Journal of Machine Learning Research 23 (2022) 1-36 Submitted 2/22; Revised 7/22; Published 10/22

Nonparametric Neighborhood Selection in Graphical Models

Hao Dong HAO_DONG@PSTAT.UCSB.EDU
Department of Statistics and Applied Probability
University of California, Santa Barbara
Santa Barbara, CA, USA

Yuedong Wang YUEDONG@PSTAT.UCSB.EDU
Department of Statistics and Applied Probability
University of California, Santa Barbara
Santa Barbara, CA, USA

Editor: Daniela Witten

Abstract

The neighborhood selection method directly explores the conditional dependence structure and has been widely used to construct undirected graphical models. However, except for some special cases with discrete data, there is little research on nonparametric methods for neighborhood selection with mixed data. This paper develops a fully nonparametric neighborhood selection method under a consolidated smoothing spline ANOVA (SS ANOVA) decomposition framework. The proposed model is flexible and contains many existing models as special cases. The proposed method provides a unified framework for mixed data without any restrictions on the type of each random variable. We detect edges by applying an L_1 regularization to interactions in the SS ANOVA decomposition. We propose an iterative procedure to compute the estimates and establish the convergence rates for conditional density and interactions. Simulations indicate that the proposed methods perform well under Gaussian and non-Gaussian settings. We illustrate the proposed methods using two real data examples.

Keywords: conditional density estimation, mixed data, regularization, reproducing kernel Hilbert space, smoothing spline ANOVA

1. Introduction

Discovering conditional independence among random variables is an essential task in statistics. Undirected probabilistic graphical models play a pivotal role in characterizing conditional independence. They have been utilized in a wide range of scientific and engineering domains, including statistical physics, computer vision, machine learning, and computational biology (Koller and Friedman, 2009). A graphical model is constructed based on an undirected graph $G = (V, E)$ with node set $V = \{1, \cdots, p\}$ representing p random variables $X_1, \cdots, X_p$ and edge set $E \subseteq V \times V$ describing the conditional dependence among $X_1, \cdots, X_p$. Let $\boldsymbol{X} = (X_1, \cdots, X_p)$ and $\boldsymbol{X}_{\setminus\{i_1,\cdots,i_k\}}$ be the sub-vector of $\boldsymbol{X}$ without elements in $\{i_1, \cdots, i_k\}$. Then, $\{i, j\} \notin E$ corresponds to the conditional independence between X_i and X_j given other variables in $\boldsymbol{X}$, denoted as $X_i \perp X_j | \boldsymbol{X}_{\setminus\{i,j\}}$.

As joint density ultimately determines the conditional relationship, methods for edge detection based on estimating joint density have been proposed (Yuan and Lin, 2007; Banerjee

et al., 2008; Friedman et al., 2008; Hsieh et al., 2014; Liu et al., 2009). Under the Gaussian assumption of $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, the task of edge detection reduces to the estimation of the precision matrix Σ^{-1}. Yuan and Lin (2007), Banerjee et al. (2008), and Friedman et al. (2008) proposed regularization methods that minimize the log-likelihood with an L_1 penalty on the entries of Σ^{-1}. Hsieh et al. (2014) proposed a fast second-order algorithm for solving the L_1-regularized Gaussian MLE. Liu et al. (2009) extended the L_1-regularized Gaussian MLE approach to the setting where there exist monotone transformations $f_1, \cdots, f_p$ such that $(f_1(X_1), \cdots, f_p(X_p)) \sim \mathcal{N}(\mathbf{0}, \Sigma)$. These parametric and semi-parametric methods may be too restrictive for some applications and cannot handle mixed data since they rely on the Gaussian assumption.

Let $f(\boldsymbol{x})$ be the joint density function of $\boldsymbol{X}$, and consider the transformation $f(\boldsymbol{x}) = e^{\eta(\boldsymbol{x})} / \int e^{\eta(\boldsymbol{x})} d\boldsymbol{x}$, where $\eta(\boldsymbol{x})$ is the logistic transformation of f. The SS ANOVA decomposition represents $\eta(\boldsymbol{x})$ as a summation of a constant, main effects, and interactions:

$$\eta(x_1, \cdots, x_p) = c + \sum_{j=1}^{p} \eta_j(x_j) + \sum_{1 \le j < k \le p} \eta_{jk}(x_j, x_k) + \cdots + \eta_{1 \cdots p}(x_1, \cdots, x_p). \tag{1}$$

The conditional independence $X_j \perp X_k | \boldsymbol{X}_{\setminus \{j,k\}}$ is equivalent to the summation of all interactions involving x_j and x_k equal to zero (Gu, 2013). Consequently, identifying edges is equivalent to identifying nonzero interactions. Jeon and Lin (2006) developed a penalized M-estimation method for edge detection based on the SS ANOVA decomposition (1). Our experience indicates that this joint density estimation approach is only computationally feasible with a small p due to large memory requirements.

The neighborhood selection approach explores structures in conditional densities and is usually more computationally efficient. By the conditional independence properties of undirected graphical models, for any node $\alpha \in V$, X_α only depends on other variables in its neighborhood set $nb_G(\alpha)$, where $nb_G(\alpha) = \{k \in V | \{\alpha, k\} \in E\}$. Consequently, the conditional independence structure of graph G can be constructed by estimating all of its neighborhoods $nb_G(\alpha)$ for $\alpha = 1, \cdots, p$. Neighborhood selection aims to identify a minimal set of variables $nb_G(\alpha)$ that X_α depends on for each node $\alpha \in V$.

Many neighborhood selection methods have been developed for learning sparse graphical models (Hastie et al., 2015; Drton and Maathuis, 2017). Flexible models were proposed for discrete data (Höfling and Tibshirani, 2009; Ravikumar et al., 2010). Methods for continuous data usually model the conditional mean (Meinshausen and Bühlmann, 2006; Voorman et al., 2014) or conditional quantiles (Ali et al., 2016). For example, Meinshausen and Bühlmann (2006) and Peng et al. (2009) considered a linear model for the conditional mean with L_1 penalties on coefficients and partial correlations, respectively. Voorman et al. (2014) considered an additive model for the conditional mean. The conditional mean approach does not assume a specific distribution for the regression error and therefore appears to be distribution-free. However, if the conditional relationships are linear, the joint distribution must be multivariate Gaussian under mild assumptions (Voorman et al., 2014). In other words, the restriction of Gaussianity has not been removed as it appears. For mixed continuous and discrete variables, Lee and Hastie (2015) considered a pairwise model that generalizes Gaussian graphical and discrete models. Chen et al. (2015) proposed a flexible pairwise graphical model where each node's conditional distribution is in the exponential

family. Gu and Ma (2011) developed a functional ANOVA method for estimating the conditional density of cross-classified responses and identifying conditional independence structure through Kullback-Leibler projection.

Modeling mixed data is challenging since it is difficult to specify a joint density. Existing neighborhood selection methods are restrictive since they model specific mixed data or assume specific conditional distributions. In this paper, we propose a new fully nonparametric neighborhood selection method. We construct an SS ANOVA model for each conditional density and select neighborhood via L_1 regularization. The new contributions of our neighborhood selection method consist of four parts. First, we directly target the neighborhood definition based on conditional density without assuming any specific family of distributions. The whole conditional density provides the most comprehensive summary of the relationship which might be missed by specific characteristics such as conditional mean and quantiles. Second, we allow the range of each random variable to be an arbitrary set and use the tensor product of reproducing kernel Hilbert spaces (RKHS) to construct a model space for each conditional density. Therefore, the proposed method provides a unified framework for mixed data types without any restrictions on the type of each random variable. The proposed model is more general and flexible than existing models. Third, we use the SS ANOVA structure to facilitate the selection of neighborhoods. Specifically, we estimate the conditional density for each node based on SS ANOVA decomposition with an L_1 penalty involving interaction components. This approach for neighborhood selection has not been studied before. Last but not least, the new neighborhood selection method based on conditional density is more computationally efficient than those based on joint density and is parallelizable.

The rest of the paper is organized as follows. Section 2 introduces the new neighborhood selection method. Section 3 presents the computational method and the implementation of the proposed algorithm. Section 4 derives the convergence rate of the conditional density estimate and its components in the SS ANOVA decomposition. Section 5 conducts simulations to compare edge detection performance with existing methods under both Gaussian and non-Gaussian settings. Section 6 illustrates the proposed methods using two real data sets. Section 7 provides some discussion. The Appendix contains proofs and auxiliary material.

2. Neighborhood Selection Through Conditional Density Estimation with L_1 Penalty

In this section, we first introduce some notation and the SS ANOVA decomposition. Then, we present our nonparametric method for edge detection.

2.1 Notation and SS ANOVA Decomposition

Consider p random variables $X_1, \cdots, X_p$ with ranges denoted as $\mathcal{X}_1, \cdots, \mathcal{X}_p$. Each range $\mathcal{X}_\alpha$ is an arbitrary set for generality. It may be a continuous interval, a discrete set, or a circle. It could even be a subset in Euclidean space or a sphere. That is, each $\mathcal{X}_\alpha$ could be a multivariate random variable. Denote $\boldsymbol{X} = (X_1, \cdots, X_p)$ as the p-dimensional random vector with range $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ and $\boldsymbol{x} = (x_1, \cdots, x_p)$ as a realization of the random vector. For a fixed $\alpha \in V = \{1, \cdots, p\}$, denote $\boldsymbol{X}_{\setminus\{\alpha\}} = (X_1, \cdots, X_{\alpha-1}, X_{\alpha+1}, \cdots, X_p)$

and $\boldsymbol{x}_{\setminus\{\alpha\}} = (x_1, \cdots, x_{\alpha-1}, x_{\alpha+1}, \cdots, x_p)$ as the vectors of $\boldsymbol{X}$ and $\boldsymbol{x}$ with the αth element being removed. Our goal is to select the neighborhood $nb_G(\alpha)$ through the estimation of the conditional density $f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})$.

Denote $\boldsymbol{X}_i = (X_{i,1}, \cdots, X_{i,p})$ and $\boldsymbol{x}_i = (x_{i,1}, \cdots, x_{i,p})$ for $i = 1, \cdots, n$ as n i.i.d. random vectors and their realizations. Let $\boldsymbol{x}_{i,\setminus\{\alpha\}} = (x_{i,1}, \cdots, x_{i,\alpha-1}, x_{i,\alpha+1}, \cdots, x_{i,p})$. Denote $\boldsymbol{x}_i^\alpha = (x_{i,1}, \cdots, x_{i,\alpha-1}, x_\alpha, x_{i,\alpha+1}, \cdots, x_{i,p})$ as the p-dimensional vector with x_α varies in $\mathcal{X}_\alpha$ and all other variables fixed at their ith realizations. For simplicity, the dependence of $\boldsymbol{x}_i^\alpha$ on x_α is not expressed explicitly.

Let $\mathcal{H}^{(j)}$ be an RKHS on $\mathcal{X}_j$ and $\mathcal{H}^{(j)} = \{1_{(j)}\} \oplus \mathcal{H}_{(j)}$, where $\{1_{(j)}\}$ is the space of the constant functions on $\mathcal{X}_j$ and $\mathcal{H}_{(j)}$ is the orthogonal complement of $\{1_{(j)}\}$. One may construct a flexible and interpretable model for a p-dimensional function through the following SS ANOVA decomposition of the tensor product space $\bigotimes_{j=1}^{p} \mathcal{H}^{(j)}$ on $\mathcal{X}$ (Wang, 2011; Gu, 2013):

$$\begin{aligned}\bigotimes_{j=1}^{p} \mathcal{H}^{(j)} &= \bigotimes_{j=1}^{p} \left\{\{1_{(j)}\} \oplus \mathcal{H}_{(j)}\right\} \\ &= \{1\} \oplus \left\{\bigoplus_{j=1}^{p} \mathcal{H}_{(j)}\right\} \oplus \left\{\bigoplus_{1\le j<k\le p} [\mathcal{H}_{(j)} \otimes \mathcal{H}_{(k)}]\right\} \oplus \cdots \oplus \left\{\mathcal{H}_{(1)} \otimes \cdots \otimes \mathcal{H}_{(p)}\right\}. \end{aligned} \tag{2}$$

The decomposition in equation (1) corresponds to the SS ANOVA decomposition to the logistic transformation of the joint density function.

