem 3.2 is

$$r_n := \sqrt{\frac{s^2 \log(dmh^{-1})}{nm^{-2}h}} + \sqrt{\frac{s^3}{m^3}} + \frac{s \log(dh^{-1})}{nm^{-5/2}} + s\sqrt{mh^2}.$$

We first establish a lemma on the estimation error of $\widehat{\varepsilon}_i$.

Lemma F.1. Let $\widehat{\varepsilon}_i = Y_i - \widehat{f}(X_{i1}, \dots, X_{id})$ for $i = 1, \dots, n$. Under Assumption (A4), we have

$$\mathbb{P}\left(\max_{i \in [n]} |\widehat{\varepsilon}_i - \varepsilon_i| < 2Cr_n\sqrt{m}\right) \geq 1 - \frac{1}{n}.$$

If $h \asymp n^{-\delta}$, $m \asymp n^\delta$ for $\delta > 1/5$, we have $r_n\sqrt{m} = o(n^{-1/5})$.

We defer the proof of the lemma to the end of this subsection. With the rate of $\max_{i \in [n]} |\widehat{\varepsilon}_i - \varepsilon_i|$, we can first bound the rate of $\widehat{\sigma}^2 - \sigma^2$. Using the triangle inequality, we have

$$|\widehat{\sigma}^2 - \sigma^2| \leq \underbrace{\frac{1}{n} \sum_{i=1}^n (\widehat{\varepsilon}_i - \varepsilon_i)^2}_{\text{I}} + \underbrace{\frac{2}{n} \sum_{i=1}^n |(\widehat{\varepsilon}_i - \varepsilon_i)\varepsilon_i|}_{\text{II}} + \underbrace{\left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 - \sigma^2 \right|}_{\text{III}}. \quad (\text{F.1})$$

From Lemma F.1, we have the convergence rate of the noise estimator

$$\mathbb{P}(\text{I} > 4cr_n^2m) \leq \mathbb{P}\left(\max_{i \in [n]} |\widehat{\varepsilon}_i - \varepsilon_i|^2 > 4cr_n^2m\right) \leq 1/n. \quad (\text{F.2})$$

Under Assumption (A4), ε_i are subgaussian random variables with variance-proxy σ_ε^2 . Using Bernstein's inequality, we have

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 - \sigma^2\right| > C_1 \sqrt{\frac{\sigma_\varepsilon^2 \log n}{n}}\right) \leq \frac{2}{n} \quad \text{and} \quad \mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n |\varepsilon_i| - \mathbb{E}|\varepsilon|\right| > c \sqrt{\frac{\sigma_\varepsilon^2 \log n}{n}}\right) \leq \frac{2}{n}. \quad (\text{F.3})$$

Suppose n is large enough, so that $\mathbb{E}|\varepsilon| \leq \sqrt{\sigma_\varepsilon^2 \log n/n}$. We now can bound the second term by

$$\begin{aligned} & \mathbb{P}(\text{II} > 2c(c+1)\mathbb{E}|\varepsilon|r_n\sqrt{m}) \\ & \leq \mathbb{P}\left(\max_{i \in [n]} |\widehat{\varepsilon}_i - \varepsilon_i| > 2cr_n\sqrt{m}\right) + \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n |\varepsilon_i| > \mathbb{E}|\varepsilon| + c\sqrt{\frac{\sigma_\varepsilon^2 \log n}{n}}\right) \leq \frac{3}{n}. \end{aligned}$$

Applying the fact that $\mathbb{E}|\epsilon_i| \leq \sigma^2$, we have the upper bound of the third term as

$$\mathbb{P}(\text{III} > 2c(c+1)\sigma^2 r_n \sqrt{m}) \leq 3/n. \quad (\text{F.4})$$

Combining (F.2), (F.3), (F.4) with (F.1), we have the estimation rate of the variance of noise as

$$\mathbb{P}(|\hat{\sigma}^2 - \sigma^2| \geq C_1 r_n \sqrt{m}) \leq 6/n. \quad (\text{F.5})$$

Now we come back to prove Lemma F.1.

Proof of Lemma F.1. Recall that the estimator of the true function is

$$\hat{f}(x_1, \dots, x_d) = \hat{f}_1(x_1) + \sum_{j=2}^d \sum_{k=1}^m \hat{\beta}_{jk} \psi_{jk}(x_j).$$

Similar to Lemma A.2, let $\delta_i = \sum_{j=2}^d f_j(X_{ji}) - f_{mj}(X_{ji})$ and the B-spline theory (see Lemma 1, Huang et al. (2010)) that $\delta_i^2 \leq sm^{-2\gamma}$. Define the event

$$\mathcal{E} = \left\{ \sup_{z \in \mathcal{X}} \left\{ \sqrt{m} |\hat{a}_z - f_1(z)| + \sum_{i=2}^d \|\hat{\beta}_i - \beta_i\|_2 \right\} \leq Cr_n \right\}.$$

From Theorem 3.2, we have $\mathbb{P}(\mathcal{E}) \geq 1 - 1/n$. Conditioning the event \mathcal{E} , we have

$$\begin{aligned} \max_{i \in [n]} |\hat{\varepsilon}_i - \varepsilon_i| &= \max_{i \in [n]} \left| f(X_{i1}, \dots, X_{id}) - \hat{f}(X_{i1}, \dots, X_{id}) \right| \\ &\leq \max_{i \in [n]} \left| \hat{f}_1(X_{i1}) - f_1(X_{i1}) \right| + \max_{i \in [n]} \left| \sum_{j=2}^d \sum_{k=1}^m (\hat{\beta}_{jk} - \beta_{jk}) \psi_{jk}(X_{ij}) \right| + \max_{i \in [n]} |\delta_i| \\ &\leq \sup_{z \in \mathcal{X}} |\hat{f}_1(z) - f_1(z)| + \sqrt{m} \sum_{i=2}^d \|\hat{\beta}_i - \beta_i\|_2 + \sqrt{sm}^{-\gamma} \leq 2C\sqrt{m}r_n, \end{aligned}$$

where the second inequality is because of Hölder inequality as well as the fact that $\psi_{jk} \leq 1$ for all j, k and the last inequality is since we are conditioning on \mathcal{E} . \square

F.2 Proof of Lemma 3.4

The high level idea of proving Lemma 3.4 is similar to the proof of Lemma E.1. We aim to bound the rate of $\sup_z \|\widehat{\Sigma}_z \boldsymbol{\theta}_z - \Sigma_z \boldsymbol{\theta}_z\|_\infty$. Therefore, we consider the random variable

$$\widetilde{Z}_{kj} = \sup_{z \in \mathcal{X}} \|\widehat{\Sigma}_z \boldsymbol{\theta}_z - \Sigma_z \boldsymbol{\theta}_z\|_\infty = \sup_{z \in \mathcal{X}} (\mathbb{E}_n - \mathbb{E}) \left[K_h(X_{i1} - z) \psi_{jk}(X_{ij}) \sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(X_{ij'}) (\boldsymbol{\theta}_z)_{j'k'} \right].$$

Recall that when j or k equals to 1, $\psi_{jk} \equiv 1$. Similar to the proof of Lemma E.1, we have three cases: (1) $j = k = 1$, (2) only one of j or k equals to 1 and (3) neither of j, k equals to 1. We only analyze the hardest case (3) in this proof and we can deal with the first two cases through a similar procedure. For the minor differences among the analysis of these three cases, we refer to the proof of Lemma E.1.

