Journal of Machine Learning Research 20 (2019) 1-38 Submitted 9/17; Revised 8/19; Published 8/19

Layer-Wise Learning Strategy for Nonparametric Tensor Product Smoothing Spline Regression and Graphical Models

Kean Ming Tan KEANMING@UMICH.EDU
Department of Statistics
University of Michigan
Ann Arbor MI, 48109

Junwei Lu JUNWEILU@HSPH.HARVARD.EDU
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Boston MA, 02115

Tong Zhang TONGZHANG@TONGZHANG-ML.ORG
Department of Computer Science and Engineering
Department of Mathematics
The Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong.

Han Liu HANLIU@NORTHWESTERN.EDU
Department of Electrical Engineering and Computer Science
Northwestern University
Evanston IL, 60208

Editor: Sara van de Geer

Abstract

Nonparametric estimation of multivariate functions is an important problem in statistical machine learning with many applications, ranging from nonparametric regression to nonparametric graphical models. Several authors have proposed to estimate multivariate functions under the smoothing spline analysis of variance (SSANOVA) framework, which assumes that the multivariate function can be decomposed into the summation of main effects, two-way interaction effects, and higher order interaction effects. However, existing methods are not scalable to the dimension of the random variables and the order of interactions. We propose a LAyer-wiSE leaRning strategy (LASER) to estimate multivariate functions under the SSANOVA framework. The main idea is to approximate the multivariate function sequentially starting from a model with only the main effects. Conditioned on the support of the estimated main effects, we estimate the two-way interaction effects only when the corresponding main effects are estimated to be non-zero. This process is continued until no more higher order interaction effects are identified. The proposed strategy provides a data-driven approach for estimating multivariate functions under the SSANOVA framework. Our proposal yields a sequence of estimators. To study the theoretical properties of the sequence of estimators, we establish the notion of post-selection persistency. Extensive numerical studies are performed to evaluate the performance of LASER.

Keywords: Persistency, nonparametric regression, nonparametric graphical models, sequential algorithm, model selection

1. Introduction

Much progress has been made in nonparametric estimation of univariate functions. However, nonparametric estimation of multivariate functions remains a challenging problem due to the curse of dimensionality. A number of algorithms were proposed to estimate low-dimensional multivariate functions, but there are few practical algorithms for estimating multivariate functions with higher dimension.

To address this issue, many authors proposed to restrict the multivariate function to some specific model classes. One popular model class is the additive model in which the high-dimensional multivariate function $f(\cdot)$ is decomposed into the sum of d one-dimensional functions (Stone, 1985; Hastie and Tibshirani, 1990). To increase the flexibility of the additive model to accommodate situation in which interactions among the variables may be present, Lin (2000) proposed to estimate the multivariate function under the smoothing spline analysis of variance (SSANOVA) framework. More specifically, Lin (2000) proposed to decompose the d-dimensional function $f(\cdot)$ as the summation of some constant μ, one-dimensional functions (main effects), two-dimensional functions (two-way interaction effects), and so on:

$$f(\boldsymbol{x}) = \mu + \sum_{j=1}^{d} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k) + \cdots . \tag{1}$$

We refer the reader to Lin (2000) and Gu (2013) for a detailed discussion of such models.

Most existing methods under the SSANOVA framework truncate (1) to the rth order interaction effects. Even so, existing methods are computationally intensive and are not scalable to the dimension of the random variables d and the order of the interaction term r. For instance, to fit a nonparametric regression model under the SSANOVA framework with the rth order interaction, it involves simultaneously fitting $O(d^r)$ terms and is often infeasible when both d and r are large (Lin and Zhang, 2006). Thus, one way to reduce the computation complexity is to consider only the two-way interaction terms and remove all of the higher order interaction terms from the model (Lin, 2000; Zhang et al., 2004; Lin and Zhang, 2006; Jeon and Lin, 2006; Yau et al., 2012).

The strong heredity assumption is often used for modeling regression with two-way interaction effects (Bien et al., 2013; Hao and Zhang, 2014; Haris et al., 2016; Hao et al., 2018; Radchenko and James, 2010). The strong heredity assumption mandates that both of the corresponding main effects must be present when an interaction term is included in the model. Under the strong heredity assumption, Hao and Zhang (2014) and Hao et al. (2018) proposed approaches that perform variable selection on the main effects, and allow interactions into the model once the main effects have been identified, in the context of linear regression. They proved theoretically that the main effects and interactions can be selected consistently as long as the variables are Gaussian with mean zero. On the other hand, Bien et al. (2013) and Haris et al. (2016) proposed penalty functions that are specifically designed for two-way interaction models with sparsity and strong heredity assumption in the context of linear regression. Similar work has also been done in the context of nonparametric regression (Radchenko and James, 2010). However, Bien et al. (2013), Haris et al. (2016), and Radchenko and James (2010) are computationally intractable when d is large since it

involves modeling $\binom{d}{2}$ terms. We refer the reader to Haris et al. (2016) for a comprehensive review of the literature.

In this paper, we propose a framework to estimate multivariate functions that take the form (1) without restricting it to only modeling the two-way interaction terms. We impose a hierarchical structure on the higher order interaction terms under the SSANOVA framework. Let $J, J' \subseteq \{1, \ldots, d\}$. Given a d-dimensional function f, we denote f_J as the $|J|$th order interaction term in f with variables $\{x_j\}_{j \in J}$. We impose the hierarchical structural assumption that

$$f_J = 0 \implies f_{J'} = 0, \qquad \text{if } J \subset J'. \tag{2}$$

In other words, we assume that a higher order interaction is not active when some of the lower order terms containing some variables that belong to the higher order term are not active. For instance, if the main effect of the jth variable $f_j = 0$, then any higher order term that involves the jth variable is not active. The hierarchical structural assumption is plausible in many data applications. For instance, in the context of nonparametric regression, the hierarchical structural assumption implies that the main effects are more important in modeling the response than the higher order interactions. Therefore, if the leading order interaction function does not have any explanatory power, the higher order interaction terms should be inactive. The hierarchical assumption in (2) can be thought of as an extension of the strong heredity assumption to modeling higher order interaction terms.

Under the hierarchical structural assumption, we propose a layer-wise learning algorithm to estimate multivariate functions under the SSANOVA framework. Our algorithm sequentially estimates the multivariate function starting from the main effects to the higher order interaction terms. In each step of our algorithm, we utilize the support of the estimated function from the previous step and estimate the function based on the hierarchical structural assumption in (2). This process is continued until no more higher order interaction effects are active. Instead of fitting the SSANOVA model with $d + \binom{d}{2} + \cdots + \binom{d}{r}$ terms, our proposal fits the SSANOVA model with at most $d + \binom{s_1}{2} + \cdots + \binom{s_{r-1}}{r}$ terms, where s_k is the cardinality of the support of the estimated kth order interaction effects. Thus, our algorithm is scalable to modeling higher order interaction terms with large dimension d. Our proposed framework can be interpreted as an extension of Hao and Zhang (2014) and Hao et al. (2018) to modeling multivariate functions, as well as modeling higher order interaction effects. Compared to Radchenko and James (2010) that involves modeling all two-way interaction terms, our approach is a multi-stage procedure and therefore is computationally efficient and scalable for problems with large dimension d.

To quantify the theoretical properties of our estimator, we propose the notion of post-selection persistency. We show that conditioned on the support of the estimator obtained from the $(r-1)$th step of the algorithm, the excessive risk between the estimator obtained from the rth step and the best rth order SSANOVA model converges to zero. Our results hold without assuming that the true multivariate function takes the form of the SSANOVA model in (1). In addition, we impose minimal distributional assumptions on the data.

We apply the proposed method to fitting nonparametric regression and graphical models. For nonparametric graphical models, our proposal is the first feasible approach to learn the graph without restricting the model to contain only pairwise interaction terms. In Section 2, we describe the problem setup and define the tensor product space for the func-

tional component in the SSANOVA decomposition in (1). We then propose the layer-wise learning algorithm for estimating multivariate function under a generic loss function. We apply the proposed algorithm to fitting nonparametric regression and graphical models in Sections 3 and 4, respectively. Numerical studies are performed in Section 5. We close with a discussion in Section 6. The proofs of the theoretical results are given in the Appendix.

2. Layer-Wise Learning Strategy

We describe the problem setup and define some notation. We then propose the layer-wise learning strategy for estimating nonparametric functions with high order interactions.

2.1 Problem Setup and Notation

We start with a brief overview of the tensor product space of Sobolev spaces and refer the reader to Lin (2000) for a detailed review. For any non-negative integer m, the mth order Sobolev space with a univariate variable $x_j \in [0,1]$ is defined as

$$H_j^m = \left\{ g \;\middle|\; g^{(\nu)} \text{ is absolutely continuous for } 0 \le \nu \le m-1;\ g^{(m)} \in L^2([0,1]);\ \int_0^1 g(u)du = 0 \right\},$$

where $g^{(\nu)}$ is the νth order derivative of g. The Sobolev norm for function $g \in H_j^m$ is defined as

$$\|g\|_{H_j^m}^2 = \sum_{\nu=0}^{m} \int_0^1 \left[g^{(\nu)}(u) \right]^2 du.$$

For notational convenience, let $[d] = \{1, \ldots, d\}$. Let $J \subseteq [d]$ be an index set with cardinality $|J| = r$, and let $\mathcal{H}_J = \otimes_{j \in J} H_j^m$ be the completed tensor product space of H_j^m for all $j \in J$. We assume that $g_J \in \mathcal{H}_J$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)^T$ be an r-dimensional vector with integer entries and let $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^r \alpha_i \le m$. Let $J = \{j_1, \ldots, j_r\}$. The Sobolev norm for the multivariate function $g_J \in \mathcal{H}_J$ is defined as

$$\|g_J\|_{\mathcal{H}_J}^2 = \sum_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_1 \le m} \|D^{\boldsymbol{\alpha}} g_J\|_2 \qquad \text{where} \qquad D^{\boldsymbol{\alpha}} = \frac{\partial^{\|\boldsymbol{\alpha}\|_1}}{\partial x_{j_1}^{\alpha_1} \cdots \partial x_{j_r}^{\alpha_r}}. \tag{3}$$

We define the smoothing spline ANOVA function class as

$$\{1\} \oplus \sum_{j=1}^{d} H_j^m \oplus \sum_{J \subseteq [d], |J|=2} \mathcal{H}_J \oplus \sum_{J \subseteq [d], |J|=3} \mathcal{H}_J \oplus \cdots.$$

Each functional component in the SSANOVA decomposition (1) lies in a subspace in the orthogonal decomposition of the tensor product space. We define the rth order smoothing spline ANOVA function class as

$$\mathcal{H}^{(r)} = \{1\} \oplus \sum_{j=1}^{d} H_j^m \oplus \sum_{J \subseteq [d], |J|=2} \mathcal{H}_J \oplus \cdots \oplus \sum_{J \subseteq [d], |J|=r} \mathcal{H}_J. \tag{4}$$

The additive model introduced in Stone (1985) is a special case of (4) with $r = 1$.

2.2 Layer-Wise Learning Algorithm

We propose the layer-wise learning algorithm to learn a d-dimensional multivariate function under the SSANOVA framework. Recall from (2) that we assume a hierarchical structure on the model, i.e., higher order interaction terms are not active when the lower order terms are not active. To this end, we define some additional notation that will be used throughout the paper. Given a set $\mathcal{S} \subseteq \{J \mid J \subseteq [d]\}$, let $|\mathcal{S}|_{\max} = \max\{|I| \mid I \in \mathcal{S}\}$ be the largest cardinality among the sets in $\mathcal{S}$. In addition, we define

$$\sigma(\mathcal{S}) = \left\{ I \;\middle|\; |I| = |\mathcal{S}|_{\max} + 1, I \subseteq \bigcup_{|J|=|\mathcal{S}|_{\max},\, J \in \mathcal{S}} J \right\}.$$

For example, if $\mathcal{S} = \{\{1\},\{2\},\{3\}\}$, then $\sigma(\mathcal{S}) = \{\{1,2\},\{1,3\},\{2,3\}\}$. For notational convenience, we also define $\delta(\mathcal{S}) = \mathcal{S} \cup \sigma(\mathcal{S})$ throughout the paper. Thus, in this example, $\delta(\mathcal{S}) = \{\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\}\}$.

The main crux of our proposal is to estimate the multivariate function sequentially starting from the main effects. Using the support of the estimated function from the previous step, we estimate the multivariate function by considering higher order terms only when the lower order terms are estimated to be non-zeros. Let D be the data and let $\mathcal{L}_n(D, f)$ be a generic loss function for estimating f. At the first step of the algorithm, we estimate the function f subject to $f \in \mathcal{H}^{(1)}$. In other words, f is assumed to be additive and we propose to estimate f by

$$\widehat{f}^{(1)} = \underset{f \in \mathcal{H}^{(1)}}{\operatorname{argmin}}\, \mathcal{L}_n(D, f), \qquad \text{subject to} \qquad \sum_{j=1}^{d} \|P_j(f)\|_2 \leq \tau, \tag{5}$$

where $P_j(f)$ is the orthogonal projection of f onto H_j^m. The penalty term $\sum_{j=1}^{d} \|P_j(f)\|_2 \leq \tau$ can be interpreted as an ℓ_1 constraint across components to encourage sparsity, and an ℓ_2 constraint within components to encourage smoothness. The tuning parameter τ controls the number of main effects that are estimated to be non-zero.

Let $\mathcal{S}^{(1)}$ be the support of $\widehat{f}^{(1)}$, that is, $\mathcal{S}^{(1)} = \{j \mid P_j(\widehat{f}^{(1)}) \neq 0\}$. Given the support $\mathcal{S}^{(1)}$, we fit the following model at the second step of our algorithm

$$\begin{aligned} \widehat{f}^{(2)} = & \underset{f \in \mathcal{H}^{(2)}, \mathcal{S}(f) = \delta(\mathcal{S}^{(1)})}{\operatorname{argmin}} \mathcal{L}_n(D, f), \\ & \text{subject to} \sum_{j \in \mathcal{S}^{(1)}} \|P_j(f)\|_2^2 + \sum_{\{j_1, j_2\} \in \sigma(\mathcal{S}^{(1)})} \|P_{j_1 j_2}(f)\|_2 \leq \tau, \end{aligned} \tag{6}$$

where $P_{j_1 j_2}(f)$ is the orthogonal projection of f onto $H_{j_1}^m \otimes H_{j_2}^m$. For notational convenience, let $\mathcal{S}(f)$ be the support of f. At the second step of our algorithm, we update both the main effects and the second order interactions, with the support of the function constrained on $\mathcal{S}(f) = \delta(\mathcal{S}^{(1)})$. Since we have selected the support for the main effects, we use a ridge penalty to encourage smoothness for the main effects.

More generally, let $\mathcal{S}^{(r-1)}$ be the support identified at the $(r-1)$th step of our proposed algorithm. Let $J \subseteq [d]$ be an index set and let $P_J(f)$ be the orthogonal projection of f onto

Algorithm 1 Layer-Wise Learning Method (LASER).

Input: : Data D.
Initialize: $\mathcal{S}^{(0)} = \emptyset$, $\sigma(\mathcal{S}^{(0)}) = \{1, \ldots, d\}$, and $r = 1$.
repeat

1. Update the function with rth order interaction effects:

$$\widehat{f}^{(r)} = \underset{f \in \mathcal{H}^{(r)}, \mathcal{S}(f) = \delta(\mathcal{S}^{(r-1)})}{\operatorname{argmin}} \mathcal{L}_n(D, f), \quad \text{subject to} \sum_{J \in \mathcal{S}^{(r-1)}} \|P_J(f)\|_2^2 + \sum_{J \in \sigma(\mathcal{S}^{(r-1)})} \|P_J(f)\|_2 \leq \tau.$$

2. Update the support $\mathcal{S}^{(r)} = \mathcal{S}(\widehat{f}^{(r)})$.
3. $r \leftarrow r + 1$.

until $\mathcal{S}^{(r-1)} = \mathcal{S}^{(r)}$.
Output: A sequence of estimators $\{\widehat{f}^{(\ell)}\}_{\ell=1}^{r}$.

the $|J|$th order interaction effect space $\otimes_{j \in J} H_j^m$. At the rth step of the algorithm, we fit the model

$$\widehat{f}^{(r)} = \underset{f \in \mathcal{H}^{(r)}, \mathcal{S}(f) = \delta(\mathcal{S}^{(r-1)})}{\operatorname{argmin}} \mathcal{L}_n(D, f), \quad \text{subject to} \sum_{J \in \mathcal{S}^{(r-1)}} \|P_J(f)\|_2^2 + \sum_{J \in \sigma(\mathcal{S}^{(r-1)})} \|P_J(f)\|_2 \leq \tau. \tag{7}$$

We continue this process until no more higher order interaction effects are estimated to be non-zero. We summarize the proposed method in Algorithm 1. Step 1 in Algorithm 1 depends on a specific loss function $\mathcal{L}_n(D, f)$. We will present the details for Step 1 in the context of nonparametric regression and nonparametric graphical models in Sections 3 and 4, respectively.