2.2 SS ANOVA Model for Conditional Density

For the conditional density of X_α, we consider the logistic density transformation

$$f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}}) = \frac{e^{\eta(\boldsymbol{x})}}{\int_{\mathcal{X}_\alpha} e^{\eta(\boldsymbol{x})} dx_\alpha} \tag{3}$$

to enforce the conditions of $f > 0$ and $\int f = 1$. The function η is the logistic transformation of f. An SS ANOVA model for η in (3) may contain any subset of components in the SS ANOVA decomposition (2). For simplicity, we assume that $\eta \in \mathcal{M}_\alpha$ where

$$\mathcal{M}_\alpha = \{1\} \oplus \left\{\bigoplus_{j=1}^{p} \mathcal{H}_{(j)}\right\} \oplus \left\{\bigoplus_{k\neq\alpha} [\mathcal{H}_{(\alpha)} \otimes \mathcal{H}_{(k)}]\right\} \tag{4}$$

is a subspace with main effects and two-way interactions only. A function $\eta \in \mathcal{M}_\alpha$ can be decomposed as follows:

$$\eta(\boldsymbol{x}) = \varsigma + \sum_{j=1}^{p} \eta_j(x_j) + \sum_{k\neq\alpha} \eta_{\alpha k}(x_\alpha, x_k), \tag{5}$$

where each functional component in (5) belongs to the corresponding subspace in (4). We note that the proposed method can be easily extended to include higher-order interactions.

Remark 1 The SS ANOVA model (1) for the joint density with main effects and two-way interaction only is a pairwise graphical model, which is commonly assumed in the existing literature.

Remark 2 We consider both the log-likelihood and pseudo log-likelihood approaches for estimating the conditional density (Gu, 2013). We present the pseudo likelihood approach in the main text since it is computationally more efficient. The log-likelihood approach is presented in Appendix A. For the pseudo log-likelihood approach, model space $\mathcal{M}_\alpha$ includes constant functions. The model space for the log-likelihood approach eliminates the constant functions for identifiability.

Remark 3 To compare the estimation between the joint and neighborhood approaches under pairwise graphical models, we consider the SS ANOVA decomposition (1) with main effects and two-way interaction only for the joint density and the SS ANOVA decomposition (5) for the conditional density. The joint density approach needs to estimate all main effects and two-way interactions simultaneously with a total number of components proportional to p^2. Our experience indicates that this joint approach is computationally infeasible even with moderately large p due to memory constraints. On the other hand, the neighborhood approach needs to estimate p main effects and $p-1$ two-way interactions for each node, which significantly reduces the computational cost and memory requirement and is parallelizable.

Remark 4 Model (5) contains many parametric models as special cases. Specifically, the Gaussian graphical model is a special case with $\mathcal{X}_j = \mathbb{R}$, $\eta_j(x_j) = \beta_j x_j - x_j^2/2$ for $j = \alpha$ and 0 otherwise, and $\eta_{\alpha k}(x_\alpha, x_k) = \beta_{\alpha k} x_\alpha x_k$ for some constants β_j and $\beta_{\alpha k}$. The Ising model for binary data is a special case with $\mathcal{X}_j = \{0, 1\}$, $\eta_j(x_j) = x_j$ for $j = \alpha$ and 0 otherwise, and $\eta_{\alpha k}(x_\alpha, x_k) = \beta_{\alpha k} x_\alpha x_k$. The Poisson graphical model for discrete data is a special case with $\mathcal{X}_j = \{0, 1, 2, \cdots\}$, $\eta_j(x_j) = x_j - \log(x_j!)$ for $j = \alpha$ and 0 otherwise, and $\eta_{\alpha k}(x_\alpha, x_k) = \beta_{\alpha k} x_\alpha x_k$. The exponential family model proposed by Suggala et al. (2017),

$$\log f(x_\alpha | \boldsymbol{x}_{\setminus\{\alpha\}}) \propto \left\{ \beta_\alpha B_\alpha(x_\alpha) + \sum_{\{\alpha,k\}\in E} \beta_{\alpha k} B_\alpha(x_\alpha) B_k(x_k) + C_\alpha(x_\alpha) \right\}, \tag{6}$$

is also a special case with $\eta_j(x_j) = \beta_j B_j(x_j) + C_j(x_j)$ for $j = \alpha$ and 0 otherwise, and $\eta_{\alpha k}(x_\alpha, x_k) = \beta_{\alpha k} B_\alpha(x_\alpha) B_k(x_k)$. Note that many existing exponential family models including (6) assume a multiplicative interaction while model (5) does not assume any specific interaction. Therefore, the proposed model is more general.

2.3 Penalized Pseudo Log-likelihood Estimation

For each node $\alpha \in V$, we assume that $\eta(\boldsymbol{x}) \in \mathcal{M}_\alpha$ where $\mathcal{M}_\alpha$ is given in (4) and η is decomposed as in (5). We further decompose $\mathcal{H}_{(j)}$ as $\mathcal{H}_{(j)} = \mathcal{H}^0_{(j)} \oplus \mathcal{H}^1_{(j)}$ where $\mathcal{H}^0_{(j)}$ is a finite dimensional space containing functions that are not subject to penalty. We estimate η in (5) by minimizing the following penalized pseudo log-likelihood in $\mathcal{M}_\alpha$:

$$l_\alpha + \frac{\lambda_1}{2} \sum_{j=1}^{p} \theta_j^{-1} ||P_j \eta_j||^2 + \tau_1 \sum_{k\neq\alpha} w_{\alpha k} ||\eta_{\alpha k}||, \tag{7}$$

where $l_\alpha = n^{-1}\sum_{i=1}^n \left\{e^{-\eta(\boldsymbol{x}_i)} + \int_{\mathcal{X}_\alpha} \eta(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha\right\}$ is the pseudo log-likelihood, $\rho(\cdot)$ is a known density of X_α conditional on $\boldsymbol{X}_{\setminus\{\alpha\}} = \boldsymbol{x}_{i,\setminus\{\alpha\}}$, P_j is the projection operator onto $\mathcal{H}^1_{(j)}$, λ_1, τ_1, and θ_j's are tuning parameters, $0 \le w_{\alpha k} < \infty$ are pre-specified weights, and $||\cdot||$ is an induced norm in $\mathcal{M}_\alpha$. The pseudo log-likelihood l_α measures the goodness-of-fit. The second element in (7) is the roughness L_2 penalty on main effects. The third element in (7) is the L_1 penalty for selecting the neighborhood $nb_G(\alpha)$. We allow different weights in the L_1 penalty for flexibility.

Remark 5 The idea of pseudo log-likelihood was first developed in Jeon and Lin (2006) for joint density estimation. Gu (2013) extended this approach to conditional density estimation. We present the pseudo log-likelihood estimation in the main text since this approach is computationally more efficient. The log-likelihood approach to conditional density estimation needs to calculate the integral $\int_{\mathcal{X}_\alpha} e^{\eta(\boldsymbol{x}_i^\alpha)}dx_\alpha$ repeatedly, which can be computationally intensive. With a proper choice of ρ, the pseudo log-likelihood approach needs to calculate an integral only once.

Remark 6 The proposed method replaces the L_2 penalty on interactions in Gu (2013) with the L_1 penalty for neighborhood selection and differs from that in Jeon and Lin (2006) in two aspects. First, Jeon and Lin (2006)'s approach is a global method that estimates the joint density; thus is computationally intensive and can only handle small dimensions p. Second, Jeon and Lin (2006) posed the L_1 penalty to both main effects and interactions. Consequently, their method selects both nodes and edges. In practice, the nodes are usually given, and the goal is to detect edges. Therefore, we consider the smoothness promoting L_2 penalty to main effects and the sparsity promoting L_1 penalty to interactions.

Let

$$\mathcal{G} = \left\{\bigoplus_{j=1}^p \mathcal{H}_{(j)}\right\} \oplus \left\{\bigoplus_{k\neq\alpha} [\mathcal{H}_{(\alpha)} \otimes \mathcal{H}_{(k)}]\right\}. \tag{8}$$

We can rewrite $\eta(\boldsymbol{x}) = \varsigma + g(\boldsymbol{x})$ where $g(\boldsymbol{x}) = \sum_{j=1}^p \eta_j(x_j) + \sum_{k\neq\alpha} \eta_{\alpha k}(x_\alpha, x_k) \in \mathcal{G}$. We first estimate ς with fixed g and then estimate g using the profiled pseudo log-likelihood. The results are summarized in the following proposition.

Proposition 1 *With fixed g, the minimizer of ς in (7) is $\hat{\varsigma} = \log\{n^{-1}\sum_{i=1}^n e^{-g(x_i)}\}$ and the penalized pseudo log-likelihood (7) reduces to the following penalized profiled pseudo log-likelihood*

$$l(\hat{\varsigma}(g), g) + \frac{\lambda_1}{2}\sum_{j=1}^p \theta_j^{-1}||P_j\eta_j||^2 + \tau_1 \sum_{k\neq\alpha} w_{\alpha k}||\eta_{\alpha k}||, \tag{9}$$

where $l(\hat{\varsigma}(g), g) = log\{n^{-1}\sum_{i=1}^n e^{-g(\boldsymbol{x}_i)}\} + n^{-1}\sum_{i=1}^n \int_{\mathcal{X}_\alpha} g(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha$ is the profiled pseudo log-likelihood.

The proof can be found in Appendix C. Instead of minimizing (9) that involves L_1 penalties on functions, as in Lin and Zhang (2006), we will solve an equivalent but more convenient minimization problem that involves L_1 penalties on the smoothing parameters.

Proposition 2 *Minimizing*

$$l(\hat{\varsigma}(g), g) + \frac{\lambda_1}{2}\Big(\sum_{j=1}^{p}\theta_j^{-1}||P_j\eta_j||^2 + \sum_{k\neq\alpha} w_{\alpha k}\theta_{\alpha k}^{-1}||\eta_{\alpha k}||^2\Big) + \lambda_2\sum_{k\neq\alpha} w_{\alpha k}\theta_{\alpha k}, \tag{10}$$

subject to $\theta_{\alpha k} \geq 0$ *for* $k = 1, \cdots, p$ *and* $k \neq \alpha$ *is equivalent to minimizing* (9).

The proof of equivalence can be found in Appendix C. Proposition 2 transforms the selection of nonzero functions $\eta_{\alpha k}$ in (9) into a selection of nonzero parameters $\theta_{\alpha k}$. The minimization problem (10) consists of L_2 penalties on functions and L_1 penalties on parameters and existing methods can be modified to solve each part. Computational details for solving (10) are presented in Section 3.

Since the pseudo log-likelihood is used for estimation, we need to compute the conditional density estimate using the following proportion.

Proposition 3 *The resulting estimate of the conditional density is* $\hat{f}(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}}) \propto e^{\hat{g}(\boldsymbol{x})}\rho(\boldsymbol{x})$ *where* $\hat{g}$ *is the minimizer of* (10).