We first study the covering number of the space

$$\overline{\mathcal{G}}_h = \left\{ h^{-1} K(h^{-1}(x_1 - z)) \psi_{jk}(x_j) \sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(x_{j'}) (\boldsymbol{\theta}_z)_{j'k'} \mid z \in \mathcal{X}, j, k \in [d] \right\}.$$

Since $\overline{\mathcal{G}}_h$ can be decomposed into a production of a few functions, we aim to apply Lemma H.1 to bound its covering number. Lemma H.3 gives us the covering number of $\{h^{-1} K(h^{-1}(\cdot - z)) \mid z \in \mathcal{X}\}$, it remains to bound the covering number of

$$\overline{\mathcal{G}}_h^{(1)} = \left\{ \overline{g}_z(\mathbf{x}) := \sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(x_{j'}) (\boldsymbol{\theta}_z)_{j'k'} \mid z \in \mathcal{X} \right\}.$$

Given any $z \in \mathcal{X}$, we can find a \tilde{z} such that $|z - \tilde{z}| \leq \epsilon$. We then have given any measure Q ,

$$\begin{aligned} \|\overline{g}_z - \overline{g}_{\tilde{z}}\|_{L^2(Q)}^2 &= \mathbb{E}_Q \left[\sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(X_{ij'}) [(\boldsymbol{\theta}_z)_{j'k'} - (\boldsymbol{\theta}_{\tilde{z}})_{j'k'}] \right]^2 \\ &\leq L^2 \|\boldsymbol{\theta}_z - \boldsymbol{\theta}_{\tilde{z}}\|_1^2 \leq L^2 d \|\boldsymbol{\theta}_z - \boldsymbol{\theta}_{\tilde{z}}\|_2^2 \leq 2L^2 d m \rho_{\min}^{-2}(B/b) L \rho_{\max} \cdot \epsilon^2, \end{aligned}$$

where the last inequality is due to Lemma F.4. Therefore,

$$\sup_Q \left(\overline{\mathcal{G}}_h^{(1)}, L^2(Q), \epsilon \right) \leq \sqrt{2L^2 d m \rho_{\min}^{-2}(B/b) L |\mathcal{X}|^2 \rho_{\max} / \epsilon}. \quad (\text{F.6})$$

According to Corollary 8 in Chapter XI of [de Boor \(2001\)](#), we have

$$\sup_{z, \mathbf{x}} \left| \sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(x_{j'}) (\boldsymbol{\theta}_z)_{j'k'} \right| \leq L \sum_{j=1}^d \sup_z \|(\boldsymbol{\theta}_z)_{j\bullet}\|_\infty = L\sqrt{d} \sup_z \|\boldsymbol{\theta}_z\|_2 \leq 4L\rho_{\min}^{-1} m\sqrt{d}, \quad (\text{F.7})$$

where the last inequality is due to Lemma [F.3](#). By Lemma [H.1](#), combining [\(F.6\)](#), [\(F.7\)](#) and Lemma [H.3](#), we have

$$\sup_Q N(\bar{\mathcal{G}}_h, L^2(Q), \epsilon) \leq d^2 \left(\frac{C|\mathcal{X}|md}{\rho_{\min}^2 b h \epsilon} \right)^5. \quad (\text{F.8})$$

We then consider the envelop function of $\bar{\mathcal{G}}_h$ as

$$\bar{F}(\mathbf{x}) = 4h^{-1} \|K\|_\infty \sup_z \left| \sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(x_{j'}) (\boldsymbol{\theta}_z)_{j'k'} \right|.$$

In order to study $\bar{F}(\mathbf{x})$, we define $\bar{\boldsymbol{\Sigma}} = \mathbb{E}[\boldsymbol{\Psi}_{1\bullet} \boldsymbol{\Psi}_{1\bullet}^T]$ and $\bar{\boldsymbol{\theta}} = \bar{\boldsymbol{\Sigma}}^{-1} \mathbf{e}_1$. We decompose $\bar{F}(\mathbf{x})$ into

$$\begin{aligned} \bar{F}^{(1)}(\mathbf{x}) &= 4h^{-1} \|K\|_\infty \left| \sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(x_{j'}) (\bar{\boldsymbol{\theta}})_{j'k'} \right| \text{ and} \\ \bar{F}^{(2)}(\mathbf{x}) &= 4h^{-1} \|K\|_\infty \sup_z \left| \sum_{j'=1}^d \sum_{k'=1}^m \psi_{j'k'}(x_{j'}) (\boldsymbol{\theta}_z - \bar{\boldsymbol{\theta}})_{j'k'} \right|. \end{aligned}$$

According to Lemma [F.2](#), we have

$$\|\bar{F}^{(1)}\|_{L^2(\mathbb{P})}^2 \leq 4mh^{-1} \|K\|_\infty \|\bar{\boldsymbol{\theta}}\|_2^2 \leq 4h^{-1} \|K\|_\infty \rho_{\min}.$$

Similarly, we also have

$$\|\bar{F}^{(2)}\|_{L^2(\mathbb{P})}^2 \leq 4mh^{-1} \|K\|_\infty \sup_z \|\boldsymbol{\theta}_z - \bar{\boldsymbol{\theta}}\|_2^2 \leq 8h^{-1} \|K\|_\infty \rho_{\min}.$$

Therefore, we have $\sigma_P^2 \leq \|F\|_{L^2(\mathbb{P})}^2 \leq 32h^{-1} \|K\|_\infty \rho_{\min}$. By Lemma [H.2](#) we have

$$\mathbb{E} \left[\max_{k,j} \tilde{Z}_{kj} \right] \leq C_1 \sqrt{\frac{(\log^2 d) \log(|\mathcal{X}|/\rho_{\min})}{nh}}. \quad (\text{F.9})$$

We can also apply Lemma H.5 to obtain

$$\mathbb{P}\left(\sqrt{n} \max_{k,j} \tilde{Z}_{kj} \geq 2\sqrt{n}\mathbb{E}[\max_{k,j} \tilde{Z}_{kj}] + Ch^{-1/2}\sqrt{t} + Ch^{-1/2}t\right) \leq t^{-1}. \quad (\text{F.10})$$

Combining (F.9) with (F.10), we have

$$\max_{k,j} \tilde{Z}_{kj} = O_P\left(\sqrt{(\log^2 d) \log(|\mathcal{X}|/\rho_{\min})/nh}\right).$$

Finally, we finish the proof of the lemma by

$$\|\widehat{\Sigma}_z \boldsymbol{\theta}_z - \mathbf{e}_1\|_{2,\infty} \leq \sqrt{m} \|\widehat{\Sigma}_z \boldsymbol{\theta}_z - \mathbf{e}_1\|_{\infty} = O_P\left(\sqrt{m(\log^2 d) \log(|\mathcal{X}|/\rho_{\min})/nh}\right).$$

F.3 Auxiliary Lemmas for Constraint Rate

In this section, we prove some auxiliary lemmas needed in the proof of Lemma 3.4.