2.3 Post-Selection Persistency

We first provide a brief review of the definition of persistency introduced by Greenshtein and Ritov (2004). We define the risk of some function f as $R(f) = \mathbb{E}[\mathcal{L}_n(D, f)]$. An estimator $\widehat{f}$ is said to be persistent relative to a class of function $\mathcal{F}$ if

$$R(\widehat{f}) - \inf_{f \in \mathcal{F}} R(f) = o_P(1). \tag{8}$$

In other words, the risk of the estimator $\widehat{f}$ is consistent to that of the oracle function under the model class $\mathcal{F}$. In the statistical literature, many authors have shown that the estimators for various statistical models are persistent (see, for instance, Greenshtein and Ritov, 2004 for the lasso regression, and Ravikumar et al., 2009 for the sparse additive model). However, most of the existing results on persistency are derived for a single estimator, and not much work has been done to characterize a sequence of estimators.

We establish the notion of post-selection persistency to characterize the theoretical properties of a sequence of estimators. Recall from Algorithm 1 that our proposed method

yields a sequence of estimators $\{\widehat{f}^{(\ell)}\}_{\ell=1}^{r}$. Also recall that we denote $\mathcal{S}^{(r-1)}$ to be the support of $\widehat{f}^{(r-1)}$. Let $\mathcal{F}^{(r)}$ be some function class with support constrained on $\delta(\mathcal{S}^{(r-1)}) = \mathcal{S}^{(r-1)} \cup \sigma(\mathcal{S}^{(r-1)})$. Conditioned on the support $\mathcal{S}^{(r-1)}$, we say that $\widehat{f}^{(r)}$ is *post-selection persistent* if

$$R(\widehat{f}^{(r)}) - \inf_{f \in \mathcal{F}^{(r)}} R(f) = o_P(1). \tag{9}$$

We will show that our proposed estimators are post-selection persistent in the context of nonparametric regression and nonparametric graphical models in Sections 3 and 4, respectively.

3. Nonparametric Regression

We apply the layer-wise learning strategy to the setting of nonparametric regression. We consider a nonparametric regression problem of a univariate response $Y \in \mathbb{R}$ on a d-dimensional covariates $\boldsymbol{X} \in [0,1]^d$:

$$Y = f(\boldsymbol{x}) + \epsilon,$$

where ϵ is the random noise variable. It is generally agreed upon in the literature that estimating a general multivariate function without restricting the function into a smaller function class $\mathcal{F}$ is infeasible.

Hastie and Tibshirani (1990) and Stone (1985) introduced a class of additive models of the form $f(\boldsymbol{x}) = \sum_{j=1}^{d} f_j(x_j)$, which decomposed the multivariate function $f(\cdot)$ into the summation of d univariate functions. One caveat of the additive model is the assumption that there are no interaction terms among the covariates. To address this issue, Lin (2000) proposed to estimate $f(\cdot)$ by assuming that it takes the form in (1) in the nonparametric regression setting. However, their proposal is infeasible for high-dimensional problem in which the number of covariates d and the order of interaction terms are large. Thus, they truncated (1) to only modeling two-way interaction terms. More specifically, they considered the following decomposition for $f(\cdot)$:

$$f(\boldsymbol{x}) = \mu + \sum_{j=1}^{d} f_j(x_j) + \sum_{j<k} f_{jk}(x_j, x_k).$$

Several authors have extended the aforementioned models to perform variable selection and estimation simultaneously (among others, Lin and Zhang, 2006; Ravikumar et al., 2009). However, Ravikumar et al. (2009) models only the main effects and Lin and Zhang (2006) is computationally infeasible for large d problems even when they model only the second order interaction terms. The nonparametric regression literature is vast and we refer the reader to several recent proposals for more references (see, Tibshirani, 2014; Fan et al., 2015; Lou et al., 2016).

We now apply the proposed layer-wise learning strategy for fitting a nonparametric regression. Our proposal is scalable to the dimension of the covariates and does not need to truncate model (1) to only modeling the two-way interaction terms. We show that the resulting sequence of estimators from our algorithm is post-selection persistent under the squared error risk in the high-dimensional setting in which $d > n$.

3.1 Method and Optimization Problem

Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be n independent pairs of observations. We assume that the covariates $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are standardized such that $\mathbf{x}_i \in [0,1]^d$ and that $y_i = f(\mathbf{x}_i) + \epsilon_i$ with $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] < \infty$. The function $f(\cdot)$ is an arbitrary function and is not assumed to take the form of (1). To approximate the function $f(\cdot)$, we fit the model in (7) with the squared error loss function $\mathcal{L}_n(D, f) = \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2/n$. This yields the optimization problem

$$\begin{aligned} &\underset{f \in \mathcal{H}^{(r)}, \mathcal{S}(f) = \delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2, \\ &\text{subject to} \quad \sum_{J \in \mathcal{S}^{(r-1)}} \|P_J(f)\|_2^2 + \sum_{J \in \sigma(\mathcal{S}^{(r-1)})} \|P_J(f)\|_2 \le \tau, \end{aligned} \tag{10}$$

where $\tau > 0$ is a positive tuning parameter.

It is useful to write the function $f(\cdot)$ in terms of its basis function. Let $\{\phi_{j\ell}, \ell = 1, 2, \ldots\}$ denote a uniformly bounded basis with respect to H_j^m. Given $J = \{j_1, \ldots, j_r\}$, for any $f_J \in \mathcal{H}_J$, the basis expansion of f_J is

$$f_J = \sum_{1 \le k_1, \ldots, k_r < \infty} \theta_{j_1 \cdots j_r}^{k_1 \cdots k_r} \phi_{j_1 k_1}(x_{j_1}) \cdots \phi_{j_r k_r}(x_{j_r}). \tag{11}$$

In practice, we approximate (11) by its kth order basis expansion

$$\widetilde{f}_J = \sum_{1 \le k_1, \ldots, k_r \le k} \theta_{j_1 \cdots j_r}^{k_1 \cdots k_r} \phi_{j_1 k_1}(x_{j_1}) \cdots \phi_{j_r k_r}(x_{j_r}) = \boldsymbol{\phi}_J^T(\boldsymbol{x}) \boldsymbol{\theta}_J, \tag{12}$$

where $\boldsymbol{\theta}_J = \text{vec}(\{\theta_{j_1 \cdots j_r}^{k_1 \cdots k_r}\})$ and $\boldsymbol{\phi}_J(\boldsymbol{x}) = \text{vec}(\{\phi_{j_1 k_1}(x_{j_1}) \cdots \phi_{j_r k_r}(x_{j_r})\})$.

Let $\boldsymbol{\Phi}_J$ denote the $n \times k^{|J|}$ matrix with rows $\boldsymbol{\phi}_J(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_J(\mathbf{x}_n)$ and let $\mathbf{y} = (y_1, \ldots, y_n)^T$. We approximate (10) in terms of the kth order basis expansion:

$$\begin{aligned} &\underset{\boldsymbol{\theta}_J, \mathcal{S}(f) = \delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \ \frac{1}{n} \Big\| \mathbf{y} - \sum_{J \in \mathcal{S}(f)} \boldsymbol{\Phi}_J \boldsymbol{\theta}_J \Big\|_2^2, \\ &\text{subject to} \quad \frac{1}{n} \sum_{J \in \mathcal{S}^{(r-1)}} \|\boldsymbol{\Phi}_J \boldsymbol{\theta}_J\|_2^2 + \frac{1}{\sqrt{n}} \sum_{J \in \sigma(\mathcal{S}^{(r-1)})} \|\boldsymbol{\Phi}_J \boldsymbol{\theta}_J\|_2 \le \tau. \end{aligned} \tag{13}$$

Instead of solving optimization problem in (13) directly, we consider solving the following problem

$$\underset{\boldsymbol{\theta}_J, \mathcal{S}(f) = \delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \ \frac{1}{n} \Big\| \mathbf{y} - \sum_{J \in \mathcal{S}(f)} \boldsymbol{\Phi}_J \boldsymbol{\theta}_J \Big\|_2^2 + \lambda \Big(\frac{1}{n} \sum_{J \in \mathcal{S}^{(r-1)}} \|\boldsymbol{\Phi}_J \boldsymbol{\theta}_J\|_2^2 + \frac{1}{\sqrt{n}} \sum_{J \in \sigma(\mathcal{S}^{(r-1)})} \|\boldsymbol{\Phi}_J \boldsymbol{\theta}_J\|_2 \Big) \tag{14}$$

Problems (13) and (14) are equivalent in the sense that for a given tuning parameter $\lambda > 0$, there exists a $\tau > 0$ such that the two problems share the same solution, and vice versa.

Algorithm 2 Block Coordinate Descent Algorithm for Solving (14).

Initialize $\widehat{\boldsymbol{\theta}}_J^{(0)}$.
repeat
  for $J \in \delta(\mathcal{S}^{(r-1)})$ **do**
    Update the coefficients:

$$\widehat{\boldsymbol{\theta}}_J^{(t)} = \left(\boldsymbol{\Phi}_J^T \boldsymbol{\Phi}_J\right)^{-1} \left(\boldsymbol{\Phi}_J^T \mathbf{y} - \boldsymbol{\Phi}_J^T \sum_{J' \in \{\delta(\mathcal{S}^{(r-1)}) \setminus J\}} \boldsymbol{\Phi}_{J'}^T \widehat{\boldsymbol{\theta}}_{J'}^{(t-1)} \right).$$

    Penalize the coefficients:

$$\widehat{\boldsymbol{\theta}}_J^{(t)} = \begin{cases} \widehat{\boldsymbol{\theta}}_J^{(t)}/(1+\lambda) & \text{if } J \in \mathcal{S}^{(r-1)}, \\ \left(1 - \frac{\sqrt{n}\lambda}{2\|\boldsymbol{\Phi}_J \widehat{\boldsymbol{\theta}}_J^{(t)}\|_2}\right)_+ \widehat{\boldsymbol{\theta}}_J^{(t)} & \text{if } J \in \sigma(\mathcal{S}^{(r-1)}), \end{cases}$$

    where $(a)_+ = \max(0, a)$.
  end for
  Update $t = t + 1$.
until converge such that $\sum_J \left\| \widehat{\boldsymbol{\theta}}_J^{(t)} - \widehat{\boldsymbol{\theta}}_J^{(t-1)} \right\|_2 \leq \epsilon$.

It can be verified that when $r = 1$, (14) is equivalent to the sparse additive model in Ravikumar et al. (2009).

Since (14) is quadratic in terms of $\boldsymbol{\theta}_J$ and both the penalty terms are convex, standard convexity theory implies the existence of a global minimizer. We propose a block coordinate descent algorithm to solve (14), which details are given in Algorithm 2. The convergence of block coordinate descent algorithm is studied in Tseng (2001). The derivation of Algorithm 2 is straightforward and hence omitted. In this section, our estimation procedure and algorithm are designed based on basis representation of the functions. We note that in principle, other nonparametric methods such as that of Wang et al. (2016) and Benkeser and van der Laan (2016) can be used to estimate the individual functions in our framework.

3.2 Post-Selection Persistency for Nonparametric Regression

In this section, we show that the sequence of estimators obtained from our proposal is post-selection persistent. The population version of the optimization problem in (10) is

$$\begin{aligned} &\underset{f \in \mathcal{H}^{(r)}, \mathcal{S}(f) = \delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \ \mathbb{E}\left[(Y - f(\boldsymbol{X}))^2\right], \\ &\text{subject to} \sum_{J \in \mathcal{S}^{(r-1)}} \mathbb{E}\left[(P_J(f))^2\right] + \sum_{J \in \sigma(\mathcal{S}^{(r-1)})} \sqrt{\mathbb{E}\left[(P_J(f))^2\right]} \leq \tau, \quad \mathbb{E}\left[P_J(f)\right] = 0, \end{aligned} \tag{15}$$

where the expectation is taken with respect to $\boldsymbol{X}$ and the noise ϵ. To simplify our theoretical analysis, let $f_J(\boldsymbol{X}_J) = \beta_J g_J(\boldsymbol{X}_J)$ and consider the following equivalent population problem

$$\begin{aligned} &\underset{g\in\mathcal{H}^{(r)},\beta_J,\mathcal{S}(g)=\delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \mathbb{E}\left[\left(Y-\sum_{J\in\delta(\mathcal{S}^{(r-1)})}\beta_J g_J(\boldsymbol{X}_J)\right)^2\right], \\ &\text{subject to} \quad \sum_{J\in\mathcal{S}^{(r-1)}}\beta_J^2+\sum_{J\in\sigma(\mathcal{S}^{(r-1)})}|\beta_J|\le\tau, \quad \mathbb{E}[P_J(g)]=0, \quad \mathbb{E}\left[(P_J(g))^2\right]=1. \end{aligned} \tag{16}$$

Problems (16) and (15) are equivalent in the sense that their solutions are equivalent. A similar formulation was also considered in Ravikumar et al. (2009) for sparse additive models.

Let $(\boldsymbol{X}, Y)$ denote a new pair of independent data and define the predictive risk as

$$R(f)=\mathbb{E}\left[(Y-f(\boldsymbol{X}))^2\right].$$

In this section, we assume that our estimator $\widehat{f}^{(r)}$ is chosen to minimize the empirical version of (16). Let

$$\mathcal{F}^{(r)}=\left\{f: f(\boldsymbol{x})=\sum_{J\in\delta(\mathcal{S}^{(r-1)})}\beta_J g_J(\boldsymbol{x}_J), \mathbb{E}[g_J]=0, \|g_J\|_{\mathcal{H}_J}\le 1, \sum_{J\in\mathcal{S}^{(r-1)}}\beta_J^2+\sum_{J\in\sigma(\mathcal{S}^{(r-1)})}|\beta_J|\le\tau\right\}.$$

The following theorem establishes that the sequence of estimators is post-selection persistent.

Theorem 1. *Let $s_0=1$ and let s_{r-1} be the cardinality of the support $\mathcal{S}^{(r-1)}$. Conditioned on $\mathcal{S}^{(r-1)}$, under the square error risk $R(f)=\mathbb{E}[(Y-f(\boldsymbol{X}))^2]$ and for any $1\le r<2m$, we have*

$$R(\widehat{f}^{(r)})-\inf_{f\in\mathcal{F}^{(r)}}R(f)=O_P\left(\tau^2\cdot\sqrt{\frac{rs_{r-1}^2\log d}{n}}\right).$$

Thus, if $\tau=o([n/(rs_{r-1}^2\log d)]^{1/4})$, the estimator $\widehat{f}^{(r)}$ is post-selection persistent.

In other words, Theorem 1 states that conditioned on the selected support on the $(r-1)$th step of our proposed method, $\mathcal{S}^{(r-1)}$, the estimator $\widehat{f}^{(r)}$ converges to the best rth order approximation of the form (1) with support constrained on $\delta(\mathcal{S}^{(r-1)})$. Given the support $\mathcal{S}^{(r-1)}$, the term s_{r-1} is a fixed constant that is much smaller than n. The proof of Theorem 1 involves obtaining the bracketing number of some function classes and applying empirical process tools to obtain an upper bound of the supremum between the empirical and the expected value of the function. The condition $r<2m$ in Theorem 1 is needed to guarantee that the integral of the log bracketing number is well defined. The details are given in Appendix A.

Theorem 1 holds without assuming that the true regression function $f(\cdot)$ takes the form of SSANOVA framework in (1). In addition, we do not impose any distributional assumptions on $(Y, \boldsymbol{X})$. We recover the persistency result in Ravikumar et al. (2009) for the sparse additive model as a special case when $r=1$ and $m=2$.

3.3 From Post-Selection Persistency to Persistency

Post-selection persistency results can serve as motivations for intermediate steps of LASER. However, the results do not concern properties of the final estimator. In this section, we establish sufficient conditions such that a post-selection persistency result can be strengthen into a persistency result for the final estimator. For simplicity, we consider nonparametric regression models with two-way interaction terms, i.e., $f \in \mathcal{H}^{(2)}$.