The proof of Proposition 3 is given in Appendix C. Notice that the minimization problem (10) involves $p-1$ two-way interaction terms. Solving (10) for all $\alpha = 1, \cdots, p$ leads to two estimates for each two-way interaction, denoted as $\hat{\eta}_{\alpha k}$ and $\hat{\eta}_{k\alpha}$ for $\alpha, k = 1, \cdots, p$ and $\alpha \neq k$. There are two commonly used rules to combine the results: AND-rule ($\{\alpha, k\} \in E$ iff $\hat{\eta}_{\alpha k} \neq 0$ and $\hat{\eta}_{k\alpha} \neq 0$) or OR-rule ($\{\alpha, k\} \in E$ iff $\hat{\eta}_{\alpha k} \neq 0$ or $\hat{\eta}_{k\alpha} \neq 0$) (Hastie et al., 2015). As discussed in Section 4.2 in Chen et al. (2015), when the αth and kth nodes are of the same type (same marginal distribution) or are both non-Gaussian, there is no clear reason to prefer one edge estimate over the other. We adopt the AND-rule in all simulations and real data examples.

3. Algorithm

In this section, we propose a computational algorithm that solves (10) iteratively. Denote $\boldsymbol{\theta}_1 = (\theta_1, \cdots, \theta_p)^T$, $\boldsymbol{\theta}_2 = (\theta_{\alpha 1}, \cdots, \theta_{\alpha(\alpha-1)}, \theta_{\alpha(\alpha+1)}, \cdots, \theta_{\alpha p})^T$, and $\boldsymbol{w} = (w_{\alpha 1}, \cdots, w_{\alpha(\alpha-1)}, w_{\alpha(\alpha+1)}, \cdots, w_{\alpha p})^T$. Let $\mathcal{H}_{(j)} = \mathcal{H}^0_{(j)} \oplus \mathcal{H}^1_{(j)}$ where $\mathcal{H}^0_{(j)}$ is a finite-dimensional space containing functions that are not subject to L_2 penalty. Denote $\phi_{j1}, \cdots, \phi_{jm_j}$ as basis functions of $\mathcal{H}^0_{(j)}$, and R^1_j, R_j, and $R_{\alpha k}$ as reproducing kernels of $\mathcal{H}^1_{(j)}$, $\mathcal{H}_{(j)}$, and $\mathcal{H}_{(\alpha k)}$, respectively. We collect all basis functions ϕ_{jk} for $j = 1, \cdots, p$ and $k = 1, \cdots, m_j$ and denote them as $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)^T$, a vector of functions of $\boldsymbol{x}$ with dimension $m = \sum_{j=1}^p m_j$.

Since in general the minimization problem (10) does not have a solution in a finite-dimensional space, as in Gu (2013), we approximate the solution by a subset of representers. Specifically, let $\{\tilde{\boldsymbol{x}}_u = (\tilde{x}_{u,1}, \cdots, \tilde{x}_{u,p}),\ u = 1, \cdots, q\}$ be a subset of all observations $\{\boldsymbol{x}_i,\ i = 1, \cdots, n\}$. Let $\xi_{1ju}(x_j) = R^1_j(\tilde{x}_{u,j}, x_j)$ and $\xi_{\alpha ku}(x_\alpha, x_k) = R_{\alpha k}((\tilde{x}_{u,\alpha}, \tilde{x}_{u,k}), (x_\alpha, x_k))$ for $u = 1, \cdots, q$, $k = 1, \cdots, p$, and $k \neq \alpha$. Let $\xi_{\boldsymbol{\theta_1},u}(\boldsymbol{x}) = \sum_{j=1}^{p}\theta_j\xi_{1ju}(x_j)$, $\boldsymbol{\xi}_{\boldsymbol{\theta}_1}(\boldsymbol{x}) = (\xi_{\boldsymbol{\theta_1},1}, \cdots, \xi_{\boldsymbol{\theta_1},q})^T$, $\xi_{\boldsymbol{\theta_2},u}(\boldsymbol{x}) = \sum_{k=1,k\neq\alpha}^{p} w_{\alpha k}^{-1}\theta_{\alpha k}\xi_{\alpha ku}(x_\alpha, x_k)$, $\boldsymbol{\xi}_{\boldsymbol{\theta}_2}(\boldsymbol{x}) = (\xi_{\boldsymbol{\theta_2},1}, \cdots, \xi_{\boldsymbol{\theta_2},q})^T$, and $\boldsymbol{\xi}(\boldsymbol{x}) = \boldsymbol{\xi}_{\boldsymbol{\theta}_1}(\boldsymbol{x}) +$

$\boldsymbol{\xi}_{\boldsymbol{\theta}_2}(\boldsymbol{x})$. The approximate solution can be represented as a linear combination of basis functions and representers:

$$\begin{aligned}\hat{g}(\boldsymbol{x}) &= \sum_{v=1}^{m} d_v \phi_v(\boldsymbol{x}) + \sum_{u=1}^{q} c_u \left\{ \sum_{j=1}^{p} \theta_j \xi_{1ju}(x_j) + \sum_{k=1,k\neq\alpha}^{p} w_{\alpha k}^{-1} \theta_{\alpha,k} \xi_{\alpha k u}(x_\alpha, x_k) \right\} \\ &= \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{d} + \boldsymbol{\xi}^T(\boldsymbol{x})\boldsymbol{c}, \end{aligned} \tag{11}$$

where $\boldsymbol{c} = (c_1, \cdots, c_q)^T$ and $\boldsymbol{d} = (d_1, \cdots, d_m)^T$ are coefficients. Let $Q = \sum_{j=1}^{p} \theta_j Q_j + \sum_{k=1,k\neq\alpha}^{p} w_{\alpha k}^{-1} \theta_{\alpha k} Q_{\alpha k}$, where $Q_j = \left\{ R_j^1(\tilde{x}_{u,j}, \tilde{x}_{v,j}) \right\}_{u,v=1}^{q}$ are kernel matrices for the main effects and $Q_{\alpha k} = \left\{ R_{\alpha k}((\tilde{x}_{u,\alpha}, \tilde{x}_{u,k}), (\tilde{x}_{v,\alpha}, \tilde{x}_{v,k})) \right\}_{u,v=1}^{q}$ are kernel matrices for the two-way interactions. We can rewrite (10) in a vector form:

$$A(\boldsymbol{c}, \boldsymbol{d}, \boldsymbol{\theta}_2) = \log \left\{ \frac{1}{n} \sum_{i=1}^{n} e^{-\boldsymbol{\phi}_i^T \boldsymbol{d} - \boldsymbol{\xi}_i^T \boldsymbol{c}} \right\} + \boldsymbol{b}_{\boldsymbol{\phi}}^T \boldsymbol{d} + \boldsymbol{b}_{\boldsymbol{\xi}}^T \boldsymbol{c} + \frac{\lambda_1}{2} \boldsymbol{c}^T Q \boldsymbol{c} + \lambda_2 \boldsymbol{w}^T \boldsymbol{\theta}_2, \tag{12}$$

where $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$, $\boldsymbol{\xi}_i = \boldsymbol{\xi}(\boldsymbol{x}_i)$, $\boldsymbol{b}_{\boldsymbol{\phi}} = n^{-1} \sum_{i=1}^{n} \int_{\mathcal{X}_\alpha} \boldsymbol{\phi}(\boldsymbol{x}_i^\alpha) \rho(\boldsymbol{x}_i^\alpha) dx_\alpha$, and $\boldsymbol{b}_{\boldsymbol{\xi}} = n^{-1} \sum_{i=1}^{n} \int_{\mathcal{X}_\alpha} \boldsymbol{\xi}(\boldsymbol{x}_i^\alpha) \rho(\boldsymbol{x}_i^\alpha) dx_\alpha$. We solve (12) by updating $\boldsymbol{c}$, $\boldsymbol{d}$, and $\boldsymbol{\theta}_2$ between two steps discussed in the following two subsections.

3.1 Newton-Raphson Procedure

We fix $\boldsymbol{\theta}_2$ and update $\boldsymbol{c}$ and $\boldsymbol{d}$ at this step. Dropping the last term which is independent of $\boldsymbol{c}$ and $\boldsymbol{d}$, (12) reduces to

$$A_1(\boldsymbol{c}, \boldsymbol{d}) = \log \left\{ \frac{1}{n} \sum_{i=1}^{n} e^{-\boldsymbol{\phi}_i^T \boldsymbol{d} - \boldsymbol{\xi}_i^T \boldsymbol{c}} \right\} + \boldsymbol{b}_{\boldsymbol{\phi}}^T \boldsymbol{d} + \boldsymbol{b}_{\boldsymbol{\xi}}^T \boldsymbol{c} + \frac{\lambda_1}{2} \boldsymbol{c}^T Q \boldsymbol{c}. \tag{13}$$

Note that (13) has the same form as (10.31) in Gu (2013). Therefore, we can solve (13) using the Newton-Raphson procedure with λ_1 and $\boldsymbol{\theta}_1$ selected by the approximate cross-validation (ACV) method (Gu, 2013). We note that $\boldsymbol{\theta}_2$ are fixed at this step. Therefore, the existing function in the `gss` R package cannot be used directly. More implementation details can be found in Appendix B.1.

3.2 Quadratic Programming

We fix $\boldsymbol{c}$, $\boldsymbol{d}$, λ_1 and $\boldsymbol{\theta}_1$ and update $\boldsymbol{\theta}_2$ at this step. We rewrite $\hat{g}$ in (11) as

$$\begin{aligned}\hat{g}(\boldsymbol{x}) &= \sum_{v=1}^{m} d_v \phi_v(\boldsymbol{x}) + \sum_{j=1}^{p} \theta_j \sum_{u=1}^{q} c_u \xi_{1ju}(x_j) + \sum_{k=1,k\neq\alpha}^{p} \theta_{\alpha k} w_{\alpha k}^{-1} \sum_{u=1}^{q} c_u \xi_{\alpha k u}(x_\alpha, x_k) \\ &= \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{d} + \boldsymbol{\psi}_1^T(\boldsymbol{x})\boldsymbol{\theta}_1 + \boldsymbol{\psi}_2^T(\boldsymbol{x})\boldsymbol{\theta}_2. \end{aligned} \tag{14}$$

Let $Q_{(2)} = \sum_{k=1,k\neq\alpha}^{p} w_{\alpha k}^{-1}\theta_{\alpha k}Q_{\alpha k}$. Plugging $\hat{g}(\boldsymbol{x}_i)$ and keeping terms involving $\boldsymbol{\theta}_2$ only, (12) reduces to

$$\log\left\{\frac{1}{n}\sum_{i=1}^{n} e^{-\boldsymbol{\phi}_i^T\boldsymbol{d}-\boldsymbol{\psi}_{1i}^T\boldsymbol{\theta}_1-\boldsymbol{\psi}_{2i}^T\boldsymbol{\theta}_2}\right\} + \boldsymbol{b}_{\boldsymbol{\psi}_2}^T\boldsymbol{\theta}_2 + \frac{\lambda_1}{2}\boldsymbol{c}^T Q_{(2)}\boldsymbol{c} + \lambda_2\boldsymbol{w}^T\boldsymbol{\theta}_2 \tag{15}$$

subject to $\boldsymbol{\theta}_2 \geq 0$, where $\boldsymbol{\psi}_{1i} = \boldsymbol{\psi}_1(\boldsymbol{x}_i)$, $\boldsymbol{\psi}_{2i} = \boldsymbol{\psi}_2(\boldsymbol{x}_i)$, and $\boldsymbol{b}_{\boldsymbol{\psi}_2} = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha}\boldsymbol{\psi}_2(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha$. Furthermore, the constraint minimization problem (15) is equivalent to

$$A_2(\boldsymbol{\theta}_2) = \log\left\{\frac{1}{n}\sum_{i=1}^{n} e^{-\boldsymbol{\phi}_i^T\boldsymbol{d}-\boldsymbol{\psi}_{1i}^T\boldsymbol{\theta}_1-\boldsymbol{\psi}_{2i}^T\boldsymbol{\theta}_2}\right\} + \boldsymbol{b}_{\boldsymbol{\psi}_2}^T\boldsymbol{\theta}_2 + \frac{\lambda_1}{2}\boldsymbol{c}^T Q_{(2)}\boldsymbol{c} \tag{16}$$

subject to $\boldsymbol{\theta}_2 \geq 0$ and $\boldsymbol{w}^T\boldsymbol{\theta}_2 \leq M$ for some constant M, where M controls the sparsity in $\boldsymbol{\theta}_2$. We note that $A_2(\boldsymbol{\theta}_2)$ is a convex function of $\boldsymbol{\theta}_2$ (see Appendix C for a brief proof). We solve (16) iteratively using quadratic programming. We apply K-fold cross-validation or BIC method to select M. Implementation details can be found in Appendix B.2.