Lemma F.2. Under Assumptions (A1), (A2) and (A6), there exists a constant $\rho_{\max} < \infty$ such that for any $\boldsymbol{\beta}_+ \in \mathbb{R}^{1+(d-1)m}$,

$$\frac{\boldsymbol{\beta}_+^T \mathbb{E}[\boldsymbol{\Psi}_{1\bullet} \boldsymbol{\Psi}_{1\bullet}^T] \boldsymbol{\beta}_+}{\|\boldsymbol{\beta}_+\|_2^2} \leq \frac{3\rho_{\max}}{2m}. \quad (\text{F.11})$$

Proof. We first derive some inequalities from (3.10). Given any $j \neq k \geq 2$, let $u_j(x_j) = \sum_{s=1}^m \boldsymbol{\beta}_{js} \psi_{js}(x_j)$, $\Delta_{jk}(x_j, x_k) := |p_{j,k}(x_j, x_k) - p_j(x_j)p_j(x_k)|$, and we have

$$\begin{aligned} |\boldsymbol{\beta}_+^T \mathbb{E}[\boldsymbol{\Psi}_{1j} \boldsymbol{\Psi}_{1k}^T] \boldsymbol{\beta}_+| &= \left| \iint u_j(x_j) u_k(x_k) p_{jk}(x_j, x_k) dx_j dx_k \right| \\ &\leq \left| \int u_j(x_j) p_j(x_j) dx_j \int u_k(x_k) p_k(x_k) dx_k \right| + \iint |u_j(x_j) u_k(x_k)| \Delta_{jk}(x_j, x_k) dx_j dx_k \\ &\leq m^{-1} \|\boldsymbol{\beta}_j\|_2 \|\boldsymbol{\beta}_k\|_2 \cdot \|\Delta_{jk}\|_2, \end{aligned} \quad (\text{F.12})$$

where the last inequality is due to $\mathbb{E}[\psi_{jk}(X_j)] = 0$ and (A.10). Therefore, we have

$$\begin{aligned} \boldsymbol{\beta}_+^T \mathbb{E}[\boldsymbol{\Psi}_{1\bullet} \boldsymbol{\Psi}_{1\bullet}^T] \boldsymbol{\beta}_+ &= \sum_{j=1}^d \boldsymbol{\beta}_j^T \mathbb{E}[\boldsymbol{\Psi}_{1j} \boldsymbol{\Psi}_{1j}^T] \boldsymbol{\beta}_j + \sum_{j \neq k} \boldsymbol{\beta}_j^T \mathbb{E}[\boldsymbol{\Psi}_{1j} \boldsymbol{\Psi}_{1k}^T] \boldsymbol{\beta}_k \\ &\leq \frac{\rho_{\max}}{m} \sum_{j=1}^d \|\boldsymbol{\beta}_j\|_2^2 + \sum_{j \neq k} m^{-1} \|\boldsymbol{\beta}_j\|_2 \|\boldsymbol{\beta}_k\|_2 \cdot \|\Delta_{jk}\|_2 \leq \frac{3\rho_{\max}}{2m} \|\boldsymbol{\beta}\|_2^2, \end{aligned}$$

where the last inequality is due to Assumption **(A6)** and the Gershgorin circle theorem. \square

Lemma F.3. Under Assumptions **(A1)**, **(A2)**, **(A3)** and **(A6)**, there exists a constant $\rho_{\max} < \infty$ such that for any $z \in \mathcal{X}$ and any $\beta_+ \in \mathbb{R}^{1+(d-1)m}$,

$$\frac{\rho_{\min}}{2m} \leq \frac{\beta_+^T \Sigma_z \beta_+}{\|\beta_+\|_2^2} \leq \frac{3\rho_{\max}}{2m} \text{ and } \sup_z \|\theta_z\|_2 \leq \frac{2m}{\rho_{\min}}. \quad (\text{F.13})$$

Proof. We first derive some inequalities from (3.10). Denote $\Delta_{1jk}(x_1, x_j, x_k) := |p_{1,j,k}(x_1, x_j, x_k) - p_1(x_1)p_j(x_j)p_k(x_k)|$. Given any $j \neq k \geq 2$, let $u_j(x_j) = \sum_{s=1}^m \beta_{js} \psi_{js}(x_j)$, similar to (F.12), we have

$$\begin{aligned} \left| \beta_j^T \mathbb{E}[K_h(X_1 - z) \Psi_{1j} \Psi_{1k}^T] \beta_k \right| &= \left| \iiint K_h(x_1 - z) u_j(x_j) u_k(x_k) p_{1,j,k}(x_1, x_j, x_k) dx_1 dx_j dx_k \right| \\ &\leq m^{-1} \|\beta_j\|_2 \|\beta_k\|_2 \|\Delta_{1jk}\|_2. \end{aligned} \quad (\text{F.14})$$

Therefore, we have

$$\begin{aligned} \beta_+^T \Sigma_z \beta_+ &= \sum_{j=1}^d \beta_j^T \mathbb{E}[K_h(X_1 - z) \Psi_{1j} \Psi_{1j}^T] \beta_j + \sum_{j \neq k} \beta_j^T \mathbb{E}[K_h(X_1 - z) \Psi_{1j} \Psi_{1k}^T] \beta_k \\ &\leq \frac{\rho_{\max}}{m} \sum_{j=1}^d \|\beta_j\|_2^2 + \sum_{j \neq k} m^{-1} \|\beta_j\|_2 \|\beta_k\|_2 \|\Delta_{1jk}\|_2 \leq \frac{3\rho_{\max}}{2m} \|\beta\|_2^2, \end{aligned}$$

where the first inequality is due to (E.9) and the last inequality is due to the Gershgorin circle theorem. Similarly, we apply the Gershgorin circle theorem to obtain a lower bound as

$$\begin{aligned} \beta_+^T \Sigma_z \beta_+ &= \sum_{j=1}^d \beta_j^T \mathbb{E}[K_h(X_1 - z) \Psi_{1j} \Psi_{1j}^T] \beta_j + \sum_{j \neq k} \beta_j^T \mathbb{E}[K_h(X_1 - z) \Psi_{1j} \Psi_{1k}^T] \beta_k \\ &\geq \frac{\rho_{\min}}{m} \sum_{j=1}^d \|\beta_j\|_2^2 - \sum_{j \neq k} m^{-1} \|\beta_j\|_2 \|\beta_k\|_2 \|\Delta_{1jk}\|_2 \geq \frac{\rho_{\min}}{2m} \|\beta\|_2^2, \end{aligned}$$

where the first inequality is due to Assumption **(A3)**. \square

The following lemma shows the Lipschitz properties of θ_z .

Lemma F.4. Under Assumptions **(A1)** - **(A6)**, we have

$$\|\theta_z - \theta_{z'}\|_2 \leq 2m\rho_{\min}^{-2}(B/p_1(z))L\rho_{\max} \cdot |z - z'|,$$

where C_K is a constant only depending on the kernel K .

Proof. The idea of proving this lemma is similar to Lemma F.3. Given any $j \neq k \geq 2$, again let $u_j(x_j) = \sum_{s=1}^m \beta_{js} \psi_{js}(x_j)$ and we have

$$\begin{aligned}
& |\beta_j^T \mathbb{E}[(K_h(X_1 - z) - K_h(X_1 - z')) \Psi_{1j} \Psi_{1k}^T] \beta_k| \\
&= \left| \int (K_h(x_1 - z) - K_h(x_1 - z')) u_j(x_j) u_k(x_k) p_{1,j,k}(x_1, x_j, x_k) dx_1 dx_j dx_k \right| \\
&\leq \left| \int K(x_1) \sup_u \mathbb{E}[u_j(X_j) u_k(X_k) | X_1 = u] (p_1(z + x_1 h) - p_1(z' + x_1 h)) dx_1 \right| \\
&\leq b^{-1} L |z - z'| m^{-1} \|\beta_j\|_2 \|\beta_k\|_2 \|\Delta_{1jk}\|_2.
\end{aligned}$$

Similarly, we also have

$$\begin{aligned}
& |\beta_j^T \mathbb{E}[(K_h(X_1 - z) - K_h(X_1 - z')) \Psi_{1j} \Psi_{1j}^T] \beta_j| \\
&\leq \left| \int K(x_1) \sup_u \mathbb{E}[u_j(X_j)^2 | X_1 = u] (p_1(z + x_1 h) - p_1(z' + x_1 h)) dx_1 \right| \\
&\leq (B/b) L |z - z'| m^{-1} \|\beta_j\|_2^2.
\end{aligned}$$