In general, without any conditions on the bivariate functions f_{jk}, it is extremely challenging to show that the main effects f_j can be identified without modeling the two-way interaction effects. This problem is related to proving model selection consistency under model misspecification for sparse additive model, and such theoretical results have not been well established in the literature. In fact, the same problem is not well understood in the context of linear regression with two-way interaction effects until recently (Hao et al., 2018).

In the following, we impose sufficient conditions on the bivariate functions such that the active main effects can be identified at the first stage of LASER in the context of nonparametric regression. Therefore, conditioned on the correctly identified main effects, the estimator obtained from the second stage is persistent. With some abuse of notation, let $\bar{\mathcal{S}}^{(1)}$ and $\bar{\mathcal{S}}^{(2)}$ be two sets containing indices for the underlying active main and bivariate effects, respectively. Assume that $\bar{\mathcal{S}}^{(2)}$ satisfies the hierarchical structural assumption in (2).

Recall from Section 3.1 that we approximate the main effects by its kth order basis expansion, i.e., $\boldsymbol{\theta}^* = \mathrm{argmin}_{\boldsymbol{\theta}=(\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_d^T)^T} \| \sum_{j=1}^d f_j - \sum_{j=1}^d \boldsymbol{\theta}_j^T \boldsymbol{\phi}_j \|_2^2$. In the following proposition, our conditions are written in terms of an approximation of both the main and the partial derivatives of the bivariate function using first order basis expansion. To this end, we define some additional notation. Given a bivariate function $g(x_1, x_2)$, let

$$g^{(2)}(x_1,x_2)\big|_{a_1,a_2} = g(a_1,a_2) + \sum_{j\in\{1,2\}} \left(\frac{\partial g(a_1,a_2)}{\partial x_j}(x_j - a_j) + \frac{1}{2}\frac{\partial^2 g(a_1,a_2)}{\partial x_j^2}(x_j-a_j)^2 \right),$$

and

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_d^T)^T}{\mathrm{argmin}} \left\| \sum_{j=1}^d f_j + \sum_{s<t} f_{st}^{(2)}\big|_{1/2,1/2} - \sum_{j=1}^d \boldsymbol{\beta}_j^T \boldsymbol{\phi}_j \right\|_2^2.$$

The following proposition establishes that the true underlying support, $\bar{\mathcal{S}}^{(1)}$, is a subset of the estimated support from first stage of LASER, $\mathcal{S}^{(1)}$.

Proposition 1. *Assume that $f \in \mathcal{H}^{(2)}$. Suppose that there exist a positive constant $C_{\min}$ such that the minimum eigenvalue $\lambda_{\min}(\boldsymbol{\Phi}_{\bar{\mathcal{S}}^{(1)}}^T \boldsymbol{\Phi}_{\bar{\mathcal{S}}^{(1)}}) \geq C_{\min} > 0$. Let $\rho_n^* = \min_{j\in\bar{\mathcal{S}}^{(1)}} \|\boldsymbol{\beta}_j^*\|_\infty$ and $q_n^* = \| \sum_{(s,t)\in\bar{\mathcal{S}}^{(2)}} \partial_{x_s x_t}^2 f_{st} \|_\infty$. Assume that $\sqrt{|\bar{\mathcal{S}}^{(1)}|k} \cdot q_n^*/\rho_n^* = o(1)$, $\sqrt{k} q_n^*/\lambda = o(1)$, and*

$$\frac{1}{\rho_n^*}\left[\sqrt{\log(|\bar{\mathcal{S}}^{(1)}|k)/n} + |\bar{\mathcal{S}}^{(1)}|^{3/2}/k + \lambda\sqrt{|\bar{\mathcal{S}}^{(1)}|k} \right] = o(1).$$

We have $\mathbb{P}(\mathcal{S}^{(1)} \supseteq \bar{\mathcal{S}}^{(1)}) \to 1$.

The proof of Proposition 1 is similar to that of Theorem 2 in Ravikumar et al. (2009): we provide a sketch proof in Appendix B. Intuitively, the conditions on q_n^* states that the bivariate functions should be sufficiently smooth relative to the signal ρ_n^*. As a reviewer

pointed out, this implies that the bivariate functions should be close to linear, relative to the main effects. We have removed the incoherence condition in Ravikumar et al. (2009) and we now allow the active and non-active main effects to be correlated. Relaxing the imposed assumptions on the bivariate functions is out of the scope of this paper, and we leave it as an open problem for future research.

4. Nonparametric Graphical Models

Undirected graphical models, also known as Markov random field, have been used extensively to model the conditional dependence relationships among a set of random variables. In a graph, each node represents a random variable and an edge between two nodes indicates that the two random variables are conditionally dependent, given all of the other variables. Let $\boldsymbol{X}$ be a d-dimensional random variable with joint density function of the form

$$p(\boldsymbol{x}) = \frac{1}{Z(f)} \exp(-f(\boldsymbol{x})), \tag{17}$$

where $Z(f) = \int \exp(-f(\boldsymbol{x}))d\boldsymbol{x}$ is the partition function such that the density $p(\boldsymbol{x})$ integrates to one. By the Hammersley-Clifford theorem, a set of random variables $\boldsymbol{X}$ forms a Markov random field with respect to a graph G if $f(\boldsymbol{x})$ takes the form $f(\boldsymbol{x}) = \sum_{J \in J_G} f_J(\boldsymbol{x}_J)$, where J_G is a set of all cliques in G and f_J is the potential function. For a set $J \subseteq [d]$, if $f_J(\boldsymbol{x}_J) \neq 0$, then the set of random variables $\boldsymbol{X}_J$ forms a clique and are conditionally dependent given all of the other variables.

Currently, most of the research on graphical models are limited to the case when the maximal clique is of size two. This is referred to as the pairwise Markov random field with the following joint density

$$p(\boldsymbol{x}) = \frac{1}{Z(f)} \exp\left\{ -\sum_{j=1}^{d} f_j(x_j) - \sum_{j<k} f_{jk}(x_j, x_k) \right\}. \tag{18}$$

Under the pairwise Markov random field, the jth and kth random variables are conditionally independent if and only if $f_{jk}(x_j, x_k) = 0$. The pairwise Markov random field in (18) is fully nonparametric and consists of many recently studied pairwise graphical models as its special cases.

Example 1. Gaussian graphical models*: Let* $\boldsymbol{X} \sim N_d(\boldsymbol{0}, \boldsymbol{\Sigma})$ *and let* $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ *be the inverse covariance matrix. The Gaussian graphical model has joint density*

$$p(\boldsymbol{x}) \propto \exp\left\{ -\frac{1}{2} \sum_{j=1}^{d} \Theta_{jj} x_j^2 - \sum_{j=1}^{d-1} \sum_{k>j} \Theta_{jk} x_j x_k \right\}.$$

Thus, the Gaussian graphical model is a special case of (18) with $f_{jk}(x_j, x_k) = \Theta_{jk} x_j x_k$ *and* $f_j(x_j) = \frac{1}{2}\Theta_{jj} x_j^2$*. The Gaussian graphical model is well studied in the literature (see, for instance, Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008; Rothman et al., 2008; Peng et al., 2009; Ravikumar et al., 2011; Cai et al., 2011; Sun and Zhang, 2013; Tan et al., 2014, 2015; Liu and Luo, 2015; Drton and Maathuis, 2017).*

Example 2. Exponential family graphical models*: The exponential family graphical model has joint density*

$$p(\boldsymbol{x}) \propto \exp\left\{\sum_{j=1}^{d}(t(x_j)+C(x_j))+\sum_{j=1}^{d-1}\sum_{k>j}\Theta_{jk}t(x_j)t(x_k)\right\},$$

where $t(x_j)$ is a univariate sufficient statistics function, $C(x_j)$ is some function of x_j specified by the exponential family distribution, and Θ_{jk} is the canonical parameter. Thus, this is a special case of (18) with $f_{jk}(x_j,x_k)=-\Theta_{jk}t(x_j)t(x_k)$ and $f_j(x_j)=-t(x_j)-C(x_j)$. This model is recently studied by many authors (see, for instance, Yang et al., 2013, 2015; Tan et al., 2016; Yang et al., 2018; Chen et al., 2014)

Example 3. Nonparanormal graphical models*: Let $g=\{g_1,\ldots,g_d\}$ be a set of monotone univariate functions. A d-dimensional random vector $\boldsymbol{X}$ has a nonparanormal distribution $\boldsymbol{X}\sim \mathrm{NPN}_d(g,\boldsymbol{\Sigma})$ if $g(\boldsymbol{X})\sim N_d(\boldsymbol{0},\boldsymbol{\Sigma})$. Let $\boldsymbol{\Theta}=\boldsymbol{\Sigma}^{-1}$. Then, the nonparanormal graphical model has joint density*

$$p(\boldsymbol{x}) \propto \exp\left\{\sum_{j=1}^{d}\left(-\frac{1}{2}\Theta_{jj}g_j(x_j)^2+\log\left|g_j'(x_j)\right|\right)-\sum_{j=1}^{d-1}\sum_{k>j}\Theta_{jk}g_j(x_j)g_k(x_k)\right\}.$$

This is a special case of (18) with $f_{jk}(x_j,x_k)=\Theta_{jk}g_j(x_j)g_k(x_k)$ and $f_j(x_j)=\Theta_{jj}g_j(x_j)^2/2-\log|g_j'(x_j)|$. This model is studied in Liu et al. (2009) and Liu et al. (2012).

We consider modeling the Markov random field in (17) under the SSANOVA framework, that is, the function $f(\cdot)$ can be decomposed as in (1). This general model has been considered in the literature and an estimate of $f(\cdot)$ can be obtained by optimizing over the penalized maximum likelihood function (see, for instance, Leonard, 1978; Silverman, 1982; Gu and Wang, 2003; Jeon and Lin, 2006). However, due to the log-partition function $Z(f)$, the proposed algorithms are not scalable to large dimension and higher order interaction terms. To the best of our knowledge, most methods involve truncating the functional decomposition (1) to only modeling the two-way interaction terms, which corresponds to pairwise nonparametric graphical models in (18).

In this section, we propose a novel method to estimate nonparametric graphical models of the form (17) without restricting it to only modeling two-way interaction terms. More specifically, we are interested in estimating nonparametric graphical models with joint density function

$$p(\boldsymbol{x})=\frac{1}{Z(f)}\exp(-f(\boldsymbol{x})) \qquad \text{and} \qquad f(\boldsymbol{x})=\mu+\sum_{j=1}^{d}f_j(x_j)+\sum_{j<k}f_{jk}(x_j,x_k)+\cdots.$$

Rather than using the penalized maximum likelihood function to estimate $f(\cdot)$, we propose to estimate $f(\cdot)$ under the score matching loss function proposed in Hyvärinen (2005) and Hyvärinen (2007), which is independent of the log-partition function $Z(f)$ that is computationally intractable. Thus, our algorithm is scalable to the dimension of the random variables as well as the size of cliques among the random variables compared to existing proposals. We also show that our proposal is post-selection persistent under the score matching risk in the high-dimensional setting.

4.1 Score Matching Loss

The score matching loss function was introduced to estimate densities of the form (17), which involves a computationally intractable log-partition function $Z(f)$. In the following, we provide a brief discussion on the score matching loss and refer the reader to Hyvärinen (2005) and Hyvärinen (2007) for more details. Let $\boldsymbol{X}$ be a d-dimensional continuous random vector with distribution $\mathcal{P}$ and joint density function $p(\cdot)$. For a twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we define the Laplacian operator and the gradient of $f(\cdot)$ as

$$\Delta f(\boldsymbol{x}) = \sum_{j=1}^{d} \frac{\partial^2}{\partial x_j^2} f(\boldsymbol{x}) \in \mathbb{R} \qquad \text{and} \qquad \nabla_j f(\boldsymbol{x}) = \frac{\partial}{\partial x_j} f(\boldsymbol{x}), \tag{19}$$

respectively. For a distribution $\mathcal{Q}$ with density $q(\cdot)$, Hyvärinen (2005) defined the score matching loss of $\mathcal{Q}$ with respect to $\mathcal{P}$ as

$$\frac{1}{2} \int_{\mathbb{R}^d} p(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) - \nabla \log q(\boldsymbol{x})\|_2^2 d\boldsymbol{x}. \tag{20}$$

Equation (20) is also referred to as the Fisher divergence. It can be seen that (20) is minimized as a function of $\mathcal{Q}$ when $\mathcal{Q} = \mathcal{P}$, which depends on the true distribution $\mathcal{P}$. Hyvärinen (2005) showed that under the condition that $\|p(\boldsymbol{x}) \nabla \log q(\boldsymbol{x})\|_2 \to 0$ as $\|\boldsymbol{x}\|_2 \to \infty$, the score matching loss can be rewritten as

$$\int_{\mathbb{R}^d} p(\boldsymbol{x}) \left[\Delta \log q(\boldsymbol{x}) + \frac{1}{2} \|\nabla \log q(\boldsymbol{x})\|_2^2 \right] d\boldsymbol{x} + C, \tag{21}$$

where C is a constant that is independent of $\mathcal{Q}$. The term in the integrand (21) is referred to as the Hyvärinen scoring rule, and no longer depends on the true distribution $\mathcal{P}$ and the log-partition function. Thus, an estimator of $f(\cdot)$ can be obtained by minimizing the Hyvärinen scoring rule. The statistical properties of the estimator obtained by minimizing the Hyvärinen scoring rule have been studied in the classical setting in which $n > d$ (among others, Hyvärinen, 2005, 2007; Forbes and Lauritzen, 2015).

Recently, Lin et al. (2016) proposed to estimate parametric pairwise graphical models in the high-dimensional setting in which $d > n$ under the score matching loss function. In addition, in his dissertation, Janofsky (2015) proposed to estimate fully nonparametric pairwise graphical models as in (18) using the score matching loss function. However, their proposal is limited to pairwise interactions between two random variables and are not able to estimate clique of size greater than two in a graph. We now generalize the aforementioned proposals to accommodate general nonparametric graphical models of the form (17) using the score matching loss function.

We start with establishing a proper score matching loss function for estimating nonparametric graphical models in (17). In the context of nonparametric graphical model setting, we consider distribution $\mathcal{P}$ that is supported on $[0, 1]^d$. The Hyvärinen scoring rule (21) no longer applies since it is derived for distribution that is supported on $\mathbb{R}^d$. We now make a modification to the Hyvärinen scoring rule for densities with support $[0, 1]^d$. To this end, we define $r_j(x_j)$ to be a function of x_j and $\mathbf{r}(\boldsymbol{x}) = (r_1(x_1), \ldots, r_d(x_d))^T$. We define $\mathbf{r}'(\boldsymbol{x}) = (r_1'(x_1), \ldots, r_d'(x_d))^T$ to be the element-wise differentiation of the vector $\mathbf{r}(\boldsymbol{x})$, that

is, $r_j'(x_j) = \partial r_j(x_j)/\partial x_j$. We define the modified score matching loss of $\mathcal{Q}$ with respect to $\mathcal{P}$ as

$$\frac{1}{2}\int_{[0,1]^d} p(\boldsymbol{x})\,\|\mathbf{r}(\boldsymbol{x}) \circ [\nabla \log p(\boldsymbol{x}) - \nabla \log q(\boldsymbol{x})]\|_2^2\, d\boldsymbol{x}, \tag{22}$$

where $\circ$ is the Hadamard product between two vectors. The following lemma establishes a scoring rule similar to that of (21) for random variables $\boldsymbol{X} \in [0,1]^d$.

Lemma 1. *Assume that the density $p(\boldsymbol{x})$ for $\mathcal{P}$ satisfies the regularity conditions that*

$$\lim_{x_j \to 0} p(\boldsymbol{x}) \cdot \nabla_j \log q(\boldsymbol{x}) r_j^2(x_j) \to 0 \qquad \text{and} \qquad \lim_{x_j \to 1} p(\boldsymbol{x}) \cdot \nabla_j \log q(\boldsymbol{x}) r_j^2(x_j) \to 0.$$

for any $1 \le j \le d$. Then, the modified score matching loss can be written as

$$\int_{[0,1]^d} p(\boldsymbol{x}) S(\boldsymbol{x}, q) d\boldsymbol{x} + C, \text{ where}$$

$$S(\boldsymbol{x}, q) = 2\left(\mathbf{r}(\boldsymbol{x}) \circ \mathbf{r}'(\boldsymbol{x})\right)^T \nabla \log q(\boldsymbol{x}) + \left(\mathbf{r}(\boldsymbol{x}) \circ \mathbf{r}(\boldsymbol{x})\right)^T \nabla^2 \log q(\boldsymbol{x}) + \frac{1}{2}\|\mathbf{r}(\boldsymbol{x}) \circ \nabla \log q(\boldsymbol{x})\|_2^2, \tag{23}$$

C is some constant independent of $\mathcal{Q}$, and $\nabla^2 \log q(\boldsymbol{x})$ is a vector of second order derivative of $\boldsymbol{x}$.