3.3 Algorithm

We summarize the whole algorithm as follows. A parameter with superscript (t) denotes its value at the tth iteration.

Algorithm 1

Input: Data frame X containing n observations with p dimensions.
Output: Estimated $\boldsymbol{c}$, $\boldsymbol{d}$, $\boldsymbol{\theta}_2$, and the neighborhood set $nb_G(\alpha)$.

1: **Initialization** $\boldsymbol{\theta}_2^{(1)} = \boldsymbol{\theta}_{2,0}$, $\boldsymbol{\theta}_2^{(0)} = \mathbf{0}$, and $t = 1$.
2: **while** $||\boldsymbol{\theta}_2^{(t)} - \boldsymbol{\theta}_2^{(t-1)}||_2/(||\boldsymbol{\theta}_2^{(t-1)}||_2 + 10^{-6}) \geq \varepsilon$ or $t = 1$ **do**:
3: Fix $\boldsymbol{\theta}_2^{(t)}$,

$$\boldsymbol{c}^{(t)}, \boldsymbol{d}^{(t)} \leftarrow \underset{\boldsymbol{c},\boldsymbol{d}}{\operatorname{argmin}}\, A_1(\boldsymbol{c}, \boldsymbol{d})$$

with tuning parameters $\lambda_1^{(t)}$ and $\boldsymbol{\theta}_1^{(t)}$ selected by the ACV method.
4: Fix $\boldsymbol{d}^{(t)}$, $\boldsymbol{c}^{(t)}$, $\lambda_1^{(t)}$, and $\boldsymbol{\theta}_1^{(t)}$,

$$\boldsymbol{\theta}_2^{(t+1)} \leftarrow \underset{\boldsymbol{\theta}_2}{\operatorname{argmin}}\, A_2(\boldsymbol{\theta}_2),$$

subject to $\boldsymbol{\theta}_2 \geq 0$ and $\boldsymbol{w}^T\boldsymbol{\theta}_2 \leq M^{(t)}$ where the tuning parameter $M^{(t)}$ is selected by K-fold cross-validation or BIC method.
5: $t \leftarrow t + 1$
6: **end while**

More implementation details can be found in Appendix B, including the initialization of $\boldsymbol{\theta}_2$, the convergence criterion, and the selection of M.

4. Theoretical Analysis

In this section, we study the theoretical properties of the proposed method. Following similar steps and under the same regularity conditions as Gu (2013), we derive the convergence rate for the conditional density estimate $\hat{g}$ subject to both L_1 and L_2 penalties. In addition, we derive the convergence rate for interactions in the SS ANOVA decomposition, which is new and important for edge detection.

Let $f_0(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}}) = e^{g_0(\boldsymbol{x})}\rho(\boldsymbol{x})$ be the true conditional density to be estimated. Let $g = g^{(1)} + g^{(2)}$ where $g^{(1)} = \sum_{j=1}^{p} \eta_j$ and $g^{(2)} = \sum_{k\neq\alpha} \eta_{\alpha k}$ are main effects and interactions respectively. Denote $\hat{g}$ as the minimizer of (9). Define

$$
\begin{aligned}
V^*(h_1, h_2) &= \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) \int_{\mathcal{X}_\alpha} h_1(\boldsymbol{x}) h_2(\boldsymbol{x}) \rho(\boldsymbol{x}) dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}}, \\
J_1(h_1, h_2) &= \sum_{j=1}^{p} \theta_j^{-1} \int_{\mathcal{X}_j} (P_j h_1)(P_j h_2) dx_j, \\
J_2(h_1, h_2) &= \sum_{k\neq\alpha} w_{\alpha k} (\int_{\mathcal{X}_\alpha} \int_{\mathcal{X}_k} |h_{1,\alpha k} h_{2,\alpha k}| dx_\alpha dx_k)^{1/2}, \\
J_2^*(h_1, h_2) &= \sum_{k\neq\alpha} \theta_{\alpha k}^{-1} \int_{\mathcal{X}_\alpha} \int_{\mathcal{X}_k} h_{1,\alpha k} h_{2,\alpha k} dx_\alpha dx_k,
\end{aligned}
$$

for any functions $h_1,\ h_2 \in \mathcal{G}$, where $f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})$ is the density of $\boldsymbol{X}_{\setminus\{\alpha\}}$ on $\mathcal{X}_{\setminus\{\alpha\}} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_{\alpha-1} \times \mathcal{X}_{\alpha+1} \times \cdots \times \mathcal{X}_p$. Furthermore, we define $V^*(g) = V^*(g, g)$, $V_1(g^{(1)}) = V^*(g^{(1)})$, $V_2(g^{(2)}) = [V^*(g^{(2)})]^{1/2}$, $J_1(g^{(1)}) = J_1(g^{(1)}, g^{(1)}) = \sum_{j=1}^{p} \theta_j^{-1} ||P_j \eta_j||^2$, $J_2(g^{(2)}) = J_2(g^{(2)}, g^{(2)}) = \sum_{k\neq\alpha} w_{\alpha k} ||\eta_{\alpha k}||$, and $J_2^*(g^{(2)}) = J_2^*(g^{(2)}, g^{(2)}) = \sum_{k\neq\alpha} \theta_{\alpha k}^{-1} ||\eta_{\alpha k}||^2$.

Without loss of generality, we assume $w_{\alpha k} = 1$ in the proof, simulations, and real data examples. We note that V^*, J_1, and J_2^* are quadratic functionals. In the proof of Corollary 1 in Appendix C, it is shown that $V^*(g)$, $J_1(g^{(1)})$, and $J_2^*(g^{(2)})$ are equivalent to $||g||_2^2$, $\sum_{j=1}^{p} ||P_j \eta_j||_2^2$, and $\sum_{k\neq\alpha}^{p} ||\eta_{\alpha k}||_2^2$, respectively, where $||\cdot||_2$ is the L_2 norm. It is also shown that $V_2(g^{(2)})$ and $J_2(g^{(2)})$ are equivalent to the square root of $V^*(g^{(2)})$ and $J_2^*(g^{(2)})$. Let $V(g) = V_1(g^{(1)}) + V_2(g^{(2)})$, $J = J_1 + J_2$, and $J^*(g) = J_1(g) + J_2^*(g)$. To derive the convergence rate, we need the following conditions.

Condition 1 *V^* is completely continuous with respect to J^*.*

From Theorem 3.1 of Weinberger (1974), there exist eigenvalues γ_v of J^* with respect to V^* and the associated eigenfunctions ζ_v such that $V^*(\zeta_v, \zeta_u) = \delta_{v,u}$ and $J^*(\zeta_v, \zeta_u) = \gamma_v \delta_{v,u}$, where $0 \leq \gamma_v \uparrow \infty$ and $\delta_{v,u}$ is the Kronecker delta. Functions satisfying $J^*(g) < \infty$ can be expressed as a Fourier series expansion $g = \sum_v a_v \zeta_v$, where $a_v = V^*(g, \zeta_v)$ are the Fourier coefficients.

Condition 2 *For v sufficiently large and some $\varphi > 0$, the eigenvalues γ_v of J^* with respect to V^* satisfy $\gamma_v > \varphi v^r$ where $r > 1$.*

Consider the quadratic functional

$$\frac{1}{n}\sum_{i=1}^{n} -e^{-g_0(\boldsymbol{X}_i)}g(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha} g(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + \frac{1}{2}V^*(g-g_0) + \frac{\lambda_1}{2}J^*(g), \tag{17}$$

and denote the minimizer of (17) as $\tilde{g}$. Plugging the Fourier series expansions $g = \sum_v a_v\zeta_v$ and $g_0 = \sum_v a_{v,0}\zeta_v$ into (17), $\tilde{g}$ has Fourier coefficients $\tilde{a}_v = (\kappa_v + a_{v,0})/(1+\lambda_1\gamma_v)$, where $\kappa_v = n^{-1}\sum_{i=1}^{n}\{e^{-g_0(\boldsymbol{X}_i)}\zeta_v(\boldsymbol{X}_i) - \int_{\mathcal{X}_\alpha}\zeta_v(\boldsymbol{x})\rho(\boldsymbol{x})dx_\alpha\}$. It is not difficult to verify that $E(\kappa_v) = 0$ and $E(\kappa_v^2) \le n^{-1}\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}\zeta_v^2(\boldsymbol{x})e^{-g_0(\boldsymbol{x})}\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}}$.

Condition 3 *For some $c_1 < \infty$, $e^{-g_0} < c_1$.*

Under Condition 3, noting that $V^*(\zeta_v) = \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}\zeta_v^2(\boldsymbol{x})\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}} = 1$ by the definition of V^* and ζ_v, we have $E(\kappa_v^2) \le n^{-1}c_1$.

Condition 4 *For g in a convex set B_0 around g_0 containing $\hat{g}$ and $\tilde{g}$, $c_2 < e^{g_0-g} < c_3$ holds uniformly for some $0 < c_2 < c_3 < \infty$.*

Condition 5 *For any $u, v = 1, 2, \cdots$, $\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}\zeta_v^2\zeta_u^2 e^{-g_0}\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}} < c_4$ for some $c_4 < \infty$.*

Conditions 1-5 are common assumptions for convergence rate analysis of the SS ANOVA estimates, which were also made in Gu (2013). Condition 2 states that the growth rate of the eigenvalues γ_v is at v^r, which controls how fast λ_1 approaches zero. Many commonly used smoothing spline models, including tensor products of cubic splines, thin-plate splines, and spherical splines, satisfy Conditions 1 and 2. See Chapter 9 in Gu (2013) for examples. Condition 4 bounds e^{g_0-g} at g in a convex set B_0 around g_0. Condition 5 requires a bounded fourth moment of ζ_v.

We consider metrics $V^* + \lambda_1 J^*$ and $V + \lambda_1 J$. Let $Y > 0$, we denote $X = O_p(Y)$ if $P(|X| > CY) \to 0$ for some constant $C < \infty$, and denote $X = o_p(Y)$ if $P(|X| > \epsilon Y) \to 0$ for $\forall \epsilon > 0$.