Therefore, for any $\beta_+ \in \mathbb{R}^{1+(d-1)m}$ and z, z' ,

$$\begin{aligned}
\beta_+^T (\Sigma_z - \Sigma_{z'}) \beta_+ &\leq (B/b) L |z - z'| \cdot m^{-1} \sum_{j=1}^d \|\beta_j\|_2^2 + b^{-1} L |z - z'| \cdot \sum_{j \neq k} m^{-1} \|\beta_j\|_2 \|\beta_k\|_2 \|\Delta_{1jk}\|_2 \\
&\leq \frac{2(B/b) L \rho_{\max}}{m} \|\beta_+\|_2^2 \cdot |z - z'|.
\end{aligned}$$

Therefore, combining with Lemma F.3, we can apply the matrix inverse perturbation inequality (see e.g., Demmel (1992)) and have

$$\|\theta_z - \theta_{z'}\|_2 \leq \|\Sigma_z^{-1}\|_2^2 \|\Sigma_z - \Sigma_{z'}\|_2 \leq 2m \rho_{\min}^{-2} (B/b) L \rho_{\max} \cdot |z - z'|.$$

□

F.4 Proof of Lemma C.4

Applying the fact that $\widehat{\boldsymbol{\theta}}_z^T \widehat{\boldsymbol{\Sigma}}_z \widehat{\boldsymbol{\theta}}_z \leq \boldsymbol{\theta}_z^T \widehat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta}_z$, we have the following inequality

$$\begin{aligned}
(\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z)^T \widehat{\boldsymbol{\Sigma}}_z (\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z) &= \widehat{\boldsymbol{\theta}}_z^T \widehat{\boldsymbol{\Sigma}}_z \widehat{\boldsymbol{\theta}}_z - 2\widehat{\boldsymbol{\theta}}_z^T \widehat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta}_z + \boldsymbol{\theta}_z^T \widehat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta}_z \\
&\leq 2\boldsymbol{\theta}_z^T \widehat{\boldsymbol{\Sigma}}_z \boldsymbol{\theta}_z - 2(\widehat{\boldsymbol{\theta}}_z^T \widehat{\boldsymbol{\Sigma}}_z - e_i) \boldsymbol{\theta}_z - 2e_i^T \boldsymbol{\theta}_z \\
&= 2\boldsymbol{\theta}_z^T (\widehat{\boldsymbol{\Sigma}}_z - \boldsymbol{\Sigma}_z) \boldsymbol{\theta}_z - 2(\widehat{\boldsymbol{\theta}}_z^T \widehat{\boldsymbol{\Sigma}}_z - e_i) \boldsymbol{\theta}_z \\
&\leq 2\|\boldsymbol{\theta}_z\|_1^2 \|\widehat{\boldsymbol{\Sigma}}_z - \boldsymbol{\Sigma}_z\|_{\max} + 2\|\widehat{\boldsymbol{\Sigma}}_z \widehat{\boldsymbol{\theta}}_z - \mathbf{e}_1\|_{2,\infty} \|\boldsymbol{\theta}_z\|_1.
\end{aligned} \tag{F.15}$$

We now study the rate of $\boldsymbol{\theta}_z$ in this subsection. We separate $\boldsymbol{\Sigma}_z$ into four blocks such that

$$\boldsymbol{\Sigma}_z = \begin{pmatrix} \boldsymbol{\Sigma}_z^{(1,1)} & \boldsymbol{\Sigma}_z^{(2,1)T} \\ \boldsymbol{\Sigma}_z^{(2,1)} & \boldsymbol{\Sigma}_z^{(2,2)} \end{pmatrix},$$

where $\boldsymbol{\Sigma}_z^{(1,1)} \in \mathbb{R}$, $\boldsymbol{\Sigma}_z^{(2,1)} \in \mathbb{R}^{(d-1)m}$ and $\boldsymbol{\Sigma}_z^{(2,2)} \in \mathbb{R}^{(d-1)m \times (d-1)m}$. By Lemma F.3, both $[\boldsymbol{\Sigma}_z^{(1,1)}]^{-1}$ and $[\boldsymbol{\Sigma}_z^{(2,2)}]^{-1}$ exist for any $z \in \mathcal{X}$. By the inversion formula of a block matrix, we have

$$\boldsymbol{\Sigma}_z^{-1} = \begin{pmatrix} \boldsymbol{\Theta}_z^{(1,1)} & \boldsymbol{\Theta}_z^{(2,1)T} \\ \boldsymbol{\Theta}_z^{(2,1)} & \boldsymbol{\Theta}_z^{(2,2)} \end{pmatrix},$$

where the concrete formulations of these four submatrices are

$$\begin{aligned}
\boldsymbol{\Theta}_z^{(1,1)} &= \left(\boldsymbol{\Sigma}_z^{(1,1)} - [\boldsymbol{\Sigma}_z^{(2,1)}]^T [\boldsymbol{\Sigma}_z^{(2,2)}]^{-1} \boldsymbol{\Sigma}_z^{(2,1)} \right)^{-1}, \\
\boldsymbol{\Theta}_z^{(2,1)} &= -\boldsymbol{\Theta}_z^{(1,1)} [\boldsymbol{\Sigma}_z^{(2,2)}]^{-1} \boldsymbol{\Sigma}_z^{(2,1)}, \\
\boldsymbol{\Theta}_z^{(2,2)} &= [\boldsymbol{\Sigma}_z^{(2,2)}]^{-1} - \boldsymbol{\Theta}_z^{(2,1)} [\boldsymbol{\Sigma}_z^{(2,1)}]^T [\boldsymbol{\Sigma}_z^{(2,2)}]^{-1}.
\end{aligned}$$

In order to bound $\boldsymbol{\theta}_z = (\boldsymbol{\Theta}_z^{(1,1)}, \boldsymbol{\Theta}_z^{(2,1)T})^T$, we first bound the ℓ_1 norm of the second part

$$\|\boldsymbol{\Sigma}_z^{(2,1)}\|_1 = \sum_{j=2}^d \sum_{k=1}^m |\mathbb{E}[K_h(X_1 - z) \psi_{jk}(X_j)]|.$$

Following a similar analysis as (F.12), we can bound the norm by $\|\boldsymbol{\Sigma}_z^{(2,1)}\|_1 \leq \frac{\rho_{\max}}{2}$ and $\|[\boldsymbol{\Sigma}_z^{(2,2)}]^{-1} \boldsymbol{\Sigma}_z^{(2,1)}\|_1 \leq$

$(\rho_{\max}/\rho_{\min}) \cdot m/2$. In fact, by Lemma F.3, $\Theta_z^{(1,1)} > 0$. Combining with $\Sigma_z^{(1,1)} = \mathbb{E}[K_h(X_1 - z)] = p_1(z) + o(1)$, we can have $\Sigma_z^{(2,1)T} [\Sigma_z^{(2,2)}]^{-1} \Sigma_z^{(2,1)} \leq p_1(z) + o(1)$ for any $z \in \mathcal{X}$.