The assumption in Lemma 1 requires that $r_j(x_j) \to 0$ as $x_j \to 0$ and $x_j \to 1$. One possible choice of r_j is $r_j(x_j) = x_j(1 - x_j)$, which was considered in Janofsky (2015) in the context of pairwise nonparametric graphical models. Thus, an estimate of $f(\cdot)$ for the nonparametric graphical model (17) can be obtained by minimizing the modified scoring rule

$$S(\boldsymbol{x}, f) = -2\sum_{j=1}^{d} r_j(x_j) r_j'(x_j) f^{(j)}(\boldsymbol{x}) - \sum_{j=1}^{d} r_j^2(x_j) f^{(jj)}(\boldsymbol{x}) + \frac{1}{2}\sum_{j=1}^{d} r_j^2(x_j)\left(f^{(j)}(\boldsymbol{x})\right)^2, \tag{24}$$

where $f^{(j)}$ and $f^{(jj)}$ are the first and second order derivative of $f(\boldsymbol{x})$ with respect to x_j, respectively. From (24), we see that the score matching loss function depends only on $\nabla \log q(\boldsymbol{x})$ and $\nabla^2 \log q(\boldsymbol{x})$, which are free of the log-partition function $Z(f)$.

4.2 Method and Optimization Problem

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n independent and identically distributed observations drawn from $\mathcal{P}$ with support $[0,1]^d$. To estimate the conditional dependencies among the random variables, we fit the model in (7) with the score matching loss function $\mathcal{L}_n(D, f) = \frac{1}{n}\sum_{i=1}^{n} S(\mathbf{x}_i, f)$, where $S(\mathbf{x}_i, f)$ is as defined in (24).

In the context of graphical models, the main effect f_j is always non-zero unless the jth random variable is uniformly distributed. Thus, we start estimating nonparametric graphical models with the second step of our algorithm since we do not have to perform

variable selection for the main effects. This yields the following optimization problem at the rth step of our proposed algorithm

$$\begin{aligned} &\underset{f\in\mathcal{H}^{(r)},\mathcal{S}(f)=\delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n} S(\mathbf{x}_i, f), \\ &\text{subject to} \quad \sum_{J\in\mathcal{S}^{(r-1)}} \|P_J(f)\|_2^2 + \sum_{J\in\sigma(\mathcal{S}^{(r-1)})} \|P_J(f)\|_2 \le \tau. \end{aligned} \tag{25}$$

Similar to Section 3.1, we solve the penalized version of (25) in terms of its basis expansion. To this end, we define additional notation. Let $\boldsymbol{\phi}_J^{(j)}(\boldsymbol{x})$ and $\boldsymbol{\phi}_J^{(jj)}$ be the first and second order derivative of $\boldsymbol{\phi}_J(\boldsymbol{x})$ with respect to x_j, respectively. Similarly, let $\boldsymbol{\Phi}_J^{(j)}$ and $\boldsymbol{\Phi}_J^{(jj)}$ denote the $n \times k^{|J|}$ matrix with rows $\boldsymbol{\phi}_J^{(j)}(\mathbf{x}_1)^T, \ldots, \boldsymbol{\phi}_J^{(j)}(\mathbf{x}_n)^T$ and $\boldsymbol{\phi}_J^{(jj)}(\mathbf{x}_1)^T, \ldots, \boldsymbol{\phi}_J^{(jj)}(\mathbf{x}_n)^T$, respectively. Writing (25) in terms of its basis expansion yields the optimization problem

$$\underset{\boldsymbol{\theta}_J,\mathcal{S}(f)=\delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \ \frac{1}{n}\sum_{i=1}^{n} S_\phi(\mathbf{x}_i, \boldsymbol{\theta}) + \lambda\Big(\frac{1}{n}\sum_{J\in\mathcal{S}^{(r-1)}} \|\boldsymbol{\Phi}_J\boldsymbol{\theta}_J\|_2^2 + \frac{1}{\sqrt{n}}\sum_{J\in\sigma(\mathcal{S}^{(r-1)})} \|\boldsymbol{\Phi}_J\boldsymbol{\theta}_J\|_2\Big), \tag{26}$$

where $S_\phi(\mathbf{x}_i, \boldsymbol{\theta})$ is a function of $\boldsymbol{\Phi}_J, \boldsymbol{\Phi}_J^{(j)}$, and $\boldsymbol{\Phi}_J^{(jj)}$. Problem (26) is convex and can be solved directly via the block coordinate descent algorithm (Tseng, 2001).

The block coordinate descent algorithm involves cycling through the updates for $\boldsymbol{\theta}_J$ for all J until convergence. Since the loss function $\frac{1}{n}\sum_{i=1}^{n} S_\phi(\mathbf{x}_i, \boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}_J$, there is a closed form update for any $J \in \mathcal{S}^{(r-1)}$. However, for $J \in \sigma(\mathcal{S}^{(r-1)})$, there is no closed form update for $\boldsymbol{\theta}_J$ due to the composite function in the group lasso penalty. In the context of pairwise nonparametric graphical models, Janofsky (2015) proposed to use the alternating direction method of multiplies algorithm to obtain updates for $\boldsymbol{\theta}_J$ with $J \in \sigma(\mathcal{S}^{(1)})$. For higher order terms, a similar algorithm can be used. We omit the details and refer the reader to Janofsky (2015) for the derivation of the algorithm.

4.3 Post-Selection Persistency for the Nonparametric Graphical Model

We now establish that the sequence of estimators $\{\widehat{f}^{(\ell)}\}_{\ell=1}^{r}$ obtained from solving (25) is post-selection persistent under the score matching risk function. For density estimation, a natural risk function is the distance between two density functions. There are various measures to quantify the distance between two density functions p and q. One of the most popular distance measure is the Kullback-Leibler divergence. Since we derive the score matching loss function based on the Fisher divergence criterion, it is natural to define the risk function using the Fisher divergence:

$$R(p, q) = \frac{1}{2}\int_{[0,1]^d} p(\boldsymbol{x}) \left\| \mathbf{r}(\boldsymbol{x}) \circ \nabla \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right\|_2^2 d\boldsymbol{x}. \tag{27}$$

The population version of the optimization problem in (25) is

$$\begin{aligned} &\underset{f\in\mathcal{H}^{(r)},\mathcal{S}(f)=\delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \ \mathbb{E}\left[S(\boldsymbol{X}, f)\right], \\ &\text{subject to} \quad \sum_{J\in\mathcal{S}^{(r-1)}} \mathbb{E}\left[(P_J(f))^2\right] + \sum_{J\in\sigma(\mathcal{S}^{(r-1)})} \sqrt{\mathbb{E}\left[(P_J(f))^2\right]} \le \tau, \quad \mathbb{E}\left[P_J(f)\right] = 0, \end{aligned} \tag{28}$$

where the expectation is taken with respect to the random variables $\boldsymbol{X}$. Similar to Section 3.2, we let $f_J(\boldsymbol{X}_J) = \beta_J g_J(\boldsymbol{X}_J)$ and consider the following equivalent population problem

$$\begin{aligned} &\underset{g\in\mathcal{H}^{(r)},\beta_J,\mathcal{S}(g)=\delta(\mathcal{S}^{(r-1)})}{\text{minimize}} \mathbb{E}\left[S(\boldsymbol{X},\boldsymbol{\beta},g)\right], \\ &\text{subject to} \quad \sum_{J\in\mathcal{S}^{(r-1)}} \beta_J^2 + \sum_{J\in\sigma(\mathcal{S}^{(r-1)})} |\beta_J| \le \tau, \quad \mathbb{E}[P_J(g)] = 0, \quad \mathbb{E}\left[(P_J(g))^2\right] = 1. \end{aligned} \tag{29}$$

For theoretical purposes, at the rth step of LASER, we assume that the estimator is chosen to minimize the empirical version of (29). Recall that

$$\mathcal{F}^{(r)} = \left\{ f : f(\boldsymbol{x}) = \sum_{J\in\delta(\mathcal{S}^{(r-1)})} \beta_J g_J(\boldsymbol{x}_J), \mathbb{E}[g_J] = 0, \|g_J\|_{\mathcal{H}_J} \le 1, \sum_{J\in\mathcal{S}^{(r-1)}} \beta_J^2 + \sum_{J\in\sigma(\mathcal{S}^{(r-1)})} |\beta_J| \le \tau \right\}.$$

We consider the following density function class

$$\mathcal{Q}^{(r)} = \{q \mid q \propto \exp(-f(\boldsymbol{x})), f \in \mathcal{F}^{(r)}\}.$$

We now state the main theorem on the post-selection persistency property of the estimator $\widehat{p}^{(r)}$ obtained from the rth step of LASER.

Theorem 2. *Let $\widehat{p}^{(r)} \propto \exp(-\widehat{f}^{(r)})$ and let s_{r-1} be the cardinality of the support $\mathcal{S}^{(r-1)}$. Given $\mathcal{S}^{(r-1)}$, for any $1 \le r < 2(m-2)$, $\widehat{p}^{(r)}$ is post-selection persistent under the Fisher divergence risk function*

$$R(p,\widehat{p}^{(r)}) - \inf_{q\in\mathcal{Q}^{(r)}} R(p,q) = O_P\left(\tau^2\sqrt{\frac{r^3 s_{r-1}^2 \log d}{n}}\right).$$

Thus, if $\tau = o([n/(r^3 s_{r-1}^2 \log d)]^{1/4})$, then the estimator $\widehat{p}^{(r)}$ is post-selection persistent given $\mathcal{S}^{(r-1)}$.

Theorem 2 states that conditioned on the support, $\mathcal{S}^{(r-1)}$, the estimator $\widehat{p}^{(r)}$ converges to the best rth order approximation of the form in $\mathcal{Q}^{(r)}$. The proof of Theorem 2 is given in Appendix D.

5. Numerical Studies

We perform numerical studies for both nonparametric regression and nonparametric graphical models.

5.1 Nonparametric Regression

We perform extensive numerical studies to evaluate the performance of our proposal for fitting multivariate nonparametric regression. In all of our numerical studies, we generate a training set and a test set. Each model is fit using the training set, and the trained model is used to predict the response on the test set. To compare the performance across different

methods, we calculate the sum of squares error between the predicted response and the true response from the test set. These results are reported in Tables 1—3. In addition, we report in Appendix F the true and false positive rates for the main effects and interaction effects, defined as the proportion of correctly estimated active variables and the proportion of inactive variables that are incorrectly estimated to be active, respectively.

Seven approaches are compared in our numerical studies: our proposal, `LASER`; the sparse additive model, `SpAM` (Ravikumar et al., 2009); the nonparametric additive regression model with two-way interactions, `VANISH` (Radchenko and James, 2010); the backtracking method for modeling high-dimensional linear regression with two-way interaction terms, `BT` (Shah, 2016); the convex modeling of interactions with strong heredity, `FAMILY` (Haris et al., 2016); the regularization approach for high-dimensional quadratic regression, RAMP (Hao et al., 2018); the oracle approach by assuming that the active variables were known a priori, `ORACLE`. In particular, the `ORACLE` is obtained by fitting nonparametric regression model (14) using only the active variables with the ridge penalty for smoothness.

Our proposal `LASER`, `SpAM`, `VANISH`, and `ORACLE` are nonparametric. `LASER` involves fitting the nonparametric regression model (14) sequentially. In each step of Algorithm 1, we select the tuning parameter using a five-fold cross-validation on the training data set. Algorithm 1 is stopped when there is no more higher order interaction terms to be estimated. We fit `LASER` using the kth basis expansion with $k = 3$. Note that the solution for `SpAM` can be obtained from the first layer of `LASER`. For `VANISH`, we simply use the default setting as in Radchenko and James (2010) and select the tuning parameter with cross-validation. For `ORACLE`, we select the tuning parameter that yields the smallest sum of squares error on the test set. In other words, `ORACLE` serves as a gold standard for nonparametric regression. The `FAMILY`, `RAMP`, and `BT` are sparse high-dimensional linear regression with two-way interaction terms. The tuning parameters for `RAMP` are selected using the extended BIC described in Hao et al. (2018). For `FAMILY` and `BT`, we consider a fine grid of tuning parameters and report the best results. In other words, we are giving unfair advantage to `FAMILY` and `BT`.

Most methods that estimate pairwise interaction terms are not computationally feasible for high-dimensional problem when d is large. Therefore, we consider both the low-dimensional and high-dimensional settings in Sections 5.1.1 and 5.1.2, respectively. We then perform numerical studies to assess how correlation among the covariates affects the performance of `LASER` in Section 5.1.3.

5.1.1 Low-Dimensional Setting with Two-Way Interactions

In our simulation studies, we generate $\epsilon \sim N(0, 1)$ and each element of $\boldsymbol{X}$ from a uniform distribution on the interval $[0, 1]$. We consider the following regression models with $d = 30$ covariates and $n = \{200, 400\}$:

A1 — A linear regression model with two-way interaction terms:

$$y = x_1 + x_2 + x_3 + 5x_1x_2 - 2x_1x_3 + 5x_2x_3 + \epsilon.$$

A2 — A non-linear regression model with two-way interaction terms (product of two individual functions):

$$f_1(z) = \sqrt{2}\left[\sin(6z) - 0.0066\right], \qquad f_2(z) = \sqrt{11}[(2z-1)^2 - 1/3],$$
$$f_3(z) = \sqrt{12}(z - 1/2), \qquad f_4(z) = \sqrt{16.6}[\exp(-5z) - 0.2],$$
$$f_5(z) = \sqrt{50}[1/(1+z) - 0.69],$$
$$y = \sum_{j=1}^{4} f_j(x_j) + f_3(x_1)f_3(x_2) + f_5(x_2)f_4(x_4) + f_3(x_3)f_4(x_4) + \sqrt{0.5}\epsilon,$$

where the constants are designed such that each function has mean zero and variance approximately one.

A3 — A non-linear regression model with two-way interaction terms (bivariate functions that cannot be decomposed as product of two individual functions):

$$y = \sum_{j=1}^{3} f_j(x_j) + \sqrt{19}(\sqrt{x_1 x_2} - 4/9) + \sqrt{50}[\exp(-5x_2x_3) - 0.438] + \sqrt{0.5}\epsilon,$$

where $f_j(x_j)$ is as defined in Scenario A2.

The results, averaged over 200 data sets, are reported in Table 1.

Table 1: The sum of squares error (standard error) out of 200 test samples for the three different scenarios in Section 5.1.1, averaged over 200 data sets. The results are for models trained with n training samples. Numbers are rounded to the nearest integer.

	Scenario A1		Scenario A2		Scenario A3	
	$n = 200$	$n = 400$	$n = 200$	$n = 400$	$n = 200$	$n = 400$
FAMILY	273 (2)	255 (2)	825 (10)	615 (7)	412 (3)	379 (3)
BT	216 (1)	203 (2)	766 (14)	751 (13)	411 (3)	407 (3)
RAMP	216 (2)	207 (2)	848 (21)	672 (22)	333 (9)	300 (9)
SpAM	323 (3)	296 (2)	825 (10)	743 (10)	91 (1)	84 (1)
VANISH	289 (3)	234 (2)	123 (3)	83 (1)	84 (1)	74 (1)
LASER	279 (4)	226 (2)	124 (4)	84 (1)	75 (1)	63 (1)
ORACLE	246 (2)	216 (2)	91 (1)	76 (1)	64 (1)	59 (1)

From Table 1, we see that BT and RAMP have the best performance in Scenario A1. This is not surprising since BT and RAMP are designed for modeling linear regression with two-way interaction terms. Both BT and RAMP outperform FAMILY that models the two-way interaction terms using a hierarchical penalty. For the nonparametric methods, SpAM has the highest sum of squares error since the true model contains two-way interaction terms that SpAM fails to model. VANISH and LASER have similar performance, and perform significantly better than SpAM. As we increase the sample size from $n = 200$ to $n = 400$, we see that the performance of LASER becomes more comparable to that of ORACLE, BT, and RAMP. For Scenarios A2—A3, FAMILY, RAMP, and BT have the highest sum of squares error since these methods are intended for modeling linear regression. Again, VANISH and LASER have similar

performance, and outperform SpAM. In summary, LASER is able to adaptively estimate the higher order terms accurately in both linear and non-linear regression settings.