Theorem 1 *Assume $\sum_v \gamma_v^l a_{v,0}^2 < \infty$ for some $l \in [1, 2]$. Under Conditions 1-5, for some $r > 1$, as $\lambda_1 \to 0$ and $n\lambda_1^{2/r} \to \infty$,*

$$(V^* + \lambda_1 J^*)(\hat{g} - g_0) = O_p(n^{-1}\lambda_1^{-1/r} + \lambda_1^l).$$

Theorem 2 *Under the conditions in Theorem 1,*

$$(V + \lambda_1 J)(\hat{g} - g_0) = O_p(n^{-1/2}\lambda_1^{-1/2r} + \lambda_1^{l/2}).$$

Corollary 1 *Assume conditions in Theorem 2 hold,* $0 < c_5 < \rho(\boldsymbol{x}) < c_6$ *and* $0 < c_7 < f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) < c_8$ *for some positive constants* c_5*,* c_6*,* c_7*, and* c_8*, we have*

$$||\hat{\eta}_{\alpha k} - \eta_{0\alpha k}||_2 = O_p(n^{-1/2}\lambda_1^{-1/2r} + \lambda_1^{l/2}), \quad k \neq \alpha,\ k = 1, \cdots, p,$$

where $\eta_{0\alpha k}$ *are two-way interactions in the true function* g_0*.*

We note that $V + \lambda_1 J$ and $V^* + \lambda_1 J^*$ are associated with the L_2 norm and its square, respectively. Consequently, the convergence rate in Theorem 2 is the square root of the rate in Theorem 1. Corollary 1 holds because V_2 and J_2 associated with two-way interactions are equivalent to the L_2 norm. Consequently, two-way interactions under the L_2 norm have the same convergence rate as that in Theorem 2. We only show the convergence rate for interactions in Corollary 1 since we are mainly interested in edge selection. Proofs of all theoretical results are in Appendix C.

5. Simulation Results

We conduct simulations to evaluate the performance of the proposed method and compare it with some existing methods. We consider four scenarios: multivariate Gaussian, multivariate skewed Gaussian, a directed acyclic graph, and a Gaussian-Bernoulli mixed graphical model.

In implementing the proposed method, we estimate the conditional density for each continuous variable on the data range and transform the data into $[0, 1]$. We construct an SS ANOVA model using the tensor product of cubic spline models. Specifically, let $\mathcal{H}^{(j)} = W_2^2[0, 1]$ where

$$W_2^2[0,1] = \left\{ f : f, f' \text{are absolutely continuous}, \int_0^1 (f'')^2 dx < \infty \right\} \tag{18}$$

is the Sobolev space for cubic spline models. Each $\mathcal{H}^{(j)}$ can be decomposed as $\mathcal{H}^{(j)} = \{1_{(j)}\} \oplus \mathcal{H}_{(j)}$ and $\mathcal{H}_{(j)} = \mathcal{H}_{(j)}^0 \oplus \mathcal{H}_{(j)}^1$, where $\mathcal{H}_{(j)}^0$ and $\mathcal{H}_{(j)}^1$ are RKHS's with reproducing kernels $R_j^0(x, z) = k_1(x)k_1(z)$ and $R_j^1(x, z) = k_2(x)k_2(z) - k_4(|x - z|)$ respectively, $k_1(x) = x - 0.5$, $k_2(x) = \frac{1}{2}(k_1^2(x) - \frac{1}{12})$, and $k_4(x) = \frac{1}{24}(k_1^4(x) - \frac{k_1^2(x)}{2} + \frac{7}{240})$. SS ANOVA decomposition of $\bigotimes_{j=1}^{p} \mathcal{H}^{(j)}$ can then be constructed based on these decompositions. More details can be found in Wang (2011). In all simulations and real data applications, when using the pseudo log-likelihood method, we set

$$\rho(x_\alpha, \boldsymbol{x}_{\setminus\{\alpha\}}) = \frac{\phi((x_\alpha - \mu(\boldsymbol{x}_{\setminus\{\alpha\}}))/\sigma)}{\Phi((1 - \mu(\boldsymbol{x}_{\setminus\{\alpha\}}))/\sigma) - \Phi((-\mu(\boldsymbol{x}_{\setminus\{\alpha\}}))/\sigma)}, \tag{19}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and cumulative distribution functions, and $\mu(\cdot)$ and σ are estimated by fitting a nonparametric regression model in model space (4) with covariates $\boldsymbol{x}_{\setminus\{\alpha\}}$. More estimation details can be found in Chapter 3 of Gu (2013). We select the tuning parameter M using the 5-fold cross-validation method in all simulations.

For the first three scenarios where all variables are continuous, we compare the proposed method with four existing parametric and semiparametric methods: space (Sparse

PArtial Correlation Estimation) (Peng et al., 2009), QUIC (QUadratic Inverse Covariance estimation) (Hsieh et al., 2011), nonparanormal (NPN) (Liu et al., 2009), and SpaCE JAM (Voorman et al., 2014). Due to memory constraints, we will not compare the proposed method with the nonparametric joint density estimation method in Gu et al. (2013).

The space method assumes that $\mathrm{E}(\boldsymbol{X}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{X}) = \Sigma$. Denote the precision matrix $\Omega = \Sigma^{-1} = (\sigma^{ij})_{p\times p}$ and $\rho^{ij} = -\sigma^{ij}/\sqrt{\sigma^{ii}\sigma^{jj}}$ as the partial correlation between X_i and X_j. Denote $\boldsymbol{x}_{(i)} = (x_{1,i}, \cdots, x_{n,i})^T$ as the vector of n observations on the ith variable, $i = 1, \cdots, p$. Peng et al. (2009) solved the following regularization problem for edge selection

$$\frac{1}{2}\Big(\sum_{i=1}^{p} w_i ||\boldsymbol{x}_{(i)} - \sum_{j\neq i} \rho^{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\boldsymbol{x}_{(j)}||^2\Big) + \lambda \sum_{1\leq i<j\leq p} |\rho^{ij}|, \tag{20}$$

where w_i's are non-negative weights. We implement the space method using the R package `space` with weights $w_i = 1$ and tuning parameter λ selected by the 5-fold cross-validation method (Lafit et al., 2019).

The QUIC method assumes that $\boldsymbol{X}$ is multivariate Gaussian and learns the precision matrix Ω by solving the following penalized negative log-likelihood

$$-\log\det(\Omega) + \mathrm{tr}(S\Omega) + \lambda||\Omega||_1, \tag{21}$$

where $||\cdot||_1$ is the L_1 penalty, S is the sample covariance matrix, and λ is the tuning parameter. We implement the QUIC method using the R package `QUIC` and select λ using the BIC method.

The NPN method assumes that there exists some monotone functions $f_1, \cdots, f_p$ such that $f(\boldsymbol{X}) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ where $f(\boldsymbol{X}) = (f_1(X_1), \cdots, f_p(X_p))^T$. The NPN is a semiparametric model since it consists of parameters $\boldsymbol{\mu}$ and Σ and nonparametric transformations f's. The graphical lasso is applied to the transformed data to estimate the undirected graph. Estimation details were given in Liu et al. (2009). We use the R package `huge` to implement the NPN method with the tuning parameter selected by an extended BIC score (Foygel and Drton, 2010).

The SpaCE JAM method models the conditional mean using additive models: $\mathrm{E}(X_j|\boldsymbol{X}_{\setminus\{j\}}) = \sum_{k\neq j} f_{jk}(X_k)$ where $f_{jk}(\cdot)$ belongs to a functional space $\mathcal{F}$ (Voorman et al., 2014). The functions f_{jk} are estimated as the minimizers of the following least squares with a group lasso type penalty:

$$\mathrm{argmin}_{f_{jk}\in\mathcal{F}} \left\{ \frac{1}{2n}\sum_{j=1}^{p} ||\boldsymbol{x}_{(j)} - \sum_{k\neq j} \boldsymbol{s}_{jk}||_2^2 + \lambda \sum_{k>j} \left(||\boldsymbol{s}_{jk}||_2^2 + ||\boldsymbol{s}_{kj}||_2^2\right)^{1/2} \right\}, \tag{22}$$

where $\boldsymbol{s}_{jk} = (f_{jk}(x_{1,k}), \cdots, f_{jk}(x_{n,k}))^T$ and $\boldsymbol{s}_{kj} = (f_{kj}(x_{1,j}), \cdots, f_{kj}(x_{n,j}))^T$. We implement the SpaCE JAM method using the R package `spacejam` (Voorman et al., 2014) with cubic basis functions for non-linear conditional relationships among variables. The tuning parameter λ is selected by the BIC method.

The last scenario comes from Chen et al. (2015), where half of the variables are Gaussian and half are Bernoulli. Chen et al. (2015) assumed a parametric mixed graphical model

where each node's conditional distribution is in the exponential family. Specifically, they considered conditional densities of the form

$$f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}}) = \exp\left\{h_\alpha(x_\alpha,\boldsymbol{\beta}_\alpha) + \sum_{k\neq\alpha}\gamma_{\alpha k}x_\alpha x_k - D_\alpha(\varpi_\alpha(\boldsymbol{x}_{\setminus\{\alpha\}},\Gamma_\alpha,\boldsymbol{\beta}_\alpha))\right\}, \tag{23}$$

where h_α is a known function of x_α with parameters $\boldsymbol{\beta}_\alpha$ and ϖ_α is a known function of $\boldsymbol{x}_{\setminus\{\alpha\}}$ with parameters $\Gamma_\alpha = (\gamma_{\alpha 1},\cdots,\gamma_{\alpha(\alpha-1)},\gamma_{\alpha(\alpha+1)},\cdots,\gamma_{\alpha p})^T$ and $\boldsymbol{\beta}_\alpha$. Chen et al. (2015) selected the neighborhood set by maximizing the following penalized log-likelihoods for each node:

$$\underset{\Gamma_\alpha,\boldsymbol{\beta}_\alpha}{\arg\min} \quad -l_\alpha(\Gamma_\alpha,\boldsymbol{\beta}_\alpha;X) + \lambda||\Gamma_\alpha||_1, \tag{24}$$

where l_α is the log-likelihood function. We refer to the method in Chen et al. (2015) as the CEF (Conditional Exponential Family) method. We implement the CEF method using author's R codes deposited at `https://github.com/ChenShizhe/MixedGraphicalModels`. We select the tuning parameter λ using the BIC method.

We note that space and CEF are neighborhood selection methods while QUIC, NPN, and SpaCE JAM are global methods. To evaluate the performance of edge detection, we compute three criteria: specificity (SPE), sensitivity (SEN), and F_1 scores, which are defined as follows:

$$\text{SPE} = \frac{\text{TN}}{\text{TN}+\text{FP}}, \qquad \text{SEN} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \qquad \text{F}_1 = \frac{2\text{TP}}{2\text{TP}+\text{FN}+\text{FP}},$$

where TP, TN, FP, and FN are the numbers of true positives, true negatives, false positives and false negatives.

We set dimension $p = 20$ and consider two sample sizes $n = 150$ and $n = 300$. All simulations are repeated for 100 times.

5.1 Multivariate Gaussian

In this section, we generate data from Gaussian distributions with different precision matrices. We first use `huge.generator` function to randomly generate a $p \times p$ sparse precision matrix Ω, where the probability p_{off} of the off-diagonal elements being nonzero is equal to 0.2 or 0.4. Then, we generate n i.i.d. samples $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n$ from $\mathcal{N}(\boldsymbol{0},\Omega^{-1})$. We apply the proposed method and compare its performance with the space, QUIC, NPN, and SpaCE JAM methods.

Table 1 presents averages and standard deviations of the sensitivity, specificity, and F_1 score. In general, the performances of all methods are better for the larger sample size. In most settings, all methods perform better when the precision matrix is sparser (i.e. $p_{\text{off}} = 0.2$). Different methods have different trade-offs between sensitivity and specificity. Overall, the NPN method has inferior performance compared to other methods, which is expected since the true distribution is Gaussian. In general, the SpaCE JAM performs better than QUIC in specificity and F_1 score. This result agrees with the observations of Voorman et al. (2014) that SpaCE JAM tends to outperform the NPN and graphical lasso methods. The space method has a similar performance as the SpaCE JAM. Unexpectedly,

even in this Gaussian case, the proposed method has larger sensitivities and F_1 scores and reasonable specificities compared to other methods. Therefore, the proposed method is efficient in edge detection and performs better with a more balanced trade-off between specificity and sensitivity even under this multivariate Gaussian scenario.