Summarizing the inequalities above, we have

$$\begin{aligned} \sup_{z \in \mathcal{X}} \|\boldsymbol{\theta}_z\|_1 &\leq \sup_{z \in \mathcal{X}} |\Theta_z^{(1,1)}| + \sup_{z \in \mathcal{X}} \|\Theta_z^{(2,1)}\|_1 \\ &= \sup_{z \in \mathcal{X}} \left| \Sigma_z^{(1,1)} - \Sigma_z^{(2,1)T} [\Sigma_z^{(2,2)}]^{-1} \Sigma_z^{(2,1)} \right|^{-1} + \sup_{z \in \mathcal{X}} \left\| \Theta_z^{(1,1)} [\Sigma_z^{(2,2)}]^{-1} \Sigma_z^{(2,1)} \right\|_1 \\ &= \sup_{z \in \mathcal{X}} \left\{ (p_1(z) + O(1))^{-1} + (p_1(z) + O(1))^{-1} \cdot O(\rho_{\min}^{-1} m^{3/2}) \right\} \leq C(b\rho_{\min})^{-1} m. \quad (\text{F.16}) \end{aligned}$$

Plugging (3.11), (F.16) and (E.8) into (F.15), we prove the lemma.

F.5 Proof of Lemma C.1

We can expand the difference between two processes as

$$\tilde{\mathbb{H}}'_n(z) - \tilde{Z}'_n(z) = \underbrace{\sqrt{n^{-1}h} \sum_{i=1}^n K_h(X_{i1} - z) \eta'_i \Psi_{i\bullet}^T \hat{\boldsymbol{\theta}}_z}_{T_1(z)} + \underbrace{\sqrt{nh} (\mathbf{e}_1^T - \hat{\boldsymbol{\theta}}_z^T \hat{\Sigma}_z) (\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}_+)}_{T_2(z)},$$

where η'_i is defined as

$$\eta'_i = f(X_{1i}, \dots, X_{di}) - \sum_{j=1}^d f_{mj}(X_{ji}) \quad \text{for any } i \in [n]. \quad (\text{F.17})$$

Using Theorem 3.2 and Lemma 3.4, with probability $1 - c/n$, we have

$$\begin{aligned} \sup_{z \in \mathcal{X}} |T_2(z)| &\leq \sqrt{nh} \|\hat{\Sigma}_z \boldsymbol{\theta}_z - \mathbf{e}_1\|_{2,\infty} \|\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta}_+\|_{2,1} \\ &\leq C\sqrt{nh} \left(\sqrt{\frac{m \log^2 d}{nh}} \right) \cdot sm \left(\sqrt{\frac{\log(dmh^{-1})}{nh}} + \frac{\sqrt{s}}{m^{5/2}} + \frac{m^{3/2} \log(dh^{-1})}{n} + \frac{h^2}{\sqrt{m}} \right). \quad (\text{F.18}) \end{aligned}$$

Since $mh = o(1)$ and $h \asymp n^{-\delta}$ for $\delta > 1/5$, we have $\sup_{z \in \mathcal{X}} |T_2(z)| = o_P(n^{-1/10})$.

To bound $T_1(z)$, we first apply the triangle inequality and Cauchy-Schwartz inequality to

decompose $T_1(z)$ into three smaller fragments

$$\begin{aligned} T_1(z) &= \sqrt{h/n} \sum_{i=1}^n K_h(X_{i1} - z) \eta_i' \Psi_{i\bullet}^T (\hat{\theta}_z - \theta_z) + \sqrt{h/n} \sum_{i=1}^n K_h(X_{i1} - z) \eta_i' \Psi_{i\bullet}^T \theta_z \\ &\leq \sqrt{nh} \cdot T_{11}^{1/2}(z) \cdot T_{12}^{1/2}(z) + \sqrt{nh} \cdot T_{13}(z), \end{aligned}$$

where the three processes T_{11} , T_{12} and T_{13} are defined as follows

$$\begin{aligned} T_{11}(z) &= \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) (\Psi_{i\bullet}^T (\hat{\theta}_z - \theta_z))^2, \quad T_{12}(z) = \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) (\eta_i')^2 \\ \text{and } T_{13}(z) &= \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \eta_i' \Psi_{i\bullet}^T \theta_z. \end{aligned}$$

From Lemma C.4, we can bound the supreme of $T_{11}(z)$ by

$$\begin{aligned} \sup_{z \in \mathcal{X}} |T_{11}(z)| &= \sup_{z \in \mathcal{X}} \left(\hat{\theta}_z - \theta_z \right)^T \hat{\Sigma}_z \left(\hat{\theta}_z - \theta_z \right) \\ &\leq C \inf_{z \in \mathcal{X}} \frac{\log(D_n/p_1(z))}{p_1^2(z)} \cdot m \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \sqrt{\frac{\log(1/h)}{nh}} \right). \end{aligned} \quad (\text{F.19})$$

Let δ , ξ_z and ζ_z be as defined in Lemma A.2 and Lemma A.3. From those two lemmas, with probability $1 - c/n$, we have

$$\sup_{z \in \mathcal{X}} |T_{12}(z)| \leq \sup_{z \in \mathcal{X}} \frac{2}{n} \|\mathbf{W}_z^{1/2} \delta\|_2^2 + \sup_{z \in \mathcal{X}} \frac{2}{n} \|\mathbf{W}_z^{1/2} (\xi_z + \zeta_z)\|_2^2 \leq C(sm^{-4} + h^2). \quad (\text{F.20})$$

Lemma A.2 and Lemma A.3 also give us

$$\begin{aligned} \sup_{z \in \mathcal{X}} |T_{13}(z)| &\leq \sup_{z \in \mathcal{X}} \frac{1}{n} \|\Psi^T \mathbf{W}_z (\delta + \xi_z + \zeta_z)\|_{2,\infty} \|\theta_z\|_1 \\ &\leq Cm \left(\sqrt{s} \cdot m^{-5/2} + \sqrt{\frac{h \log(dh^{-1})}{n}} + \frac{m^{3/2} \log(dh^{-1})}{n} + \frac{h^2}{\sqrt{m}} \right). \end{aligned} \quad (\text{F.21})$$

Combining (F.19), (F.20) with (F.21), if the scaling condition of Theorem 3.7 is satisfied, there exists a constant c such that $\mathbb{P}(\sup_{z \in \mathcal{X}} |T_1(z)| > Cn^{-c}) < C/n$. Combining this inequality with the rate of $\sup_{z \in \mathcal{X}} |T_1(z)|$ in (F.18), we have our lemma proved.

F.6 Proof of Lemma C.3

We first bound the difference between $\widehat{\boldsymbol{\theta}}_z^T \boldsymbol{\Sigma}'_z \widehat{\boldsymbol{\theta}}_z$ and $\boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z$ by applying triangle inequality and Cauchy-Schwartz inequality

$$\begin{aligned} & \left| \widehat{\boldsymbol{\theta}}_z^T \boldsymbol{\Sigma}'_z \widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z \right| \\ & \leq (\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z)^T \boldsymbol{\Sigma}'_z (\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z) + 2(\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z)^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z \\ & \leq h^{-1}(\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z)^T \widehat{\boldsymbol{\Sigma}}_z (\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z) + 2h^{-1/2} \sqrt{(\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z)^T \widehat{\boldsymbol{\Sigma}}_z (\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z)} \sqrt{\boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z}. \end{aligned} \quad (\text{F.22})$$

From Lemma C.4, we have the desired upper bound in the lemma that

$$\sup_{z \in \mathcal{X}} (\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z)^T \widehat{\boldsymbol{\Sigma}}_z (\widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z) \leq m \left(\sqrt{\frac{m \log(dm)}{nh}} + \frac{m}{nh} + \sqrt{\frac{\log(1/h)}{nh}} \right). \quad (\text{F.23})$$

The following lemma gives us a bound on the term $\boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z$.