5.1.2 High-Dimensional Setting with Three-Way Interactions

In this section, we consider the high-dimensional setting in which the number of variables d is potentially larger than the number of observations. FAMILY and VANISH are computationally infeasible since there are a total of $d + \binom{d}{2}$ parameters and functions to estimate. Moreover, we consider settings with a three-way interaction effect to illustrate the flexibility of our proposal compared to existing methods such as BT and RAMP that are limited to modeling two-way interaction terms. We generate ϵ and $\boldsymbol{X}$ as in Section 5.1.1. We consider three regression models with $d = \{200, 400\}$ covariates and $n = \{350, 700\}$:

B1 — A linear regression model with three-way interaction terms:

$$y = x_1 + x_2 + x_3 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3 + 25x_1x_2x_3 + \sqrt{0.5}\epsilon.$$

B2 — A non-linear regression model with three-way interaction terms (product of individual functions):

$$y = \sum_{j=1}^{3} f_j(x_j) + f_3(x_1)f_3(x_2) + f_5(x_2)f_4(x_3) + f_1(x_1)f_4(x_3) + 25(x_1x_2x_3 - 1/8) + \sqrt{0.5}\epsilon,$$

where $f_j(\cdot)$ is as defined in Scenario A2.

B3 — A non-linear regression model with three-way interaction terms (bivariate functions that cannot be decomposed as product of two individual functions):

$$\begin{aligned} y = \sum_{j=1}^{3} f_j(x_j) + \sqrt{19}(\sqrt{x_1x_2} - 4/9) + \sqrt{50}[\exp(-5x_2x_3) - 0.438] \\ + f_3(x_1)f_4(x_3) + 25(x_1x_2x_3 - 1/8) + \sqrt{0.5}\epsilon, \end{aligned}$$

where $f_j(\cdot)$ is as defined in Scenario A2.

The results, averaged over 200 data sets, are reported in Table 2.

From Table 2, we see that LASER has the best performance across all scenarios since it is the only method that models the three-way interaction term. The sum of squares error is quite close to that of ORACLE even in the high-dimensional setting when $n = 350$ and $d = 400$. As we increase the sample size, we see that the performance of LASER becomes more similar to that of ORACLE.

5.1.3 Two-Way Interactions With Correlated Data

In this section, we assess the performance of LASER when the covariates $\boldsymbol{X}$ are correlated. We generate $\epsilon \sim N(0, 1)$ and $\boldsymbol{X} \sim N_d(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{jk} = \rho^{|j-k|}$ for $1 \leq j, k \leq d$. For Scenario C2, we normalize the covariates such that the observed values lie within the unit interval. We consider the following regression models with $d = 30$, $n = 200$, and $\rho = \{0, 0.2, 0.4, 0.6, 0.8\}$:

Table 2: The sum of squares error (standard error) out of 200 test samples for the three different scenarios in Section 5.1.2, averaged over 200 data sets. The results are for models trained with n training samples. Numbers are rounded to the nearest integer.

		Scenario B1		Scenario B2		Scenario B3	
		$n = 350$	$n = 700$	$n = 350$	$n = 700$	$n = 350$	$n = 700$
$d = 200$	BT	134 (1)	127 (1)	543 (6)	513 (5)	570 (5)	539 (4)
	RAMP	128 (1)	125 (1)	336 (6)	309 (3)	308 (4)	291 (3)
	SpAM	745 (11)	663 (9)	2203 (28)	2008 (21)	876 (13)	780 (10)
	LASER	95 (4)	71 (2)	149 (11)	102 (3)	127 (7)	82 (2)
	ORACLE	73 (1)	62 (1)	107 (3)	90 (2)	86 (1)	72 (1)
$d = 400$	BT	136 (1)	128 (1)	548 (6)	521 (6)	578 (5)	550 (5)
	RAMP	127 (1)	124 (1)	402 (9)	309 (4)	339 (7)	288 (3)
	SpAM	755 (11)	670 (9)	2224 (30)	2017 (24)	896 (13)	792 (11)
	LASER	97 (3)	70 (1)	166 (11)	95 (2)	122 (6)	80 (2)
	ORACLE	74 (1)	62 (1)	109 (3)	87 (2)	87 (1)	71 (1)

C1 — A linear regression model with two-way interaction terms:

$$y = x_1 + x_6 + x_{11} + 0.5x_1x_6 - 0.5x_1x_{11} + 0.5x_6x_{11} + \epsilon.$$

C2 — A non-linear regression model with two-way interaction terms:

$$y = \sum_{j=1}^{4} f_j(x_j) + f_3(x_1)f_3(x_6) + f_5(x_6)f_4(x_{16}) + f_3(x_{11})f_4(x_{16}) + \sqrt{0.5}\epsilon,$$

where the $f_j(\cdot)$ is as defined in Scenario A2.

The results, averaged over 200 data sets, are reported in Table 3. From Table 3, we see that LASER and ORACLE have similar sum of squares error. Moreover, LASER outperforms SpAM significantly. These results suggest that LASER is able to select the important main effects and interaction effects even when there are correlation among covariates.

5.2 Nonparametric Graphical Models

In this section, we perform some numerical studies to estimate nonparametric graphical models. We compare LASER to the graphical lasso (glasso) which estimate pairwise Gaussian graphical models (Friedman et al., 2008). We also consider the proposal of Liu et al. (2012) (kendall), a semiparametric approach for estimating nonparanormal graphical models. To evaluate the performance across different methods, we define the true positive rate as the proportion of correctly identified non-zeros, and the false positive rate as the proportion of zeros that are incorrectly identified to be non-zeros. In addition, we illustrate the main advantage of LASER in a stock price data by modeling three way cliques that quantify conditional dependencies among three variables, conditioned on the others.

Since existing approaches are limited to estimating pairwise graphical models, to compare across different methods, we first perform some numerical studies to estimate pairwise graphical models. In his dissertation, Janofsky (2015) has also conducted some simulation

Table 3: The sum of squares error (standard error) out of 200 test samples for the two scenarios in Section 5.1.3, averaged over 200 data sets. The results are for models trained with $n = 200$ training samples with $d = 30$ covariates. Numbers are rounded to the nearest integer.

		$\rho = 0$	$\rho = 0.2$	$\rho = 0.4$	$\rho = 0.6$	$\rho = 0.8$
Scenario C1	FAMILY	236 (2)	236 (2)	235 (2)	236 (2)	240 (2)
	BT	218 (2)	217 (2)	217 (2)	218 (2)	217 (1)
	RAMP	210 (1)	209 (2)	209 (1)	210 (1)	215 (2)
	SpAM	448 (5)	452 (5)	448 (5)	431 (5)	355 (4)
	VANISH	731 (17)	852 (66)	749 (28)	745 (21)	621 (25)
	LASER	345 (7)	353 (7)	348 (8)	340 (5)	339 (4)
	ORACLE	309 (5)	314 (6)	312 (7)	305 (6)	301 (5)
Scenario C2	FAMILY	309 (5)	315 (6)	306 (5)	318 (5)	328 (5)
	BT	254 (4)	261 (5)	250 (4)	247 (5)	225 (5)
	RAMP	250 (6)	261 (7)	256 (6)	254 (6)	269 (9)
	SpAM	213 (4)	224 (6)	216 (5)	214 (5)	191 (4)
	VANISH	379 (15)	393 (17)	403 (19)	413 (22)	452 (22)
	LASER	137 (3)	143 (5)	140 (4)	141 (4)	156 (4)
	ORACLE	117 (3)	122 (4)	119 (4)	122 (4)	120 (4)

studies to assess the performance of pairwise nonparametric graphical models using the score matching approach. More specifically, we consider two different simulation settings:

1. Gaussian graphical models: we simulate $\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is generated such that $(\boldsymbol{\Sigma}^{-1})_{jk} = 0.4$ for $|k - j| = 1$, $(\boldsymbol{\Sigma}^{-1})_{jj} = 1$, and setting the other elements to zero.

2. Nonparametric graphical models: we simulate the data from the joint density

$$p(\boldsymbol{x}) \propto \exp\left(-\sum_{j=1}^{d} x_j - \sum_{j<k} \beta_{jk} x_j^2 x_k^2\right), \tag{30}$$

   where $\beta_{jk} = 1$ for $|k - j| = 1$ and $\beta_{jk} = 0$ otherwise. Noting that the conditional distribution for each variable on the others is Gaussian, we employ a Gibbs sampler to simulate data from (30).

All of the aforementioned methods involve a sparsity tuning parameter. We applied a fine grid of tuning parameter values for all methods to obtain the curves in Figure 1. For Gaussian and nonparametric graphical models, we present results for $n = 100$ and $p = 25$, and $n = 300$ and $p = 25$, respectively. Results are averaged over 200 data sets.

For Gaussian graphical models, we see from Figure 1 that the graphical lasso and the proposal of Liu et al. (2012) outperform LASER when $n = 100$. This is not surprising since both of the methods are developed based on the Gaussian assumption and Gaussian copula assumption, whereas LASER is entirely nonparametric. We essentially loses some efficiency relative to the parametric and semiparametric approaches when the parametric and semiparametric assumptions are satisfied. For nonparametric graphical models, both Friedman et al. (2008) and Liu et al. (2012) are no longer able to estimate the graph accurately since

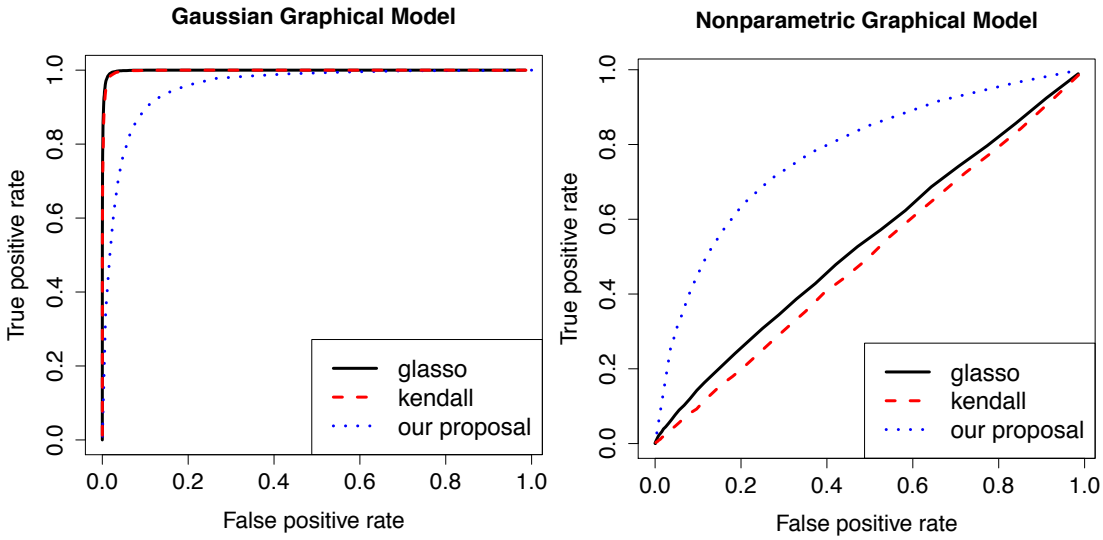

Figure 1: True and false positive rates, averaged over 200 data sets, for pairwise Gaussian and nonparametric graphical models. Left panel: Gaussian graphical models with $n = 100$ and $d = 25$. The curves are obtained by varying the tuning parameter. Right panel: nonparametric graphical models with $n = 300$ and $d = 25$.

the joint density in (30) is clearly not multivariate Gaussian. Their performances are similar to random guess even when we increase n by two-fold. LASER clearly outperforms the parametric approaches in this case. In conclusion, we sacrifice some performance in the parametric setting to gain flexibility in modeling nonparametric graphical models.

Next, we illustrate the main advantage of LASER by modeling three-way cliques that quantify conditional dependencies among three variables, conditioned on all of the other variables. To this end, we analyze the stock price data from Yahoo! Finance, which consists of daily closing prices for stocks in the S&P 500 index between January 1, 2003 and January 1, 2008. Stocks that are not consistently listed in the S&P 500 index during this time period are removed, leaving us with $n = 1258$ daily closing prices with 452 stocks. In this study, we categorize the stocks into six Global Industry Classification Standard sectors: Financials, Energy, Health Care, Information Technology, Materials, and Utilities.

The goal of our analysis is to understand the conditional dependence relationships among the six sectors. More specifically, we seek to learn the three-way conditional dependence relationships among the $d = 6$ sectors by modeling the three-way interaction terms in (17). Note that existing approaches for modeling graphical models in the literature are not able to model three-way cliques.

We first estimate a nonparametric graphical model with two-way interaction terms, corresponding to pairwise conditional dependencies. For the ease of interpretation, we pick the tuning parameter λ such that there are six edges in the estimated conditional independence graph. The results are summarized in Figure 2(a). We see that the three pairs of sectors "health care—materials", "health care—financials", and "financials—materials"

are conditionally dependent, given the other sectors. However, since we are estimating only the second order term at the second step of LASER, we cannot conclude that the three sectors financials, materials, and health care are jointly conditionally dependent.

To assess whether the three sectors are jointly conditionally dependent, we proceed to the next step of LASER for estimating three way interaction terms. We use the same tuning parameter to fit the model at the second step. The results are shown in Figure 2(b). Since the three way interaction term for health care, materials, and financials is estimated to be non-zero, we conclude that the three sectors are jointly conditionally dependent. Similar results hold for the sectors financials, materials, and information technology. Finally, LASER is terminated since there is no potential four way interaction terms to be estimated.

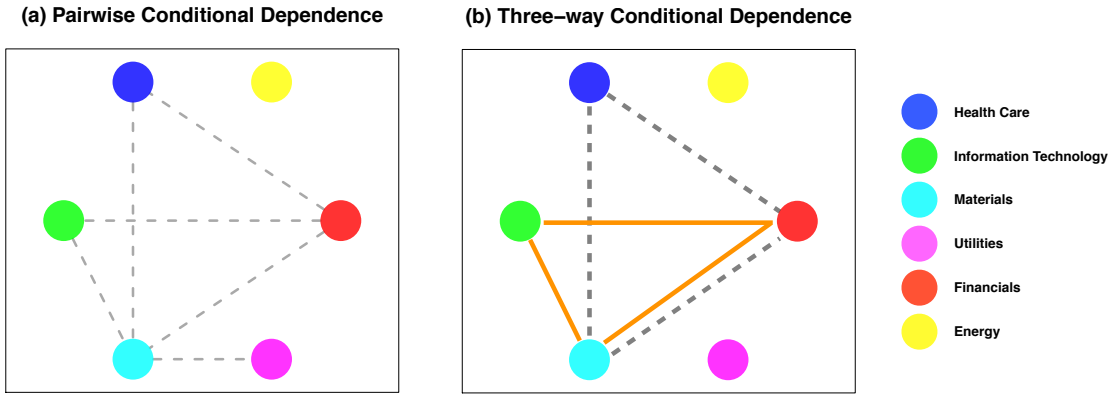

Figure 2: Estimated Conditional Dependence Graphs using the proposed method. Panel (a): Estimated pairwise conditional dependence relationships between two variables, conditioned on the others. Panel (b): Estimated three-way conditional dependence relationships among three variables, conditioned on the others.

6. Discussion

In this paper, we propose a layer-wise learning strategy (LASER) for fitting multivariate function under the SSANOVA framework. LASER provides a computationally feasible framework for estimating SSANOVA models with higher order interaction effects. In addition, we have shown that the estimators obtained from LASER is post-selection persistent. We illustrate LASER in the context of nonparametric regression and nonparametric graphical models problems. In the graphical modeling literature, most work have focused on estimating pairwise graphical models, which corresponds to estimating the conditional dependence relationships between pairs of variables. LASER provides an alternative way to estimate conditional dependence relationships among a set of more than two variables. More generally, LASER can be easily applied to other problems that involves estimating multivariate function such as the generalized nonparametric regression.