	Proposed Method			space			QUIC			NPN			SpaCE JAM		
	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1
$p_{\text{off}} = 0.2$															
n=150	0.912	0.968	0.834	0.986	0.654	0.762	0.768	0.964	0.666	0.820	0.751	0.521	0.939	0.798	0.776
	(0.028)	(0.037)	(0.050)	(0.011)	(0.096)	(0.075)	(0.036)	(0.034)	(0.042)	(0.108)	(0.402)	(0.281)	(0.035)	(0.147)	(0.068)
n=300	0.929	0.998	0.870	0.989	0.869	0.907	0.813	0.982	0.712	0.762	0.995	0.668	0.945	0.954	0.875
	(0.026)	(0.008)	(0.046)	(0.01)	(0.059)	(0.04)	(0.032)	(0.025)	(0.04)	(0.042)	(0.016)	(0.042)	(0.022)	(0.041)	(0.037)
$p_{\text{off}} = 0.4$															
n=150	0.866	0.815	0.807	0.969	0.461	0.599	0.668	0.797	0.691	0.793	0.617	0.474	0.945	0.508	0.608
	(0.040)	(0.066)	(0.046)	(0.013)	(0.062)	(0.058)	(0.047)	(0.058)	(0.030)	(0.142)	(0.411)	(0.312)	(0.025)	(0.135)	(0.126)
n=300	0.883	0.968	0.903	0.961	0.564	0.681	0.689	0.827	0.717	0.673	0.951	0.706	0.905	0.704	0.727
	(0.042)	(0.027)	(0.028)	(0.015)	(0.058)	(0.049)	(0.043)	(0.046)	(0.026)	(0.039)	(0.036)	(0.021)	(0.031)	(0.128)	(0.076)

Table 1: Averages and standard deviations (in parentheses) of specificity (SPE), sensitivity (SEN), and F_1 score for the multivariate Gaussian scenario.

5.2 Multivariate Skewed Gaussian

In this section, we consider the scenario when $\boldsymbol{X}$ follows a multivariate skewed Gaussian distribution with density function (Azzalini and Valle, 1996)

$$f(\boldsymbol{x}) = 2\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)\Phi(\boldsymbol{a}^T\boldsymbol{x}), \tag{25}$$

where $\phi_p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$ is the p-dimensional normal density with mean $\boldsymbol{\mu}$ and covariance matrix Σ, $\Phi(\cdot)$ is the cumulative distribution function of the standard Gaussian distribution, and $\boldsymbol{a}$ is a p-dimensional vector that controls the skewness of the multivariate Gaussian distribution. When $\boldsymbol{a} = \boldsymbol{0}$, the distribution reduces to the multivariate Gaussian distribution. We set $\boldsymbol{a} = a\boldsymbol{1}$ and consider two choices of a: $a = 1$ and $a = 4$, where $\boldsymbol{1}$ is a p-dimensional vector of all ones. We let $\boldsymbol{\mu} = 0.5\boldsymbol{1}$ and randomly generate Σ^{-1} as a $p \times p$ matrix, where the probability of the off-diagonal elements being nonzero equals 0.4. True edges correspond to nonzero off-diagonal elements of the precision matrix Σ^{-1}.

Table 2 presents averages and standard deviations of sensitivity, specificity, and F_1 score. All methods have better performances under the larger sample size. Again, different methods have different trade-offs between sensitivity and specificity. The space, NPN, and SpaCE JAM methods have small sensitivities and F_1 scores when $n = 150$. As expected, the proposed method has the best overall performance with significantly larger sensitivity and F_1 score and reasonable specificity.

	Proposed Method			space			QUIC			NPN			SpaCE JAM		
	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1
$a = 1$															
n=150	0.892	0.854	0.847	0.957	0.302	0.440	0.688	0.798	0.704	0.873	0.332	0.286	0.926	0.378	0.486
	(0.039)	(0.058)	(0.044)	(0.023)	(0.053)	(0.060)	(0.041)	(0.055)	(0.027)	(0.160)	(0.412)	(0.352)	(0.039)	(0.160)	(0.164)
n=300	0.927	0.976	0.937	0.949	0.470	0.608	0.718	0.827	0.737	0.668	0.932	0.769	0.845	0.728	0.739
	(0.034)	(0.022)	(0.025)	(0.027)	(0.061)	(0.061)	(0.045)	(0.047)	(0.027)	(0.049)	(0.047)	(0.024)	(0.088)	(0.133)	(0.069)
$a = 4$															
n=150	0.898	0.860	0.854	0.958	0.299	0.438	0.692	0.800	0.707	0.861	0.357	0.306	0.929	0.384	0.493
	(0.034)	(0.056)	(0.040)	(0.023)	(0.054)	(0.063)	(0.040)	(0.055)	(0.027)	(0.164)	(0.415)	(0.355)	(0.037)	(0.161)	(0.166)
n=300	0.917	0.979	0.931	0.950	0.473	0.610	0.772	0.828	0.739	0.673	0.934	0.773	0.834	0.748	0.745
	(0.034)	(0.019)	(0.026)	(0.026)	(0.060)	(0.059)	(0.041)	(0.047)	(0.027)	(0.049)	(0.045)	(0.023)	(0.103)	(0.140)	(0.067)

Table 2: Averages and standard deviations (in parentheses) of specificity (SPE), sensitivity (SEN), and F_1 score for the multivariate skewed Gaussian scenario.

5.3 Directed Acyclic Graph

It is generally difficult to construct a flexible multivariate nonparametric distribution, as discussed in Section 2 in Voorman et al. (2014). To overcome this problem, we use the same approach in Voorman et al. (2014) to generate a graphical model using a directed acyclic graph (DAG) and conditional distributions. We use the `rdag` function in the `spacejam` package to create a DAG of $\boldsymbol{X}$ and denote E_D as the directed edge set. The conditional relationships among variables can be created via $\mathrm{E}(X_j|\boldsymbol{X}_{\setminus\{j\}}) = \sum_{k\neq j} f_{jk}(X_k)$. The distribution of $\boldsymbol{X}$ is usually not a well-known multivariate distribution except for the particular case when all f_{jk}s are linear for multivariate Gaussian distribution.

We decompose $\boldsymbol{X}^T = (\boldsymbol{Y}^T, \boldsymbol{Z}^T)$ where $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are random vectors of dimensions 5 and 15 respectively. We first generate a DAG with $p = 20$ nodes and m edges selected at random from all possible $p(p-1)/2$ possible edges. We consider two choices of m: $m = 20$ and $m = 40$. Given a DAG, we generate data as follows:

$$Z_j|\{Z_k, Y_s : \{k,j\}, \{s,j\} \in E_D\} = \sum_{\{k,j\}\in E_D} f_{jk}^{(1)}(Z_k) + \sum_{\{s,j\}\in E_D} f_{js}^{(1)}(Y_s) + \epsilon_j$$
$$Y_j|\{Y_k : \{k,j\} \in E_D\} = \sum_{\{k,j\}\in E_D} f_{jk}^{(2)}(Y_k) + \epsilon_j,$$

where ϵ_j's are i.i.d. random noises from the standard normal distribution, $f_{jk}^{(1)}(t) = b_{jk,1}^{(1)} t$ with $b_{jk,1}^{(1)}$ generated from the standard Gaussian distribution, and $f_{jk}^{(2)}(t) = b_{jk,1}^{(2)} t + b_{jk,2}^{(2)} t^2 + b_{jk,3}^{(2)} t^3$ with $b_{jk,1}^{(2)}, b_{jk,2}^{(2)}$ and $b_{jk,3}^{(2)}$ independently generated from the Gaussian distributions with mean zero and variances 1, 0.3, and 0.1, respectively.

Simulation results are shown in Table 3. The performances of all methods are better when the sample size is larger. Different methods have different trade-offs between sensitivity and specificity. Since data are generated according to a model assumed by the SpaCE JAM method, as expected, the SpaCE JAM method performs better in F_1 score than the space, QUIC, and NPN methods. Remarkably, in all cases, the proposed method has larger F_1 scores than all methods, including the SpaCE JAM. It is interesting to note that the denser graph (i.e. $m = 40$) reduces the sensitivity of the proposed method and the specificity of the SpaCE JAM method.

	Proposed Method			space			QUIC			NPN			SpaCE JAM		
	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1	SPE	SEN	F_1
$m = 20$															
$n = 150$	0.970	0.835	0.840	0.997	0.588	0.730	0.808	0.838	0.588	0.83	0.791	0.588	0.963	0.697	0.736
	(0.020)	(0.074)	(0.066)	(0.004)	(0.084)	(0.068)	(0.034)	(0.079)	(0.050)	(0.066)	(0.125)	(0.054)	(0.019)	(0.079)	(0.062)
$n = 300$	0.984	0.917	0.915	0.998	0.631	0.767	0.859	0.854	0.660	0.818	0.894	0.629	0.979	0.716	0.786
	(0.013)	(0.064)	(0.050)	(0.003)	(0.075)	(0.058)	(0.035)	(0.079)	(0.055)	(0.041)	(0.082)	(0.052)	(0.057)	(0.089)	(0.072)
$m = 40$															
$n = 150$	0.970	0.598	0.725	0.984	0.359	0.517	0.705	0.671	0.627	0.671	0.708	0.634	0.690	0.710	0.643
	(0.020)	(0.064)	(0.05)	(0.012)	(0.039)	(0.041)	(0.039)	(0.053)	(0.031)	(0.042)	(0.060)	(0.031)	(0.111)	(0.112)	(0.044)
$n = 300$	0.985	0.685	0.800	0.982	0.430	0.587	0.740	0.673	0.645	0.707	0.724	0.662	0.654	0.814	0.692
	(0.014)	(0.067)	(0.049)	(0.013)	(0.054)	(0.050)	(0.046)	(0.049)	(0.036)	(0.045)	(0.048)	(0.038)	(0.07)	(0.051)	(0.031)

Table 3: Averages and standard deviations (in parentheses) of specificity (SPE), sensitivity (SEN), and F_1 score for the directed acyclic graph scenario.

5.4 Gaussian-Bernoulli Mixed Graphical Model

In this section, we consider a mixed graphical model used in Section 6.1 of Chen et al. (2015). The graph used to generate the data is shown in Figure 1 of Chen et al. (2015). Specifically, there are m Gaussian nodes labeled as $1, \cdots, m$ and m Bernoulli nodes labeled as $m + 1, \cdots, 2m$. For $j = 1, \cdots, m$, the jth and $(j + m)$th node are connected to its adjacent nodes of the same type, and the jth node and the $(j + m)$th node are connected to each other. Consider the following model

$$f(\boldsymbol{x}) \propto \exp\left\{\sum_{j=1}^{p} h_j(x_j) + \frac{1}{2}\sum_{k=1}^{p}\sum_{j\neq k}\gamma_{jk}x_j x_k\right\}, \tag{26}$$

where h_j is the node potential and γ_{jk} are edge potentials. The edge potentials γ_{jk} and γ_{kj} are generated as

$$\gamma_{jk} = \gamma_{kj} = y_{jk}r_{jk}, \quad P(y_{jk} = 1) = P(y_{jk} = -1) = 0.5, \quad r_{jk} \sim \text{Unif}(0.3, 0.6),$$

and $\gamma_{jk} = \gamma_{kj} = 0$ if $(j,k) \notin E$. Gibbs sampling is employed to sample data from (26). In this simulation scenario, we compare the proposed method with the CEF method only since it performed better than other existing methods for mixed data (Chen et al., 2015).

	Proposed Method			CEF		
	SPE	SEN	F_1	SPE	SEN	F_1
$n = 150$	0.897	0.752	0.656	0.947	0.450	0.509
	(0.021)	(0.053)	(0.044)	(0.020)	(0.015)	(0.035)
$n = 300$	0.900	0.804	0.675	0.934	0.467	0.504
	(0.022)	(0.042)	(0.039)	(0.021)	(0.017)	(0.032)

Table 4: Averages and standard deviations (in parentheses) of specificity (SPE), sensitivity (SEN), and F_1 score for the Gaussian-Bernoulli mixed graphical model.