Lemma F.5. Under Assumption (A1), for any $z \in \mathcal{X}$,

$$\mathbb{E}[\boldsymbol{\Sigma}'_z] = h^{-1} \lambda(K) [\boldsymbol{\Sigma}_z + o(h)],$$

where $\lambda(K) = \int K^2(u) du$. Furthermore, with probability at least $1 - c/n$,

$$\sup_{z \in \mathcal{X}} \|\boldsymbol{\Sigma}'_z - \mathbb{E}[\boldsymbol{\Sigma}'_z]\|_{\max} \leq C \left(\frac{1}{nh^2} + \frac{1}{\sqrt{nh^3}} + \sqrt{\frac{\log(dm)}{nmh^3}} \right).$$

We defer the proof of the lemma to the end of the section. Using Lemma F.5, we have

$$\begin{aligned} \boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z & \geq \boldsymbol{\theta}_z^T \mathbb{E}[\boldsymbol{\Sigma}'_z] \boldsymbol{\theta}_z - \|\boldsymbol{\Sigma}'_z - \mathbb{E}[\boldsymbol{\Sigma}'_z]\|_{\max} \|\boldsymbol{\theta}_z\|_1^2 \\ & \geq h^{-1} \lambda(K) \boldsymbol{\theta}_z^T \boldsymbol{\Sigma}_z \boldsymbol{\theta}_z - \|\boldsymbol{\Sigma}'_z - \mathbb{E}[\boldsymbol{\Sigma}'_z]\|_{\max} \|\boldsymbol{\theta}_z\|_1^2 - o(1) \\ & \geq h^{-1} \lambda(K) \mathbf{e}_1^T \boldsymbol{\theta}_z - Cm^2 \left(\frac{1}{nh^2} + \frac{1}{\sqrt{nh^3}} + \sqrt{\frac{\log(dm)}{nmh^3}} \right) - o(1). \end{aligned} \quad (\text{F.24})$$

We can also bound from the other direction as

$$\boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z \leq h^{-1} \lambda(K) \mathbf{e}_1^T \boldsymbol{\theta}_z + Cm^2 \left(\frac{1}{nh^2} + \frac{1}{\sqrt{nh^3}} + \sqrt{\frac{\log(dm)}{nmh^3}} \right) + o(1). \quad (\text{F.25})$$

Combining (F.22), (F.23), (F.24) and (F.25), if $mh = o(1)$, $h = n^{-\delta}$ for $\delta > 1/5$, there exists a constant c such that for any $z \in \mathcal{X}$,

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_z^T \boldsymbol{\Sigma}'_z \widehat{\boldsymbol{\theta}}_z &\geq \boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z - \left| \widehat{\boldsymbol{\theta}}_z^T \boldsymbol{\Sigma}'_z \widehat{\boldsymbol{\theta}}_z - \boldsymbol{\theta}_z^T \boldsymbol{\Sigma}'_z \boldsymbol{\theta}_z \right| \\ &= h^{-1} (\lambda(K) \mathbf{e}_1^T \boldsymbol{\theta}_z + o(1)) \geq ch^{-1} \mathbf{e}_1^T \boldsymbol{\theta}_z\end{aligned}$$

Similarly, we also have $\widehat{\boldsymbol{\theta}}_z^T \boldsymbol{\Sigma}'_z \widehat{\boldsymbol{\theta}}_z \leq Ch^{-1} \mathbf{e}_1^T \boldsymbol{\theta}_z$. The proof will be done once we prove Lemma F.5.

Proof of Lemma F.5. For any $j, j' \in [d]$ and $k, k' \in [m]$, we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{\Sigma}'_z]_{jj'kk'} &= \int K_h^2(x - z) (\psi_{jk}(x_j) \psi_{j'k'}(x_{j'})) p_{1,j,j'}(x_1, x_j, x_{j'}) dx_1 dx_j dx_{j'} \\ &= h^{-1} \int K^2(u) (\psi_{jk}(x_j) \psi_{j'k'}(x_{j'})) p_{1,j,j'}(z + uh, x_j, x_{j'}) du dx_j dx_{j'} \\ &= h^{-1} \lambda(K) \int K(u) (\psi_{jk}(x_j) \psi_{j'k'}(x_{j'})) (p_{1,j,j'}(z, x_j, x_{j'}) + o(h)) du dx_j dx_{j'} \\ &= h^{-1} \lambda(K) [\boldsymbol{\Sigma}_z + o(h)]_{jj'kk'}.\end{aligned}$$

The second part of the proof is similar to the proof of Lemma E.1. Consider the random variable

$$Z_{kk'jj'} = \sup_{z \in \mathcal{X}} (\mathbb{E}_n - \mathbb{E}) [K_h^2(X_{i1} - z) \psi_k(X_{ij}) \psi_{k'}(X_{ij'})].$$

Define the following two function classes

$$\begin{aligned}\mathcal{G}_h &= \left\{ g_z(x_1, x_2, x_3) = h^{-2} K^2(h^{-1}(x_1 - z)) \psi_k(x_2) \psi_{k'}(x_3) \mid z \in \mathcal{X}, x_1, x_2, x_3 \in \mathcal{X} \right\} \text{ and} \\ \mathcal{F}_h^2 &= \left\{ h^{-2} K^2(h^{-1}(\cdot - z)) \mid z \in \mathcal{X} \right\}.\end{aligned}$$

Using Lemma H.3, we bound the covering number by

$$\sup_Q N(\mathcal{F}_h, L^2(Q), \epsilon) \leq \left(\frac{8 \|K\|_{\text{TV}}^2 A^2}{h^2 \epsilon} \right)^8,$$

where Q is any measure on \mathbb{R} . Therefore the covering number for \mathcal{G}_h satisfies

$$N(\mathcal{G}_h, L^2(\mathbb{P}), \epsilon) \leq \left(\frac{8 \|K\|_{\text{TV}}^2 A^2}{h^2 \epsilon} \right)^8.$$

The envelope of \mathcal{G}_h is $U = 4h^{-2}\|K\|_\infty$ and we bound the variance of the process by

$$\begin{aligned}\sigma_P^2 &:= \mathbb{E} \left[\left(K_h^2(X_1 - z) (\psi_k(X_j) \psi_{k'}(X_{j'})) \right)^2 \right] \\ &= h^{-3} \mathbb{E} \left[K^2(h^{-1}(X_1 - z)) \mathbb{E} [(\psi_k^2(X_j) \psi_{k'}^2(X_{j'})) | X_1] \right] \leq Cm^{-2}h^{-3}.\end{aligned}$$

Using Lemma H.2, we obtain the upper bound of the expectation

$$\mathbb{E}[Z_{kk'jj'}] \leq C_1 \sqrt{\frac{\log(C_2 m)}{nm^2 h^3}}.$$

As $|Z_{kk'jj'}| \leq 4h^{-2}$ and $\sigma^2 \leq Cm^{-2}h^{-3}$, Lemma H.4 gives us

$$\mathbb{P} \left(Z_{kk'jj'} \geq \mathbb{E}[Z_{kk'jj'}] + t \sqrt{Cm^{-2}h^{-3} + 4h^{-2}\mathbb{E}[Z_{kk'jj'}] + 4t^2h^{-2}/3} \right) \leq \exp(-nt^2).$$

By letting $t = 3\sqrt{\log(dm)/n}$, we obtain

$$\sup_{z \in \mathcal{X}} \max_{j, j' \geq 2} |\Sigma'_z(j, j') - \mathbb{E}\Sigma'_z(j, j')| = O_P \left(\frac{1}{nh^2} + \sqrt{\frac{\log(dm)}{nm^2 h^3}} \right). \quad (\text{F.26})$$