In the context of lasso regression with two-way interaction terms, Shah (2016) considered the scenario in which the main effects are useful for prediction only when certain two-

way interaction effects are present. In this case, two-stage methods that perform variable screening on the main effects at the first stage, and then include the interaction effects based on the identified main effects may fail to identify some important covariates. To address this issue, Shah (2016) proposed a backtracking algorithm. The main idea is to first select a few variables that are most correlated with the response, and then include its corresponding interaction effects into the model. This process is repeated until no more variables are included in the model.

As a reviewer pointed out, the backtracking algorithm of Shah (2016) can be modified to accommodate higher order terms in the context of nonparametric regression. The main idea is as follows: (i) select a few univariate functions that are most highly predictive of the response by fitting the additive model for the main effects; (ii) include its associated two-way interaction effects, and subsequently the higher order interaction effects into the model; (iii) repeat (i) and (ii) until no more variables are included in the model. It is out of the scope of this paper to study such an extension carefully and we leave it for future work.

Acknowledgments

We thank two reviewers and the associate editor for providing helpful comments that improved the quality of this manuscript. We thank Peter Radchenko for providing R code to fit the model proposed in Radchenko and James (2010). This research is partially supported by the NSF DMS-1811315.

Appendix A. Proof of Theorem 1

Let $\mathcal{S}^{(r-1)}$ be the support of $\widehat{f}^{(r-1)}$ and recall from (9) that conditional on the support $\mathcal{S}^{(r-1)}$, $\widehat{f}^{(r)}$ is post-selection persistent if

$$R(\widehat{f}^{(r)}) - \inf_{f \in \mathcal{F}^{(r)}} R(f) = o_P(1), \text{ where}$$

$$\mathcal{F}^{(r)} = \left\{ f : f(\boldsymbol{x}) = \sum_{J \in \delta(\mathcal{S}^{(r-1)})} \beta_J g_J(\boldsymbol{x}_J), \mathbb{E}[g_J] = 0, \|g_J\|_{\mathcal{H}_J} \le 1, \sum_{J \in \mathcal{S}^{(r-1)}} \beta_J^2 + \sum_{J \in \sigma(\mathcal{S}^{(r-1)})} |\beta_J| \le \tau \right\}.$$

For notational convenience, let $f^* = \arg\inf_{f \in \mathcal{F}^{(r)}} R(f)$. The goal is to show that $R(\widehat{f}^{(r)}) - R(f^*) = o_P(1)$.

Under the squared error loss, we define the risk $R(f)$ and empirical risk $\widehat{R}(f)$ as

$$R(f) = \mathbb{E}[(Y - f(\boldsymbol{X}))^2] \qquad \text{and} \qquad \widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2,$$

respectively. By the definition of f^*, we have $R(f^*) \le R(\widehat{f}^{(r)})$. Thus, by the triangle inequality, we have

$$\begin{aligned} 0 &\le R(\widehat{f}^{(r)}) - R(f^*) \le R(\widehat{f}^{(r)}) - \widehat{R}(\widehat{f}^{(r)}) + \widehat{R}(\widehat{f}^{(r)}) - R(f^*) \\ &\le |R(\widehat{f}^{(r)}) - \widehat{R}(\widehat{f}^{(r)})| + |\widehat{R}(f^*) - R(f^*)| \\ &\le 2 \sup_{f \in \mathcal{F}^{(r)}} |\widehat{R}(f) - R(f)|, \end{aligned} \tag{31}$$

where the third inequality holds by the definition that $\widehat{f}^{(r)}$ is the minimizer of $\widehat{R}(f)$, that is, $\widehat{R}(\widehat{f}^{(r)}) \leq \widehat{R}(f^*)$. Thus, it suffices to obtain an upper bound on $\sup_{f\in\mathcal{F}^{(r)}}|\widehat{R}(f) - R(f)|$.

Let $\mathcal{J}_r = \delta(\mathcal{S}^{(r-1)}) \cup \{\emptyset\}$. With some abuse of notation, we write $g_J(\boldsymbol{x}_J) = y$ when J is an empty set. In addition, when $f(\boldsymbol{x}) = \sum_{J\in\delta(\mathcal{S}^{(r-1)})}\beta_J g_J(\boldsymbol{x}_J)$, we also write the risk function as $R(\beta, g)$. Then, for any $f \in \mathcal{F}^{(r)}$, the risk and empirical risk can be rewritten as

$$R(\beta,g) = \sum_{J,J'\in\mathcal{J}_r}\beta_J\beta_{J'}\mathbb{E}[g_J(\boldsymbol{X}_J)g_{J'}(\boldsymbol{X}_{J'})] \text{ and } \widehat{R}(\beta,g) = \frac{1}{n}\sum_{i=1}^{n}\sum_{J,J'\in\mathcal{J}_r}\beta_J\beta_{J'}g_J(\mathbf{x}_{iJ})g_{J'}(\mathbf{x}_{iJ'}),$$

respectively. Thus, for all (β, g), we have

$$|\widehat{R}(\beta,g) - R(\beta,g)| \leq \left(\sum_{J\in J_r}|\beta_J|\right)^2 \max_{J\in\mathcal{J}_r}\sup_{g_J\in\mathcal{H}_J, g_{J'}\in\mathcal{H}_{J'}}(\mathbb{E}_n - \mathbb{E})[g_J g_{J'}],$$

where $\mathbb{E}_n[g_J g_{J'}] = n^{-1}\sum_{i=1}^n g_J(\mathbf{x}_{iJ})g_{J'}(\mathbf{x}_{iJ'})$. Let s_{r-1} be the cardinality of the set $\mathcal{S}^{(r-1)}$. We have

$$\begin{aligned}\left(\sum_{J\in J_r}|\beta_J|\right)^2 &\leq 2\left(\sum_{J\in\mathcal{S}^{(r-1)}}|\beta_J|\right)^2 + 2\left(\sum_{J\in\sigma(\mathcal{S}^{(r-1)})}|\beta_J|\right)^2 \\ &\leq 2s_{r-1}\left(\sum_{J\in\mathcal{S}^{(r-1)}}\beta_J^2\right)^2 + 2\tau^2 \leq (2s_{r-1}+2)\tau^2,\end{aligned} \tag{32}$$

where we use the inequality $2ab \leq a^2 + b^2$ for any $a > 0$ and $b > 0$, and the constrained in the function class.

To obtain an upper bound, we begin with some notation. For a function class $\mathcal{F}$ and for any measure Q, the L^∞ bracketing number $\mathcal{N}_{[]}(\mathcal{F}, L^\infty(Q), \epsilon)$ is defined as the smallest number of pairs $B = \{(l_1, u_1), \ldots, (l_k, u_k)\}$ such that $\|u_j - l_j\|_\infty \leq \epsilon$ for $1 \leq j \leq k$, and such that for every $f \in \mathcal{F}$, there exists $(l, u) \in B$ such that $l \leq f \leq u$. Define the function class

$$\mathcal{W} = \{g_J g_{J'} \mid g_J \in \mathcal{H}_J, g_{J'} \in \mathcal{H}_{J'}, \|g_J\|_{\mathcal{H}_J} \leq 1, \|g_{J'}\|_{\mathcal{H}_{J'}} \leq 1, J, J' \in \mathcal{J}_r\}. \tag{33}$$

By Lemma 4, we have

$$\log\mathcal{N}_{[]}(\mathcal{W}, L^\infty(Q), \epsilon) \leq C(r\log d + \epsilon^{-r/m}), \tag{34}$$

where $C > 0$ is some constant. By Corollary 19.35 of Van der Vaart (2000) and (34), we obtain

$$\begin{aligned}\mathbb{E}\left(\max_{J\in\mathcal{J}_r}\sup_{g_J\in\mathcal{H}_J, g_{J'}\in\mathcal{H}_{J'}}(\mathbb{E}_n - \mathbb{E})[g_J g_{J'}]\right) &\leq \frac{C}{\sqrt{n}}\int_0^C \sqrt{\log\mathcal{N}_{[]}(\mathcal{W}, L^\infty(Q), \epsilon)}d\epsilon \\ &\leq \frac{C}{\sqrt{n}}\int_0^C \sqrt{r\log d + \epsilon^{-r/m}}d\epsilon \\ &\leq \frac{C}{\sqrt{n}}\int_0^C \sqrt{r\log d} + \epsilon^{-r/2m}d\epsilon \\ &\leq C\sqrt{\frac{r\log d}{n}},\end{aligned} \tag{35}$$

where the third inequality follows from the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, the last inequality follows from the assumption that $r < 2m$, and C is a constant that may vary line to line.

By an application of Markov's inequality, we have

$$\max_{J \in \mathcal{J}_r} \sup_{g_J \in \mathcal{H}_J, g_{J'} \in \mathcal{H}_{J'}} (\mathbb{E}_n - \mathbb{E})[g_J g_{J'}] = O_P\left(\sqrt{\frac{r \log d}{n}}\right). \tag{36}$$

Combining the above with (32), we have that for all (β, g),

$$|\widehat{R}(\beta, g) - R(\beta, g)| = O_P\left(\tau^2 \sqrt{\frac{r s_{r-1}^2 \log d}{n}}\right),$$

as desired.

Appendix B. Proof of Proposition 1

The proof of Proposition 1 is similar to that of the proof of Theorem 2 in Ravikumar et al. (2009). The main difference is that we have additional bivariate interaction terms in the true underlying model. Recall that $\mathcal{S}^{(1)}$ is the support of the estimator obtained from solving (14) with $r = 1$. Recall from Section 3.1 that we approximate the main effects f_j by its kth order basis expansion, i.e., $\widetilde{f}_j = \sum_{l=1}^{k} \theta_j^l \phi_{jl}(x_j) = \boldsymbol{\theta}_j^T \boldsymbol{\phi}_j(x_j)$. Let $\boldsymbol{\theta}_j^*$ be the underlying coefficients corresponding to the kth order basis expansion of the main effects. To simplify the notation, we let $S = \bar{\mathcal{S}}^{(1)}$.

Let $\boldsymbol{\Phi}_S$ be the $n \times k|S|$ matrix with rows $\boldsymbol{\phi}_S(\mathbf{x}_1), \ldots, \boldsymbol{\phi}_S(\mathbf{x}_n)$, where $\boldsymbol{\phi}_S(\cdot)$ is obtained by concatenating $\boldsymbol{\phi}_j(\cdot)$ for all $j \in S$. Similarly, $\boldsymbol{\theta}_S^*$ is obtained by concatenating $\boldsymbol{\theta}_j^*$ for all $j \in S$. Recall that we denote the projection operator $P_{\bar{\mathcal{S}}^{(2)}} f = \sum_{(s,t) \in \bar{\mathcal{S}}^{(2)}} f_{st}$. We now consider a second order Taylor expansion of f_{st} at $(1/2, 1/2)$ for each $(s,t) \in \bar{\mathcal{S}}^{(2)}$:

$$\begin{aligned} f_{st}(x_s, x_t) =& f_{st}(0.5, 0.5) + \sum_{j \in \{s,t\}} \left(\partial_{x_j} f_{st}(0.5, 0.5)(x_j - 0.5) + \frac{1}{2} \partial_{x_j}^2 f_{st}(0.5, 0.5)(x_j - 0.5)^2 \right) \\ &+ \partial_{x_s x_t}^2 f_{st}(0.5 + \eta(x_s - 0.5), 0.5 + \eta(x_t - 0.5))(x_s - 0.5)(x_t - 0.5), \end{aligned} \tag{37}$$

where $\eta \in [0, 1]$. Since $\boldsymbol{\Phi}_S$ is the design matrix generated from B-spline polynomials, we can represent the vector $\mathbf{u} := (P_{\bar{\mathcal{S}}^{(2)}} f(\mathbf{x}_1), \ldots, P_{\bar{\mathcal{S}}^{(2)}} f(\mathbf{x}_n))^T$ as $\mathbf{u} = \boldsymbol{\Phi}_S \boldsymbol{\gamma}_S^* + \Delta$, where the ith entry of $\boldsymbol{\Phi}_S \boldsymbol{\gamma}_S^*$ represents the leading terms in (37) and $\boldsymbol{\gamma}_S^*$ is the corresponding basis coefficients vector:

$$[\boldsymbol{\Phi}_S \boldsymbol{\gamma}_S^*]_i = \textstyle\sum_{(s,t) \in \bar{\mathcal{S}}^{(2)}} \left[f_{st}(0.5, 0.5) + \sum_{j \in \{s,t\}} (\partial_{x_j} f_{st}(0.5, 0.5)(x_{ij} - 0.5) + \frac{1}{2} \partial_{x_j}^2 f_{st}(0.5, 0.5)(x_{ij} - 0.5)^2) \right].$$

The remainder Δ corresponds to the last term in (37), i.e., for each $i = 1, \ldots, n$,

$$\Delta_i = \textstyle\sum_{(s,t) \in \bar{\mathcal{S}}^{(2)}} \partial_{x_s x_t}^2 f_{st}(0.5 + \eta(x_{is} - 0.5), 0.5 + \eta(x_{it} - 0.5))(x_{is} - 0.5)(x_{it} - 0.5).$$

Thus, $\|\Delta\|_\infty \leq q_n^*$, where $q_n^* := \| \sum_{(s,t) \in \bar{\mathcal{S}}^{(2)}} \partial_{x_s x_t}^2 f_{st} \|_\infty$.

Denote $\boldsymbol{\Sigma}_{SS} = \frac{1}{n} \boldsymbol{\Phi}_S^T \boldsymbol{\Phi}_S$. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, and let $\mathbf{v} = \mathbf{y} - \boldsymbol{\Phi}_S(\boldsymbol{\theta}_S^* + \boldsymbol{\gamma}_S^*) - \Delta - \boldsymbol{\epsilon}$ denote the error due to finite truncation of the orthogonal basis. Compared to the proof of

Theorem 2 in Ravikumar et al. (2009), we have an additional term Δ. Therefore, following the same proof, we have

$$\|\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\theta}^*_S - \boldsymbol{\gamma}^*_S\|_\infty \leq \left\|\boldsymbol{\Sigma}^{-1}_{SS}\left(\frac{1}{n}\boldsymbol{\Phi}^T_S\boldsymbol{\epsilon}\right)\right\|_\infty + \left\|\boldsymbol{\Sigma}^{-1}_{SS}\left(\frac{1}{n}\boldsymbol{\Phi}^T_S(\mathbf{v}+\Delta)\right)\right\|_\infty + \lambda\|\boldsymbol{\Sigma}^{-1}_{SS}\widehat{\mathbf{g}}_S\|_\infty, \quad (38)$$

where $\boldsymbol{\theta}_S$ and $\widehat{\mathbf{g}}_S$ are subvectors of the estimator in (14) for $r=1$ and the gradient of the penalty in (14) for $r=1$, respectively. If we can show that

$$\|\widehat{\boldsymbol{\theta}}_S - \boldsymbol{\beta}^*_S\|_\infty < \frac{1}{2}\min_{j\in S}\|\boldsymbol{\theta}^*_j + \boldsymbol{\gamma}^*_j\|_\infty = \frac{1}{2}\min_{j\in S}\|\boldsymbol{\beta}^*_j\|_\infty = \rho^*_n,$$

then $\|\widehat{\boldsymbol{\theta}}_j\|_\infty > 0$ for each $j \in S$, i.e., $\mathcal{S}^{(1)} \supseteq S$. Therefore, it suffices to show that the right hand side of (38) is smaller than ρ^*_n. Compared to the right hand side of Equation (83) in Ravikumar et al. (2009), we only have an additional term $\|\boldsymbol{\Sigma}^{-1}_{SS}(\frac{1}{n}\boldsymbol{\Phi}^T_S\Delta)\|_\infty$ in (38).

Since $\|\Delta\|_\infty \leq q^*_n$, we can bound $\|\boldsymbol{\Sigma}^{-1}_{SS}(\frac{1}{n}\boldsymbol{\Phi}^T_S\Delta)\|_\infty$ following the same procedure as $\|\boldsymbol{\Sigma}^{-1}_{SS}(\frac{1}{n}\boldsymbol{\Phi}^T_S\mathbf{v})\|_\infty$. According to Equations (99) and (111) in Ravikumar et al. (2009), if $\sqrt{|\bar{\mathcal{S}}^{(1)}|}kq^*_n/\rho^*_n = o(1)$ and $\sqrt{k}q^*_n/\lambda = o(1)$, terms related to Δ will be dominated by terms related to $\boldsymbol{\epsilon}$ and $\mathbf{v}$, and the remaining proof in Ravikumar et al. (2009) follows through.