Table 4 shows that the proposed and the CEF methods have different trade-offs between sensitivity and specificity. The proposed method has better sensitivity, while the CEF method has better specificity. The proposed method has slightly better F_1 scores.

6. Applications

We illustrate our neighborhood selection method using two real datasets. Section 6.1 applies our method to Arabidopsis Thaliana gene expression data and compares the estimated graph with those from space, QUIC, NPN, and SpaCE JAM. In addition, we present a diagnostic procedure for some existing methods. Section 6.2 illustrates our method using a dataset with mixed data types.

6.1 Isoprenoid Gene Network in Arabidopsis Thaliana

In this section, we consider the gene expression data for *Arabidopsis thaliana*, an important plant species in molecular biology and genetics studies. There are $n = 118$ observations of Affymetrix GeneChip microarrays in the dataset, where a subset of $p = 39$ genes from the isoprenoid pathway is selected for analysis. The dataset was introduced in Wille et al. (2004) and was downloaded at `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC545783/`. Lafferty et al. (2012) analyzed this dataset using the nonparanormal method.

All observations are preprocessed by log-transformation and standardization as in Lafferty et al. (2012). Using the proposed method, we build a graph for all 39 gene expression levels and compare its structure with those from space, QUIC, NPN, and SpaCE JAM. Wille et al. (2004) stated that the Gaussian graphical model selection with the BIC choice of the tuning parameter usually detects too many edges for biologically-relevant analysis. Therefore, we limit the number of edges in the graph by controlling the regularization parameters as in Lafferty et al. (2012). Specifically, we tune M such that the number of edges $|E| = 52$. Similarly, by tuning the regularization parameters in space, QUIC, NPN, and SpaCE JAM, we select the graphs with the same number of edges $|E| = 52$.

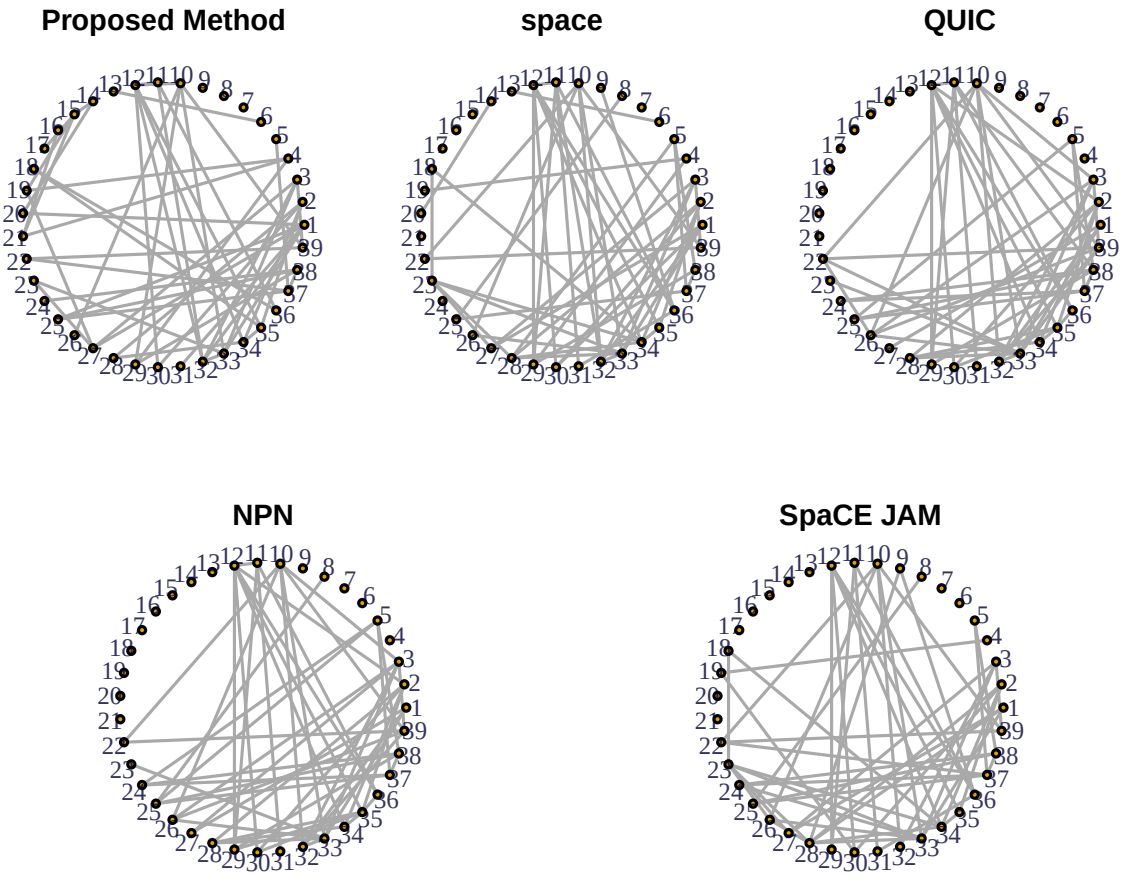

Figure 1: The estimated graph with 52 edges from the proposed (*top left*), the space (*top middle*), the QUIC (*top right*), the NPN (*bottom left*), and the SpaCE JAM (*bottom right*) methods.

Figure 1 presents graphs with $|E| = 52$ for all methods. These five graphs have some common edges, for example, edges `1-27`, `1-33`, `2-28`, `2-30`, `2-34`, `2-35`, `3-32`, `3-33`, `3-39`, `5-37`, `10-26`, `10-33`, `10-39`, `11-36`, `12-29`, `12-30`, `12-34`, `12-35`, `22-39`, `23-33`, `25-37`, `28-34`, `34-35`, and `37-38`. There are also some interesting differences. For instance, only our proposed method detects edge `16-21`. We now describe a general diagnostic procedure that explains why other methods miss this edge.

We first extend the squared error projection in Gu (2013) for diagnostics on any subspaces of $\mathcal{M}_\alpha$. Let

$$\tilde{V}(\hat{g} - g) = \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) \int_{\mathcal{X}_\alpha} \left\{ (\hat{g} - g)(\boldsymbol{x}) - \int_{\mathcal{X}_\alpha} (\hat{g} - g)(\boldsymbol{x}) \rho(\boldsymbol{x}) \right\}^2 \rho(\boldsymbol{x}) dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}} \tag{27}$$

where $\hat{g} \in \mathcal{M}_\alpha \ominus \{1\}$. We remove the constant functions from the model space since they are not relevant to the diagnostics on interactions. $\tilde{V}(\hat{g} - g)$ can be treated as a proxy of

the symmetrized Kullback-Leibler distance (Gu, 2013). For any decomposition $\mathcal{M}_\alpha \ominus \{1\} = \mathcal{M}_\alpha^0 \oplus \mathcal{M}_\alpha^1$, the squared error projection of $\hat{g}$ in $\mathcal{M}_\alpha^0$ is defined as $\tilde{g} = \underset{g \in \mathcal{M}_\alpha^0}{\arg\min} \left\{ \tilde{V}(\hat{g} - g) \right\}$. It can be shown that $\tilde{V}(\hat{g} - g_u) = \tilde{V}(\hat{g} - \tilde{g}) + \tilde{V}(\tilde{g} - g_u)$ when $g_u = -\log \rho(\mathbf{x}) \in \mathcal{M}_\alpha^0$. The ratio $\tilde{V}(\hat{g} - \tilde{g})/\tilde{V}(\hat{g} - g_u)$ represents the contribution of functions in subspace $\mathcal{M}_\alpha^1$ which can be dropped when the ratio is small (Gu et al., 2013).

Now we apply the diagnostic procedure to explain why our proposed method detects edge `16-21`, which is missed by other methods. Note that the interaction space $\mathcal{H}_{(\alpha k)} = \mathcal{H}_{(\alpha k)}^{(0)} \oplus \mathcal{H}_{(\alpha k)}^{(1)} \oplus \mathcal{H}_{(\alpha k)}^{(2)} \oplus \mathcal{H}_{(\alpha k)}^{(3)}$ where $\mathcal{H}_{(\alpha k)}^{(0)} = \mathcal{H}_{(\alpha)}^0 \otimes \mathcal{H}_{(k)}^0$, $\mathcal{H}_{(\alpha k)}^{(1)} = \mathcal{H}_{(\alpha)}^0 \otimes \mathcal{H}_{(k)}^1$, $\mathcal{H}_{(\alpha k)}^{(2)} = \mathcal{H}_{(\alpha)}^1 \otimes \mathcal{H}_{(k)}^0$, and $\mathcal{H}_{(\alpha k)}^{(3)} = \mathcal{H}_{(\alpha)}^1 \otimes \mathcal{H}_{(k)}^1$ correspond to linear-linear, linear-smooth, smooth-linear, and smooth-smooth interactions (Wang, 2011). The QUIC and space are special cases with $\eta_{\alpha k} \in \mathcal{H}_{(\alpha k)}^{(0)}$, and the SpaCE JAM is a special cases with $\eta_{\alpha k} \in \mathcal{H}_{(\alpha k)}^{(0)} \oplus \mathcal{H}_{(\alpha k)}^{(1)}$. Therefore, for diagnostics of QUIC and SpaCE JAM methods, we consider the contribution of $\mathcal{M}_\alpha^1 = \mathcal{M}_\alpha \ominus \{1\} \ominus \mathcal{H}_{(\alpha k)}^{(0)}$ and the contribution of $\mathcal{M}_\alpha^1 = \mathcal{M}_\alpha \ominus \{1\} \ominus \mathcal{H}_{(\alpha k)}^{(0)} \ominus \mathcal{H}_{(\alpha k)}^{(1)}$, respectively. For edge `16-21`, we have $\tilde{V}(\hat{g} - \tilde{g})/\tilde{V}(\hat{g} - g_u) = 0.352$ for QUIC and $\tilde{V}(\hat{g} - \tilde{g})/\tilde{V}(\hat{g} - g_u) = 0.340$ for SpaCE JAM, respective. These non-ignorable contributions suggest that the assumptions of the QUIC and SpaCE JAM methods are likely violated.

6.2 Conditional Dependence Among Demographic, Clinical, Laboratory and Treatment Variables of Hemodialysis Patients

In this section, we illustrate the application of the proposed methods to mixed binary and continuous variables using a data set collected from hemodialysis patients. The data include patients who received dialysis treatments during 2010-2014 and stayed at Fresenius Medical Care - North America throughout their treatments. To reduce heterogeneity, we include $n = 2932$ non-diabetic and non-Hispanic patients who used arteriovenous fistula for dialysis access and survived longer than two years. We use the averages of measurements in the second year of dialysis for analysis. We consider the following 23 variables: demographic variables including `race` (white and non-white) and `gender` (male and female); clinical variables including `height` (cm), `weight` (kg), `sbp` (systolic blood pressure, mmHg), `dbp` (diastolic blood pressure, mmHg), and `temp` (temperature, Celsius); laboratory variables including `albumin` (g/dL), `ferritin` (ng/mL), `hgb` (hemoglobin, g/dL), `lymphocytes` (%), `neutrophils` (%), `nlr` (neutrophils to lymphocytes ratio, unitless), `sna` (serum sodium concentration, mEq/L), `wbc` (white blood cell, 1000/mc); and treatment variables including `qb` (blood flow, mL/min), `qd` (dialysis flow, mL/min), `saline` (mL), `olc` (on-line clearance, unitless), `idwg` (interdialytic weight gain, kg), `ufv` (ultrafiltration volume, L), `ufr` (ultrafiltration rate, mL/hr/kg), and `epodose` (erythropoietin dose, unit).