We also define the empirical process

$$\bar{Z}_{kj} = \sup_{z \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n K_h(X_{i1} - z) \psi_k(X_{ij}) - \mathbb{E}[K_h(X_1 - z) \psi_k(X_j)].$$

As above, we can show that the suprema of the empirical process has the convergence rate as

$$\sup_{z \in \mathcal{X}} \max_{j \geq 2} |\Sigma'_z(j, 1) - \mathbb{E}\Sigma'_z(j, 1)| = O_P \left(\frac{1}{nh^2} + \sqrt{\frac{\log(dm)}{nmh^3}} \right). \quad (\text{F.27})$$

Finally, we have the following upper bound

$$\begin{aligned}\sup_{z \in \mathcal{X}} |\Sigma'_z(1, 1) - \mathbb{E}\Sigma'_z(1, 1)| &\leq \sup_{z \in \mathcal{X}} \left| \frac{1}{n} \sum_{i=1}^n K_h^2(X_{i1} - z) - \mathbb{E}[K_h^2(X_1 - z)] \right| \\ &\leq C \left(\frac{1}{\sqrt{nh^3}} + \frac{1}{nh^2} \right).\end{aligned} \quad (\text{F.28})$$

Combining (F.26), (F.27) and (F.28), with probability at least $1 - c/n$, we have

$$\sup_{z \in \mathcal{X}} \|\Sigma'_z - \mathbb{E}[\Sigma'_z]\|_{\max} \leq C \left(\frac{1}{nh^2} + \frac{1}{\sqrt{nh^3}} + \sqrt{\frac{\log(dm)}{nmh^3}} \right),$$

which completes the proof of the Lemma. \square

G Proof of Proposition 3.3: Examples for Assumption (A6)

In this section, we give concrete examples under which Assumption (A6) is satisfied. Given some $\rho \in [0, 1/2]$, denote $\mathbf{M}(\rho, p)$ as a $p \times p$ 3-banded matrix with $[\mathbf{M}(\rho, p)]_{kk} = 1 + \rho$ for all $k \in [d]$, $[\mathbf{M}(\rho, p)]_{st} = \rho$ if $0 < |s - t| \leq 1$ and $[\mathbf{M}(\rho, p)]_{st} = 0$ if $|s - t| > 1$.

For any $j \in \{2, \dots, d\}$, we consider a covariance matrix $\Sigma^{(j)}$ such that $\Sigma^{(2)} = \text{diag}(\mathbf{I}_2, \mathbf{M}(\rho, d - 2)) + (\rho/d)(\mathbf{e}_1 \mathbf{e}_2^T + \mathbf{e}_2 \mathbf{e}_1^T)$ and

$$\Sigma^{(j)} = \text{diag}(\mathbf{I}_2, \mathbf{M}(\rho, j - 3), 1 + \rho, \mathbf{M}(\rho, d - j)) + (\rho/d)(\mathbf{e}_1 \mathbf{e}_j^T + \mathbf{e}_j \mathbf{e}_1^T) \text{ for } j > 2,$$

where \mathbf{e}_j is the j -th canonical basis in \mathbb{R}^d and $\delta_j \in \{0, 1\}$. We assume the covariances are sparse in the sense that $J := \sum_{j=2}^d \delta_j = O(1)$ and $\delta_2 = 1$. Given some $\pi \in (0, 1/2)$, we consider the mixture distribution

$$p(\mathbf{x}) = \frac{1 - \pi}{\pi^2(1 - \pi^{d-1})} \sum_{j=2}^d \frac{\pi^j}{\sqrt{(2\pi)^d |\Sigma^{(j)}|}} \exp\left(-\frac{1}{2} \mathbf{x}^T (\Sigma^{(j)})^{-1} \mathbf{x}\right). \quad (\text{G.1})$$

We can easily check that

$$\sum_{j=2}^d \text{Cov}(X_1, X_j) \geq \pi^2 \rho, \text{Cov}(X_j, X_k) \geq \pi^2 \rho \text{ for all } |j - k| \leq 1.$$

We first show a property for 3-tuple joint density p_{1jk} . For simplicity, we denote $\mathbf{w} = (x_1, x_j, x_k)^T$ and $p_{1,j,k} \propto \exp(-\frac{1}{2} \mathbf{w}^T \mathbf{S}^{-1} \mathbf{w})$. Denote $\mathbf{S}_D = \text{diag}(\mathbf{S})$. We have

$$\delta(\mathbf{S})^2 := \|p_{1jk} - p_1 p_j p_k\|_2^2 = \frac{1}{\sqrt{(2\pi)^3}} \left[\frac{1}{\sqrt{8|\mathbf{S}|}} - \frac{2}{\sqrt{|\mathbf{S}^{-1} - \mathbf{S}_D^{-1}| |\mathbf{S}| |\mathbf{S}_D|}} + \frac{1}{\sqrt{8|\mathbf{S}_D|}} \right], \quad (\text{G.2})$$

where the right hand side of the equality is obtained through integrating normal density functions

and we omit the details.

Given the joint density $p(\mathbf{x})$ in (G.1), we have

$$\|p_{1jk} - p_1 p_j p_k\|_2 = \frac{1 - \pi}{\pi^2(1 - \pi^{d-1})} \sum_{s=2}^d \pi^s \delta(\mathbf{S}_{(s,j,k)}), \text{ where } \mathbf{S}_{(s,j,k)} = \begin{pmatrix} \mathbf{S}_{11}^{(s)} & \mathbf{S}_{1j}^{(s)} & \mathbf{S}_{1k}^{(s)} \\ \mathbf{S}_{j1}^{(s)} & \mathbf{S}_{jj}^{(s)} & \mathbf{S}_{jk}^{(s)} \\ \mathbf{S}_{k1}^{(s)} & \mathbf{S}_{kj}^{(s)} & \mathbf{S}_{kk}^{(s)} \end{pmatrix}.$$

Suppose $j < k$, and have three cases:

$$\mathbf{S}_{(s,j,k)} = \begin{cases} (1 + \rho)\mathbf{I}_3 + (\rho/d) \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, & \text{if } s = j; \\ (1 + \rho)\mathbf{I}_3 + (\rho/d) \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, & \text{if } s = k; \\ (1 + \rho)\mathbf{I}_3 + \rho \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, & \text{if } s \neq j, k, \text{ and } j > 2, k = j + 1; \\ (1 + \rho)\mathbf{I}_3, & \text{otherwise.} \end{cases} \quad (\text{G.3})$$

For case (i) in (G.3), applying (G.2) and we have when $\rho \in [0, 1/2]$,

$$\|p_{1jk} - p_1 p_j p_k\|_2^2 = \frac{1}{\sqrt{8(2\pi)^2}} \left[\frac{1}{\sqrt{(1 + \rho)^2 - (\rho/d)^2}} - \frac{4}{\sqrt{4(1 + \rho)^2 - (\rho/d)^2}} + \frac{1}{1 + \rho} \right] \leq (\rho/d)^2.$$

Case (ii) in (G.3) is the same. For case (iii) in (G.3), they are just case (i) with $d = 1$, i.e.,

$$\|p_{1jk} - p_1 p_j p_k\|_2^2 \leq \rho^2. \text{ For case (iv), } \|p_{1jk} - p_1 p_j p_k\|_2^2 = 0.$$

Summarizing the results above, since $\rho, \pi \in (0, 1/2)$, we have for any $k > 0$,

$$\begin{aligned} \sum_{j \neq k} \|p_{1jk} - p_1 p_j p_k\|_2 &= \frac{1 - \pi}{\pi^2(1 - \pi^{d-1})} \sum_{j \neq k} \sum_{s=2}^d \pi^s \delta(\mathbf{S}_{(s,j,k)}) \\ &\leq \frac{1 - \pi}{\pi^2(1 - \pi^{d-1})} \left[\sum_{s=2}^d \pi^s \rho + 2\rho\pi^2 \right] \leq 5\rho \leq \rho_{\min}/(2B), \end{aligned}$$

if $\rho/\pi \leq \rho_{\min}/(6B)$.