Appendix C. Proof of Lemma 1

Recall from Section 4.3 that we define $\mathbf{r}'(\boldsymbol{x})$ to be the element-wise differentiation of the vector $\mathbf{r}(\boldsymbol{x})$. Also, recall from (22) that the modified Fisher divergence is defined as

$$\frac{1}{2}\int_{[0,1]^d} p(\boldsymbol{x})\,\|\mathbf{r}(\boldsymbol{x})\circ[\nabla\log p(\boldsymbol{x}) - \nabla\log q(\boldsymbol{x})]\|_2^2\,d\boldsymbol{x} = T_1 + T_2 + T_3, \text{ where} \quad (39)$$

$$\begin{aligned}
T_1 &= \frac{1}{2}\int_{[0,1]^d} p(\boldsymbol{x})\|\mathbf{r}(\boldsymbol{x})\circ\nabla\log q(\boldsymbol{x})\|_2^2 d\boldsymbol{x},\\
T_2 &= -\int_{[0,1]^d} p(\boldsymbol{x})\left(\mathbf{r}(\boldsymbol{x})\circ\nabla\log p(\boldsymbol{x})\right)^T\left(\mathbf{r}(\boldsymbol{x})\circ\nabla\log q(\boldsymbol{x})\right)d\boldsymbol{x},\\
T_3 &= \frac{1}{2}\int_{[0,1]^d} p(\boldsymbol{x})\|\mathbf{r}(\boldsymbol{x})\circ\nabla\log p(\boldsymbol{x})\|_2^2 d\boldsymbol{x}.
\end{aligned}$$

By some algebraic manipulation, we have

$$\begin{aligned}
T_2 &= -\int_{[0,1]^d} p(\boldsymbol{x})\left(\mathbf{r}(\boldsymbol{x})\circ\nabla\log p(\boldsymbol{x})\right)^T\left(\mathbf{r}(\boldsymbol{x})\circ\nabla\log q(\boldsymbol{x})\right)d\boldsymbol{x}\\
&= -\int_{[0,1]^d}\left(\nabla p(\boldsymbol{x})\right)^T\left(\mathbf{r}(\boldsymbol{x})\circ\mathbf{r}(\boldsymbol{x})\right)\circ\nabla\log q(\boldsymbol{x})d\boldsymbol{x}\\
&= \sum_{j=1}^d\left[-p(\boldsymbol{x})\nabla_j\log q(\boldsymbol{x})r_j^2(x_j)\big|_0^1 + \int_{[0,1]^d} p(\boldsymbol{x})\frac{\partial}{\partial x_j}\left\{r_j^2(x_j)\nabla_j\log q(\boldsymbol{x})\right\}d\boldsymbol{x}\right]\\
&= \int_{[0,1]^d} p(\boldsymbol{x})\left[\left(2\mathbf{r}(\boldsymbol{x})\circ\mathbf{r}'(\boldsymbol{x})\right)^T\nabla\log q(\boldsymbol{x}) + \left(\mathbf{r}(\boldsymbol{x})\circ\mathbf{r}(\boldsymbol{x})\right)^T\nabla^2\log q(\boldsymbol{x})\right]d\boldsymbol{x}
\end{aligned}$$

where the second equality uses the fact that $\nabla \log p(\boldsymbol{x}) = \nabla p(\boldsymbol{x})/p(\boldsymbol{x})$, the third equality holds using integration by parts, and the last equality holds by the assumption that

$$\lim_{x_j \to 0} p(\boldsymbol{x}) \cdot \nabla_j \log q(\boldsymbol{x}) r_j^2(x_j) \to 0 \qquad \text{and} \qquad \lim_{x_j \to 1} p(\boldsymbol{x}) \cdot \nabla_j \log q(\boldsymbol{x}) r_j^2(x_j) \to 0.$$

Since T_3 is independent of $q(\boldsymbol{x})$, we obtain

$$F(p,q) = \int_{[0,1]^d} p(\boldsymbol{x}) S(\boldsymbol{x}, q) d\boldsymbol{x} + C$$

with

$$S(\boldsymbol{x}, q) = 2\left(\mathbf{r}(\boldsymbol{x}) \circ \mathbf{r}'(\boldsymbol{x})\right)^T \nabla \log q(\boldsymbol{x}) + (\mathbf{r}(\boldsymbol{x}) \circ \mathbf{r}(\boldsymbol{x}))^T \nabla^2 \log q(\boldsymbol{x}) + \frac{1}{2}\|\mathbf{r}(\boldsymbol{x}) \circ \nabla \log q(\boldsymbol{x})\|_2^2,$$

and C is a constant that does not depend on $q(\boldsymbol{x})$.

Appendix D. Proof of Theorem 2

Recall that we define the risk function as the Fisher's divergence between two densities

$$R(p,q) = \frac{1}{2} \int_{[0,1]^d} p(\boldsymbol{x}) \left\| \mathbf{r}(\boldsymbol{x}) \circ \nabla \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right\|_2^2 d\boldsymbol{x}.$$

For notational convenient, let $q^* = \arg\inf_{q \in \mathcal{Q}^{(r)}} R(p,q)$ be the oracle estimator and let $\widehat{p}^{(r)} \propto \exp(-\widehat{f}^{(r)})$. The goal is to show that $R(p, \widehat{p}^{(r)}) - R(p, q^*) = o_P(1)$. By Lemma 1, the difference between the two risk functions can be rewritten as

$$R(p, \widehat{p}^{(r)}) - R(p, q^*) = \mathbb{E}[S(\boldsymbol{X}, \widehat{p}^{(r)})] - \mathbb{E}[S(\boldsymbol{X}, q^*)],$$

since the constant term in Lemma 1 is independent of q. By an argument similar to (31), we have

$$\begin{aligned}
0 &\le \mathbb{E}[S(\boldsymbol{X}, \widehat{p}^{(r)})] - \mathbb{E}[S(\boldsymbol{X}, q^*)] \\
&\le \left| \mathbb{E}[S(\boldsymbol{X}, \widehat{p}^{(r)})] - \frac{1}{n}\sum_{i=1}^n S(\mathbf{x}_i, \widehat{p}^{(r)}) \right| + \left| \frac{1}{n}\sum_{i=1}^n S(\mathbf{x}_i, q^*) - \mathbb{E}[S(\boldsymbol{X}, q^*)] \right| \\
&\le 2 \sup_{q \in \mathcal{Q}^{(r)}} \left| \frac{1}{n}\sum_{i=1}^n S(\mathbf{x}_i, q) - \mathbb{E}[S(\boldsymbol{X}, q)] \right|,
\end{aligned}$$

where the second inequality holds from the fact that $\widehat{p}^{(r)}$ is a minimizer of $n^{-1}\sum_{i=1}^n S(\mathbf{x}_i, q)$. Thus, it suffices to establish the rate of convergence for the above quantity.

Let $\mathcal{J}_r = \delta(\mathcal{S}^{(r-1)}) \cup \{\emptyset\}$. Recall that for any $f \in \mathcal{F}^{(r)}$, we have $f(\boldsymbol{x}) = \sum_{J \in \mathcal{J}_r} \beta_J g_J(\boldsymbol{x}_J)$. Then, $S(\boldsymbol{x}, q)$ can be rewritten as

$$\begin{aligned}
S(\boldsymbol{x}, q) &= 2(\mathbf{r}(\boldsymbol{x}) \circ \mathbf{r}(\boldsymbol{x}))^T \nabla \log q(\boldsymbol{x}) + (\mathbf{r}(\boldsymbol{x}) \circ \mathbf{r}(\boldsymbol{x}))^T \nabla^2 \log q(\boldsymbol{x}) + \frac{1}{2}\|\mathbf{r}(\boldsymbol{x}) \circ \nabla \log q(\boldsymbol{x})\|_2^2 \\
&= -2\sum_{j=1}^d \sum_{J \in \mathcal{J}_r} r_j r_j' \beta_J g_J^{(j)} - \sum_{j=1}^d \sum_{J \in \mathcal{J}_r} r_j^2 \beta_J g_J^{(jj)} + \frac{1}{2}\sum_{j=1}^d \sum_{J \in \mathcal{J}_r} \sum_{J' \in \mathcal{J}_r} r_j^2 \beta_J \beta_{J'} g_J^{(j)} g_{J'}^{(j)},
\end{aligned}$$

where we denote $g_J^{(j)}$ and $g_J^{(jj)}$ as the first and second order derivatives of g_J with respect to X_j, respectively. For notational convenience, we suppress the dependencies of $\boldsymbol{x}$ in $r_j(\cdot)$ and $g_J(\cdot)$, and use the fact that $g_J^{(j)} = 0$ if $j \notin J$.

Thus, for any $q \in \mathcal{Q}^{(r)}$, we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} S(\mathbf{x}_i, q) - \mathbb{E}[S(\boldsymbol{X}, q)] \right| \leq 2I_1 + I_2 + I_3,$$

where the three terms are

$$I_1 = \max_{J \in \mathcal{J}_r} \sup_{g_J \in \mathcal{H}_J} \left| \sum_{j=1}^{d} \sum_{J \in \mathcal{J}_r} \beta_J (\mathbb{E}_n - \mathbb{E})[r_j r_j' g_J^{(j)}] \right|,$$

$$I_2 = \max_{J \in \mathcal{J}_r} \sup_{g_J \in \mathcal{H}_J} \left| \sum_{j=1}^{d} \sum_{J \in \mathcal{J}_r} \beta_J (\mathbb{E}_n - \mathbb{E})[r_j^2 g_J^{(jj)}] \right|,$$

$$I_3 = \max_{J, J' \in \mathcal{J}_r} \sup_{g_J \in \mathcal{H}_J, g_{J'} \in \mathcal{H}_{J'}} \left| \sum_{j=1}^{d} \sum_{J \in \mathcal{J}_r} \sum_{J' \in \mathcal{J}_r} \beta_J \beta_{J'} (\mathbb{E}_n - \mathbb{E})[r_j^2 g_J^{(j)} g_{J'}^{(j)}] \right|.$$

We now obtain upper bounds for I_1, I_2, and I_3 separately. To this end, we define three function classes

$$\mathcal{W}_1 = \left\{ r_j r_j' g_J^{(j)} \mid \|g_J\|_{\mathcal{H}_J} \leq 1, J \in \mathcal{J}_r, j \in [d] \right\},$$

$$\mathcal{W}_2 = \left\{ r_j^2 g_J^{(jj)} \mid \|g_J\|_{\mathcal{H}_J} \leq 1, J \in \mathcal{J}_r, j \in [d] \right\},$$

$$\mathcal{W}_3 = \left\{ r_j^2 g_J^{(j)} g_{J'}^{(j)} \mid \|g_J\|_{\mathcal{H}_J} \leq 1, \|g_{J'}\|_{\mathcal{H}_{J'}} \leq 1, J, J' \in \mathcal{J}_r, j \in [d] \right\}.$$

Upper bound for I_1: We first note that

$$I_1 \leq r \left(\sum_{J \in \mathcal{J}_r} |\beta_J| \right) \max_{j \in [d]} \max_{J \in \mathcal{J}_r} \sup_{g_J \in \mathcal{H}_J} \left| (\mathbb{E}_n - \mathbb{E})[r_j r_j' g_J^{(j)}] \right|,$$

since $\sum_{j=1}^{d} g_J^{(j)}$ can be expressed as the sum of a maximum of r terms for any $J \in \mathcal{J}_r$.

By Lemma 5, the bracketing number for the function class $\mathcal{W}_1$ is

$$\log \mathcal{N}_{[]}(\mathcal{W}_1, L^\infty(Q), \epsilon) \leq C(r \log d + \epsilon^{-r/(m-1)}).$$

Following the arguments in (35), we have

$$\mathbb{E}\left(\max_{j \in [d]} \max_{J \in \mathcal{J}_r} \sup_{g_J \in \mathcal{H}_J} (\mathbb{E}_n - \mathbb{E})[r_j r_j' g_J^{(j)}] \right) \leq C \sqrt{\frac{r \log d}{n}}.$$

Similar to the proof of (36), we obtain

$$\max_{j \in [d]} \max_{J \in \mathcal{J}_r} \sup_{g_J \in \mathcal{H}_J} (\mathbb{E}_n - \mathbb{E})[r_j r_j' g_J^{(j)}] = O_P\left(\sqrt{\frac{r \log d}{n}} \right).$$

Similar to (32), we have

$$\sum_{J\in\mathcal{J}_r} |\beta_J| \le \sqrt{2s_{r-1}+2}\tau,$$

which implies that $I_1 = O_P(\tau\sqrt{(r^3 s_{r-1}\log d)/n})$.

Upper bound for I_2: We have

$$I_2 \le r\left(\sum_{J\in\mathcal{J}_r}|\beta_J|\right)\max_{j\in[d]}\max_{J\in\mathcal{J}_r}\sup_{g_J\in\mathcal{H}_J}\left|(\mathbb{E}_n-\mathbb{E})[r_j^2 g_J^{(jj)}]\right|.$$

By Lemma 5, the bracketing number for the function class $\mathcal{W}_2$ is

$$\log\mathcal{N}_{[]}(\mathcal{W}_2, L^\infty(Q), \epsilon) \le C(r\log d + \epsilon^{-r/(m-2)}).$$

Following the arguments for the upper bound on I_1, we have

$$\max_{j\in[d]}\max_{J\in\mathcal{J}_r}\sup_{g_J\in\mathcal{H}_J}(\mathbb{E}_n-\mathbb{E})[r_j^2 g_J^{(jj)}] = O_P\left(\sqrt{\frac{r\log d}{n}}\right)$$

and that $I_2 = O_P(\tau\sqrt{(r^3 s_{r-1}\log d)/n})$.

Upper bound for I_3: To obtain an upper bound for I_3, we have

$$I_3 \le r\left(\sum_{J\in\mathcal{J}_r}|\beta_J|\right)^2\max_{j\in[d]}\max_{J\in\mathcal{J}_r}\sup_{g_J\in\mathcal{H}_J}\left|(\mathbb{E}_n-\mathbb{E})[r_j^2 g_J^{(jj)}]\right|.$$

Similar to (32), we have

$$\left(\sum_{J\in\mathcal{J}_r}|\beta_J|\right)^2 \le (2s_{r-1}+2)\tau^2.$$

By an application of Lemma 5, the bracketing number for the function class $\mathcal{W}_3$ is

$$\log\mathcal{N}_{[]}(\mathcal{W}_3, L^\infty(Q), \epsilon) \le C(r\log d + \epsilon^{-r/(m-1)}).$$

Thus, following the same arguments for the upper bound on I_1, we have

$$\max_{j\in[d]}\max_{J\in\mathcal{J}_r}\sup_{g_J\in\mathcal{H}_J}\left|(\mathbb{E}_n-\mathbb{E})[r_j^2 g_J^{(jj)}]\right| = O_P\left(\sqrt{\frac{r\log d}{n}}\right),$$

which implies that $I_3 = O_P(\tau^2\sqrt{(r^3 s_{r-1}^2\log d/n)})$.

Combining the upper bounds on I_1, I_2, and I_3, we have

$$R(p,\widehat{p}^{(r)}) - R(p,q^*) = O_P\left(\tau^2\sqrt{\frac{r^3 s_{r-1}^2\log d}{n}}\right),$$

as desired.

Appendix E. Auxiliary Results on Bracketing Number

In this section, we provide some technical results on bracketing number for some function classes. Lemma 2 provides the bracketing number of the Sobolev space. Lemma 3 provides an upper bound on the bracketing number for function classes generated from the product and addition of two function classes. With applications of Lemmas 2 and 3, we obtain upper bounds on bracketing number for function classes arise in the proof of Theorems 1 and 2.