For the bivariate density $\|p_{jk} - p_j p_k\|_2$, we can similarly get

$$\sum_{j \geq 2} \|p_{1j} - p_1 p_j\|_2 \leq \frac{1 - \pi}{\pi^2(1 - \pi^{d-1})} \rho \pi^2 \leq 2\rho \leq \rho_{\min}/(2B).$$

Therefore, if we choose $\pi = 1/3$, then if $\rho \leq \rho_{\min}/(18B)$, our example above satisfies

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

and Assumption **(A6)** holds.

H Results on Empirical Processes

Lemma H.1 (Lemma H.2, [Lu et al. \(2015\)](#)). Let \mathcal{F}_1 and \mathcal{F}_2 be two function classes satisfying

$$N(\mathcal{F}_1, \|\cdot\|_{L_2(Q)}, a_1\epsilon) \leq C_1\epsilon^{-v_1} \quad \text{and} \quad N(\mathcal{F}_2, \|\cdot\|_{L_2(Q)}, a_2\epsilon) \leq C_2\epsilon^{-v_2}$$

for some $C_1, C_2, a_1, a_2, v_1, v_2 > 0$ and any $0 < \epsilon < 1$. Define $\|\mathcal{F}_\ell\|_\infty = \sup\{\|f\|_\infty, f \in \mathcal{F}_\ell\}$ for $\ell = 1, 2$ and $U = \|\mathcal{F}_1\|_\infty \vee \|\mathcal{F}_2\|_\infty$. For the function classes $\mathcal{F}_\times = \{f_1 f_2 \mid f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ and $\mathcal{F}_+ = \{f_1 + f_2 \mid f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$, we have for any $\epsilon \in (0, 1)$,

$$\begin{aligned} N(\mathcal{F}_\times, \|\cdot\|_{L_2(Q)}, \epsilon) &\leq C_1 C_2 \left(\frac{2a_1 U}{\epsilon} \right)^{v_1} \left(\frac{2a_2 U}{\epsilon} \right)^{v_2}; \\ N(\mathcal{F}_+, \|\cdot\|_{L_2(Q)}, \epsilon) &\leq C_1 C_2 \left(\frac{2a_1}{\epsilon} \right)^{v_1} \left(\frac{2a_2}{\epsilon} \right)^{v_2}. \end{aligned}$$

Lemma H.2 (Corollary 5.1, [Chernozhukov et al. \(2014b\)](#)). Assume that the functions in \mathcal{F} defined on \mathcal{X} are uniformly bounded by an envelope function $F(\cdot)$ such that $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$. Define $\sigma_P^2 = \sup_{f \in \mathcal{F}} \mathbb{E}[f^2]$. Let Q be any measure over \mathcal{X} . If for some $A \geq e, V \geq 0$ and for all $\varepsilon > 0$, the covering entropy satisfies

$$\sup_Q N(\mathcal{F}, L^2(Q); \epsilon) \leq \left(\frac{A\|F\|_{L^2(Q)}}{\varepsilon} \right)^V,$$

then for any i.i.d. subgaussian mean zero random variables $\varepsilon_1, \dots, \varepsilon_n$ there exists a universal constant C such that

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(X_{i1}) - \mathbb{E}f(X)) \right] \leq C \left[\sqrt{\frac{V}{n}} \sigma_P \sqrt{\log \frac{A\|F\|_{L^2(\mathbb{P})}}{\sigma_P}} + \frac{V\|F\|_{L^2(\mathbb{P})}}{\sqrt{n}} \log \frac{A\|F\|_{L^2(\mathbb{P})}}{\sigma_P} \right].$$

Lemma H.3 (Lemma 3, [Giné and Nickl \(2009\)](#)). Let $K : \mathbb{R} \mapsto \mathbb{R}$ be a bounded variation function. Define the function class $\mathcal{F}_h = \{K((t - \cdot)/h) \mid t \in \mathbb{R}\}$. There exists $A < \infty$ such that for all probability measures Q on \mathbb{R} , we have

$$\sup_Q N(\mathcal{F}_h, L^2(Q), \epsilon) \leq \left(\frac{2\|K\|_{\text{TV}}A}{\epsilon} \right)^4, \text{ for any } \epsilon \in (0, 1).$$

Lemma H.4 ([Bousquet \(2002\)](#)). Let X_1, \dots, X_n be independent random variables and \mathcal{F} is a function class such that there exist η_n and τ_n^2 satisfying

$$\sup_{f \in \mathcal{F}} \|f\|_\infty \leq \eta_n \quad \text{and} \quad \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \text{Var}(f(X_{i1})) \leq \tau_n^2.$$

Define the random variable Z being the suprema of an empirical process

$$Z = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n (f(X_{i1}) - \mathbb{E}f(X_{i1})) \right|. \quad (\text{H.1})$$

Then for any $z > 0$, we have the following concentration inequality on the suprema

$$\mathbb{P} \left(Z \geq \mathbb{E}Z + z\sqrt{2(\tau_n^2 + 2\eta_n\mathbb{E}Z)} + 2z^2\eta_n/3 \right) \leq \exp(-nz^2).$$

The following lemma gives the deviation inequality when \mathcal{F} is not universally bounded.

Lemma H.5 (Theorem 5.1, [Chernozhukov et al. \(2014b\)](#)). Let $F(\cdot)$ be the envelope function of \mathcal{F} such that $F \in L^2(\mathbb{P})$ and Z is defined in [\(H.1\)](#), For every $t \geq 1$, there exists a universal constant C such that

$$\mathbb{P} \left(Z \geq 2\mathbb{E}Z + C(\sigma_P + \|F\|_{L^2(\mathbb{P})})z + \|F\|_{L^2(\mathbb{P})}z^2 \right) \leq 1/z^2.$$

References

- BOUSQUET, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique* **334** 495–500.
- 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.
- DE BOOR, C. (2001). *A practical guide to splines*, vol. 27 of *Applied Mathematical Sciences*. Revised ed. Springer-Verlag, New York.
- DEMMELE, J. (1992). The componentwise distance to the nearest singular matrix. *SIAM Journal on Matrix Analysis and Applications* **13** 10–19.
- FARRELL, M. H. (2013). Robust inference on average treatment effects with possibly more covariates than observations. *arXiv preprint arXiv:1309.4686* .
- FRIEDMAN, J. H., HASTIE, T. J. and TIBSHIRANI, R. J. (2010). A note on the group lasso and a sparse group lasso. *ArXiv e-prints, arXiv:1001.0736* .
- GINÉ, E. and GUILLOU, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincaré (B) Probability and Statistics*, vol. 38. Elsevier.
- GINÉ, E. and NICKL, R. (2009). An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Probab. Theory Related Fields* **143** 569–596.
- HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.
- HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Stat.* **38** 2282–2313.
- LU, J., KOLAR, M. and LIU, H. (2015). Post-regularization inference for dynamic nonparanormal graphical models. *arXiv preprint arXiv:1512.08298* .
- QIN, Z., SCHEINBERG, K. and GOLDFARB, D. (2013). Efficient block-coordinate descent algorithms for the group lasso. *Mathematical Programming Computation* **5** 143–169.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
- YUAN, M. and LIN, Y. (2007). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* **68** 49–67.
- ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Stat.* **26** 1760–1782.