Lemma 2 (Corollary 1 in Nickl and Pötscher, 2007)**.** *Let Ω be a bounded, convex subset of $\mathbb{R}^r$ with non-empty interior and $Q(\Omega) < \infty$ for some measure Q. Denote $\mathbb{B}_m(\Omega) = \{f \in \mathcal{H}^m(\Omega) \mid \|f\|_{\mathcal{H}^m} \leq 1\}$. There exists a constant K such that*

$$\log \mathcal{N}_{[]}(\mathbb{B}_m(\Omega), L_\infty(Q), \epsilon) \leq K \cdot \epsilon^{-r/m}.$$

Lemma 3 (Lemma 9.24 in Kosorok, 2007)**.** *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two function classes. Define $\|\mathcal{F}_\ell\|_\infty = \sup_{f \in \mathcal{F}_\ell} \|f\|_\infty$ for $\ell = 1, 2$ and $U = \|\mathcal{F}_1\|_\infty \vee \|\mathcal{F}_2\|_\infty$. For the function classes $\mathcal{F}_+ = \{f_1 + f_2 \mid f_1 \in \mathcal{F}_1,\ f_2 \in \mathcal{F}_2\}$ and $\mathcal{F}_\times = \{f_1 f_2 \mid f_1 \in \mathcal{F}_1,\ f_2 \in \mathcal{F}_2\}$, we have for any $\epsilon \in (0, 1)$,*

$$\mathcal{N}_{[]}(\mathcal{F}_+, L_\infty(Q), \epsilon) \leq \mathcal{N}_{[]}(\mathcal{F}_1, L_\infty(Q), \epsilon) \cdot \mathcal{N}_{[]}(\mathcal{F}_2, L_\infty(Q), \epsilon),$$
$$\mathcal{N}_{[]}(\mathcal{F}_\times, L_\infty(Q), \epsilon) \leq \mathcal{N}_{[]}(\mathcal{F}_1, L_\infty(Q), \epsilon/U) \cdot \mathcal{N}_{[]}(\mathcal{F}_2, L_\infty(Q), \epsilon/U).$$

Together with Lemmas 2 and 3, we can now obtain the bracketing number

$$\mathcal{W} = \left\{ g_J g_{J'} \mid g_J \in \mathcal{H}_J, g_{J'} \in \mathcal{H}_{J'}, \|g_J\|_{\mathcal{H}_J} \leq 1, \|g_{J'}\|_{\mathcal{H}_{J'}} \leq 1, J, J' \in \mathcal{J}_r \right\}$$

as defined in (33). The result is summarized in the following lemma.

Lemma 4. *For any measure Q and $\epsilon > 0$, there exists a constant $C > 0$ such that*

$$\log \mathcal{N}_{[]}(\mathcal{W}, L^\infty(Q), \epsilon) \leq C(r \log d + \epsilon^{-r/m}).$$

Proof. Let J be an index set with $|J| \leq r$. For a given J, By Lemma 2, the covering number for the function class $\mathcal{W}^{(J)} = \{g_J \mid g_J \in \mathcal{H}_J, \|g_J\|_{\mathcal{H}_J} \leq 1\}$ is

$$\log \mathcal{N}_{[]}(\mathcal{W}^{(J)}, L_\infty(Q), \epsilon) \leq K \cdot \epsilon^{-r/m}.$$

By an application of Lemma 3, there exists a constant $C > 0$ such that

$$\begin{aligned} \log \mathcal{N}_{[]}(\mathcal{W}, L_\infty(Q), \epsilon) &\leq \log \left(\binom{d}{r} \cdot [\mathcal{N}_{[]}(\mathcal{W}^{(J)}, L_\infty(Q), \epsilon)]^2 \right) \\ &\leq r \log d + 2K\epsilon^{-r/m} \leq C(r \log d + \epsilon^{-r/m}). \end{aligned}$$

The term $r \log d$ is an upper bound of the log cardinality of $\mathcal{J}_r$.

□

Next, we obtain the bracketing number for the following function classes:

$$\mathcal{W}_1 = \left\{ r_j r'_j g_J^{(j)} \mid \|g_J\|_{\mathcal{H}_J} \leq 1, J \in \mathcal{J}_r, j \in [d] \right\},$$
$$\mathcal{W}_2 = \left\{ r_j^2 g_J^{(jj)} \mid \|g_J\|_{\mathcal{H}_J} \leq 1, J \in \mathcal{J}_r, j \in [d] \right\},$$
$$\mathcal{W}_3 = \left\{ r_j^2 g_J^{(j)} g_{J'}^{(j)} \mid \|g_J\|_{\mathcal{H}_J} \leq 1, \|g_{J'}\|_{\mathcal{H}_{J'}} \leq 1, J, J' \in \mathcal{J}_r, j \in [d] \right\}.$$

Lemma 5. *For any measure Q and $\epsilon > 0$, there exists a constant $C > 0$ such that*

$$\log \mathcal{N}_{[]}(\mathcal{W}_1, L_\infty(Q), \epsilon) \leq C \cdot (r \log d + \epsilon^{-r/(m-1)}),$$
$$\log \mathcal{N}_{[]}(\mathcal{W}_2, L_\infty(Q), \epsilon) \leq C \cdot (r \log d + \epsilon^{-r/(m-2)}),$$
$$\log \mathcal{N}_{[]}(\mathcal{W}_3, L_\infty(Q), \epsilon) \leq C \cdot (r \log d + \epsilon^{-r/(m-1)}).$$

Proof. For a given index set J with $|J| \leq r$, we have $\|g_J\|_{\mathcal{H}_J} \leq 1$ by the definition of the function classes $\mathcal{W}_1$, $\mathcal{W}_2$, and $\mathcal{W}_3$. This implies that $\|g_J^{(j)}\|_{\mathcal{H}_J} \leq 1$, and that $g_J^{(j)} \in \mathbb{B}_{m-1}$. By Lemma 2, the log bracketing number for the function class $\{g_J^{(j)} \mid g_J^{(j)} \in \mathbb{B}_{m-1}\}$ is upper bounded by $K\epsilon^{-r/(m-1)}$. Since r_j and r_j' are both fixed and bounded functions on $[0, 1]$, by Lemma 3, we have

$$\log \mathcal{N}_{[]}(\mathcal{W}_1, L_\infty(Q), \epsilon) \leq C \cdot (r \log d + \epsilon^{-r/(m-1)}),$$

where the term $r \log d$ arises from the upper bound on the log cardinality of $\mathcal{J}_r$.

Similarly, since $\|g_J\|_{\mathcal{H}_J} \leq 1$, we have $\|g_J^{(jj)}\|_{\mathcal{H}_J} \leq 1$, and that $g_J^{(jj)} \in \mathbb{B}_{m-2}$. By Lemma 2, for a given index set J with $|J| \leq r$, the log bracketing number for the function class $\{g_J^{(j)} \mid g_J^{(j)} \in \mathbb{B}_{m-2}\}$ is upper bounded by $K\epsilon^{-r/(m-2)}$. Following the same argument, we have

$$\log \mathcal{N}_{[]}(\mathcal{W}_2, L_\infty(Q), \epsilon) \leq C \cdot (r \log d + \epsilon^{-r/(m-2)}).$$

For two given index set J and J' with $|J| \leq r$ and $|J'| \leq r$, the function class $\{g_J^{(j)} g_{J'}^{(j)} \mid g_J^{(j)}, g_{J'}^{(j)} \in \mathbb{B}_{m-1}\}$ is upper bounded by $2K\epsilon^{-r/(m-1)}$ by an application of Lemmas 2 and 3. Similarly, since r_j is fixed and bounded function on $[0, 1]$, by Lemma 3, there exists a constant $C > 0$ such that

$$\log \mathcal{N}_{[]}(\mathcal{W}_3, L_\infty(Q), \epsilon) \leq C \cdot (r \log d + \epsilon^{-r/(m-1)}).$$

□

Appendix F. Results on Model Selection for Numerical Studies in Section 5.1

As noted in Sections 3.2 and 3.3, it is vital for `LASER` to include all of the active main effects and the lower-order interaction effects to achieve persistency. In this section, we assess the model selection performance of `LASER` by reporting the true and false positive rates. We consider the simulation studies in Sections 5.1.1 and 5.1.2. The results, averaged over 100 replications, are summarized in Tables 4—6. Note that the results for `FAMILY` are omitted since `FAMILY` is outperformed by both `BT` and `RAMP` in terms of sum of squared error.

In the low-dimensional setting, we see from Table 4 that `LASER` has true positive rate of one for all scenarios. This illustrates that `LASER` is persistent for Scenarios A1—A3 since it is able to identify all of the active main effects before estimating the second-order interaction terms. We notice that `BT` tends to have high false positive rate on all scenarios, and that `RAMP` has a low true positive rate for Scenarios A2 and A3 when the active variables are non-linear. `LASER` has high TPRs for all scenarios for the interaction effects. Moreover, it

achieves a TPR of approximately one for Scenario A3 when other methods fail. `SPAM` has NAs on the interaction effects since it estimates only the main effects. Similar results are observed for the high-dimensional setting in Tables 5 and 6. The results for `VANISH` are omitted for the high-dimensional setting due to computational reasons.

Table 4: The true and false positive rates (TPR and FPR) for the main effects and the second-order interaction effects for three different scenarios in Section 5.1.1, averaged over 100 data sets. The results are for models trained with n training samples and $d = 30$ variables. Most of the standard errors are approximately zero and are omitted.

		$n = 200$				$n = 400$			
		Main Effects		Interaction Effects		Main Effects		Interaction Effects	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Scenario A1	`BT`	1	0.861	0.977	0.013	1	0.922	1	0.017
	`RAMP`	1	0.004	0.663	0.001	1	0.005	0.690	0.001
	`SPAM`	1	0.077	NA	NA	1	0.039	NA	NA
	`VANISH`	1	0.001	0.537	0	1	0.001	0.759	0
	`LASER`	1	0.077	0.940	0.008	1	0.039	0.990	0.006
Scenario A2	`BT`	0.903	0.546	0.497	0.015	0.920	0.620	0.480	0.019
	`RAMP`	0.8	0.043	0.227	0.002	0.830	0.012	0.493	0.005
	`SPAM`	1	0.039	NA	NA	1	0.035	NA	NA
	`VANISH`	1	0.001	1	0.001	1	0.001	1	0
	`LASER`	1	0.039	1	0.016	1	0.035	1	0.013
Scenario A3	`BT`	0.677	0.304	0.120	0.009	0.793	0.460	0.145	0.013
	`RAMP`	0.427	0.058	0.005	0.003	0.613	0.087	0.155	0.005
	`SPAM`	1	0.068	NA	NA	1	0.039	NA	NA
	`VANISH`	0.967	0.001	0.420	0.001	1	0.001	0.365	0.001
	`LASER`	1	0.068	0.985	0.008	1	0.039	1	0.007

References

David Benkeser and Mark van der Laan. The highly adaptive lasso estimator. In *Proceedings of the International Conference on Data Science and Advanced Analytics*, pages 689–696, 2016.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1113–1141, 2013.

Tony Cai, Weidong Liu, and Xi Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494): 594–607, 2011.

Shizhe Chen, Daniela M Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2014.

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Table 5: The true and false positive rates (TPR and FPR) for the main effects and the second-order interaction effects for three different scenarios in Section 5.1.2, averaged over 100 data sets. The results are for models trained with n training samples and $d = 200$. Most of the standard errors are approximately zero and are omitted.

		$n = 350$				$n = 700$			
		Main Effects		Interaction Effects		Main Effects		Interaction Effects	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Scenario B1	BT	1	0.477	0.997	0.001	1	0.502	0.973	0.001
	RAMP	1	0.001	1	0	1	0.001	1	0
	SPAM	1	0.003	NA	NA	1	0.003	NA	NA
	LASER	1	0.003	1	0.001	1	0.003	1	0.001
Scenario B2	BT	0.997	0.457	0.906	0.001	1	0.453	0.963	0.001
	RAMP	1	0.018	0.667	0.001	1	0.001	0.670	0.001
	SPAM	0.997	0.002	NA	NA	1	0.001	NA	NA
	LASER	0.997	0.002	0.993	0.001	1	0.001	1	0.001
Scenario B3	BT	0.993	0.565	0.870	0.001	1	0.507	0.987	0.001
	RAMP	1	0.013	0.667	0.001	1	0.001	0.670	0.001
	SPAM	1	0.003	NA	NA	1	0.003	NA	NA
	LASER	1	0.003	1	0.001	1	0.003	1	0.001

Table 6: The true and false positive rates (TPR and FPR) for the main effects and the second-order interaction effects for three different scenarios in Section 5.1.2, averaged over 100 data sets. The results are for models trained with n training samples and $d = 400$. Most of the standard errors are approximately zero and are omitted.

		$n = 350$				$n = 700$			
		Main Effects		Interaction Effects		Main Effects		Interaction Effects	
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Scenario B1	BT	1	0.208	1	0	1	0.435	0.99	0.001
	RAMP	1	0	1	0	1	0.001	1	0
	SPAM	1	0.002	NA	NA	1	0.002	NA	NA
	LASER	1	0.002	1	0.001	1	0.002	1	0.001
Scenario B2	BT	1	0.143	0.880	0	1	0.335	0.953	0.001
	RAMP	1	0.007	0.673	0.001	1	0.001	0.670	0.001
	SPAM	1	0.003	NA	NA	1	0.001	NA	NA
	LASER	1	0.003	1	0.001	1	0.001	1	0.001
Scenario B3	BT	1	0.200	0.883	0.001	1	0.677	0.907	0.001
	RAMP	1	0.008	0.634	0.001	1	0.001	0.670	0.001
	SPAM	1	0.002	NA	NA	1	0.001	NA	NA
	LASER	1	0.002	1	0.001	1	0.001	1	0.001

Yingying Fan, Gareth M James, and Peter Radchenko. Functional additive regression. *The Annals of Statistics*, 43(5):2296–2325, 2015.

Peter GM Forbes and Steffen Lauritzen. Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and its Applications*, 473:261–283, 2015.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Eitan Greenshtein and Ya'Acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004.

Chong Gu. *Smoothing Spline ANOVA Models*, volume 297. Springer Science & Business Media, 2013.

Chong Gu and Jingyuan Wang. Penalized likelihood density estimation: Direct cross-validation and scalable approximation. *Statistica Sinica*, 13(3):811–826, 2003.

Ning Hao and Hao Helen Zhang. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014.

Ning Hao, Yang Feng, and Hao Helen Zhang. Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522): 615–625, 2018.

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, 2016.

Trevor J Hastie and Robert J Tibshirani. *Generalized Additive Models*, volume 43. CRC Press, 1990.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

Eric Janofsky. Exponential series approaches for nonparametric graphical models. *arXiv preprint arXiv:1506.03537*, 2015.

Yongho Jeon and Yi Lin. An effective method for high-dimensional log-density ANOVA estimation, with application to nonparametric graphical model building. *Statistica Sinica*, 16(2):353–374, 2006.

Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Science & Business Media, 2007.

Tom Leonard. Density estimation, stochastic processes and prior information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(2):113–146, 1978.

Lina Lin, Mathias Drton, and Ali Shojaie. High-dimensional inference of graphical models using regularized score matching. *Electronic Journal of Statistics*, 10(1):806–854, 2016.

Yi Lin. Tensor product space ANOVA models. *The Annals of Statistics*, 28(3):734–755, 2000.

Yi Lin and Hao Helen Zhang. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Weidong Liu and Xi Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis*, 135:153–162, 2015.

Yin Lou, Jacob Bien, Rich Caruana, and Johannes Gehrke. Sparse partially linear additive models. *Journal of Computational and Graphical Statistics*, 25(4):1126–1140, 2016.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462, 2006.

Richard Nickl and Benedikt M Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov-and Sobolev-type. *Journal of Theoretical Probability*, 20(2):177–199, 2007.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Peter Radchenko and Gareth M James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105 (492):1541–1553, 2010.

Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Rajen D Shah. Modelling interactions in high-dimensional data with backtracking. *The Journal of Machine Learning Research*, 17(207):1–31, 2016.

Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, 10(3):795–810, 1982.

Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418, 2013.

Kean Ming Tan, Palma London, Karthik Mohan, Su-In Lee, Maryam Fazel, and Daniela Witten. Learning graphical models with hubs. *The Journal of Machine Learning Research*, 15(1):3297–3331, 2014.

Kean Ming Tan, Daniela Witten, and Ali Shojaie. The cluster graphical lasso for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis*, 85: 23–36, 2015.

Kean Ming Tan, Yang Ning, Daniela M Witten, and Han Liu. Replicates in high dimensions, with applications to latent variable graphical models. *Biometrika*, 103(4):761–777, 2016.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

Yu-Xiang Wang, James Sharpnack, Alexander J Smola, and Ryan J Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016.

Eunho Yang, Pradeep K Ravikumar, Genevera I Allen, and Zhandong Liu. On Poisson graphical models. In *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013.

Eunho Yang, Pradeep Ravikumar, Genevera I Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16(1):3813–3847, 2015.

Zhuoran Yang, Yang Ning, and Han Liu. On semiparametric exponential family graphical models. *Journal of Machine Learning Research*, 19(57):1–59, 2018.

Paul Yau, Robert Kohn, and Sally Wood. Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, 12(1), 2012.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Hao Helen Zhang, Grace Wahba, Yi Lin, Meta Voelker, Michael Ferris, Ronald Klein, and Barbara Klein. Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association*, 99(467):659–672, 2004.