We have 2 binary variables, `race` and `male`, and 21 continuous variables. We apply the logistic regression approach described in the Supplement to estimate the conditional density of each binary variable and the pseudo log-likelihood to estimate the conditional density of each continuous variable. We apply the BIC method to select the tuning parameter M. The left panel in Figure 2 shows the estimated graph which contains some of the expected dependencies between variables such as `gender` and `height`, `weight` and `height`, and `sbp` and `dbp`. The link between `ufv` and `idwg` is also well-known (Uduagbamen et al., 2021).

Many other edges corroborate with existing literature. For example, anemia is a common complication of dialysis patients, and its management is a major challenge. A central aim of anemia management is to maintain patients' hemoglobin levels consistently within a target range. Erythropoietin has been used to raise hemoglobin levels, which is revealed by the edge between `epodose` and `hgb`. Serum albumin has been found to be strongly associated with erythropoietin sensitivity (Agarwal et al., 2008), which is corroborated by the edge between `epodose` and `albumin`. It has been found that black patients receive greater doses of erythropoietin than white patients (Lacson et al., 2008), which is corroborated by the edge between `epodose` and `race`. The estimated graph from our proposed method in Figure 2 (left panel) provides a holistic view of complex relationships between the demographic, clinical, laboratory, and treatment variables and helps build new theories to be tested in future studies. For comparison, we apply the CEF method to fit the Gaussian-Bernoulli model (26) with the BIC choice of the tuning parameter to this data and show the resulting graph on the right panel in Figure 2. The CEF method leads to a very dense graph.

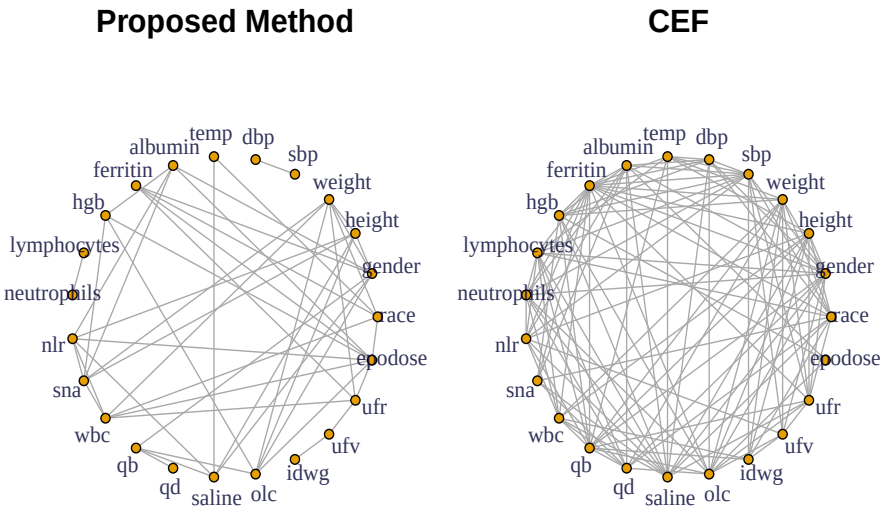

Figure 2: The estimated graph for dialysis data from the proposed (*left*) and the CEF (*right*) methods.

7. Conclusion

This paper develops a fully nonparametric method for neighborhood selection in pairwise graphical models. Since the range of each random variable is an arbitrary set, the proposed method provides a unified framework for mixed data types. The proposed SS ANOVA models are more general than existing parametric and semiparametric models. We de-

velop penalized log-likelihood and pseudo log-likelihood methods with an L_1 penalty to select edges. As illustrated in Section 6.1, in addition to providing more flexible alternatives, the proposed method also serves as a new diagnostic tool for existing graphical models. We establish convergence rates of the conditional density function estimate and interaction components in the SS ANOVA decomposition. Simulation results show that the proposed method is efficient in edge detection and performs well under Gaussian and non-Gaussian situations. Applications to real data indicate that the proposed method could detect edges that may provide new perspectives for researchers. We note that as a nonparametric method, even though it is parallelizable, the proposed method takes much longer CPU time than parametric and semiparametric methods compared in this paper.

We note that the proposed methods can be easily extended to select variables in nonparametric conditional density estimation, which has not been studied to the best of our knowledge. The proposed method can also be extended to incorporate prior knowledge of the conditional density of a node using a model-based penalty or a semiparametric model (Shi et al., 2019; Yu et al., 2020). For example, it may be known that the conditional density of X_α is close to but not necessarily a Gaussian distribution. We may consider a quintic thin-plate spline space for $\mathcal{H}^{(\alpha)}$ with a tensor sum decomposition $\mathcal{H}^{(\alpha)} = \mathcal{H}^0_{(\alpha)} \oplus \mathcal{H}^1_{(\alpha)}$, where $\mathcal{H}^0_{(\alpha)} = \{1_{(\alpha)}, x_{(\alpha)}, x^2_{(\alpha)}\}$ corresponds to the space for logistic density of a Gaussian distribution. The edge selection consistency and the control of false positives also warrant further investigation.

Acknowledgments

We would like to thank the editor and three anonymous reviewers for their insightful comments that significantly improved the manuscript. This research is partially supported by NIH R01 DK130067. We thank Fresenius Medical Care North America for providing deidentified data.

Appendix A. Penalized Log-likelihood Estimation

We describe the general penalized log-likelihood approach in Section A.1 and how it is applied for binary variables in Section A.2.

A.1 Log-likelihood Approach for Conditional Density Estimation

In this section, we describe the log-likelihood approach for estimating conditional density with L_1 penalty. For identifiability, the constant and main effects of $\boldsymbol{x}_{\setminus\{\alpha\}}$ are removed from the model space. Specifically, the model space for η in (3) is assumed to be

$$\mathcal{M}_\alpha^* = \left\{\mathcal{H}_{(\alpha)}\right\} \oplus \left\{\bigoplus_{k\neq\alpha}[\mathcal{H}_{(\alpha)} \otimes \mathcal{H}_{(k)}]\right\}. \tag{28}$$

A function $\eta \in \mathcal{M}_\alpha^*$ can be decomposed as follows:

$$\eta(\boldsymbol{x}) = \eta_j(x_\alpha) + \sum_{k\neq\alpha} \eta_{\alpha k}(x_\alpha, x_k), \tag{29}$$

where each component in (29) belongs to the corresponding subspace in (28). We further decompose $\mathcal{H}_{(\alpha)}$ as $\mathcal{H}_{(\alpha)} = \mathcal{H}_{(\alpha)}^0 \oplus \mathcal{H}_{(\alpha)}^1$ where $\mathcal{H}_{(\alpha)}^0$ is a finite-dimensional space containing functions that are not subject to the L_2 penalty. We estimate η by minimizing the following penalized log-likelihood in $\mathcal{M}_\alpha^*$:

$$l_\alpha^* + \frac{\lambda_1}{2}\left(\theta_\alpha^{-1}||P_\alpha\eta_\alpha||^2 + \sum_{k\neq\alpha} w_{\alpha k}\theta_{\alpha k}^{-1}||\eta_{\alpha k}||^2\right) + \lambda_2 \sum_{k\neq\alpha} w_{\alpha k}\theta_{\alpha k}, \tag{30}$$

where $l_\alpha^* = -n^{-1}\sum_{i=1}^n \left\{\eta(\boldsymbol{x}_i) - \log\int_{\mathcal{X}_\alpha} e^{\eta(x_\alpha, \boldsymbol{x}_{i,\setminus\{\alpha\}})}dx_\alpha\right\}$, and P_α is the projection operator onto $\mathcal{H}_{(\alpha)}^1$. Let $\boldsymbol{\theta}_2 = (\theta_{\alpha 1}, \cdots, \theta_{\alpha(\alpha-1)}, \theta_{\alpha(\alpha+1)}, \cdots, \theta_{\alpha p})^T$, $\boldsymbol{\phi}_\alpha = (\phi_{\alpha 1}, \cdots, \phi_{\alpha m_\alpha})^T$ be a vector of basis functions of $\mathcal{H}_{(\alpha)}^0$, $\xi_u(\boldsymbol{x}) = \theta_\alpha \xi_{1\alpha u}(x_\alpha) + \sum_{k=1,k\neq\alpha}^p w_{\alpha k}^{-1}\theta_{\alpha,k}\xi_{\alpha k u}(x_\alpha, x_k)$ for $u = 1, \cdots, q$, and $\boldsymbol{\xi}(\boldsymbol{x}) = (\xi_1(\boldsymbol{x}), \cdots, \xi_q(\boldsymbol{x}))^T$.

Similar to (11), the approximate solution can be represented as

$$\begin{aligned}\hat{\eta}(\boldsymbol{x}) &= \sum_{v=1}^{m_\alpha} d_v\phi_{\alpha v}(\boldsymbol{x}) + \sum_{u=1}^q c_u\left\{\theta_\alpha\xi_{1\alpha u}(x_\alpha) + \sum_{k=1,k\neq\alpha}^p w_{\alpha k}^{-1}\theta_{\alpha,k}\xi_{\alpha k u}(x_\alpha, x_k)\right\} \\ &= \boldsymbol{\phi}_\alpha^T(\boldsymbol{x})\boldsymbol{d} + \boldsymbol{\xi}^T(\boldsymbol{x})\boldsymbol{c},\end{aligned} \tag{31}$$

where $\boldsymbol{c} = (c_1, \cdots, c_q)^T$ and $\boldsymbol{d} = (d_1, \cdots, d_{m_\alpha})^T$. Plugging $\hat{\eta}(\boldsymbol{x}_i)$ in (31) into (30), we need to compute $\boldsymbol{c}$, $\boldsymbol{d}$, and $\boldsymbol{\theta}_2$ as the minimizers of

$$-\frac{1}{n}\sum_{i=1}^n(\boldsymbol{\phi}_{\alpha,i}^T\boldsymbol{d} + \boldsymbol{\xi}_i^T\boldsymbol{c}) + \frac{1}{n}\sum_{i=1}^n \log\int_{\mathcal{X}_\alpha} e^{\boldsymbol{\phi}_\alpha^T(x_\alpha, \boldsymbol{x}_{i,\setminus\{\alpha\}})\boldsymbol{d} + \boldsymbol{\xi}^T(x_\alpha, \boldsymbol{x}_{i,\setminus\{\alpha\}})\boldsymbol{c}}dx_\alpha + \frac{\lambda_1}{2}\boldsymbol{c}^T Q\boldsymbol{c} + \lambda_2\boldsymbol{w}^T\boldsymbol{\theta}_2 \tag{32}$$

subject to $\boldsymbol{\theta}_2 \geq 0$. Similar to the Algorithm in Section 3, we can update $\boldsymbol{c}$, $\boldsymbol{d}$, and $\boldsymbol{\theta}_2$ sequentially. The estimate of $\boldsymbol{\theta}_2$ provides the selection results. We skip the details here because the derivation is similar to Section 3. The algorithm can be similarly implemented as described in Appendix B. With minor changes in Conditions 1-5 and the definition of V, following similar steps in Appendix C, it can be shown that Theorem 1, Theorem 2, and Corollary 1 hold for the log-likelihood approach with the same convergence rates.

A.2 Logistic Regression for Binary Variables

In this section, we consider the penalized log-likelihood approach for the special case when x_α is a binary variable taking values 0 or 1. Consider the logit function

$$\begin{aligned}
\nu(\boldsymbol{x}) &= \log\{f(1|\boldsymbol{x}_{\setminus\{\alpha\}})/f(0|\boldsymbol{x}_{\setminus\{\alpha\}})\} \\
&= \eta(1, \boldsymbol{x}_{\setminus\{\alpha\}}) - \eta(0, \boldsymbol{x}_{\setminus\{\alpha\}}) \\
&= \eta_\alpha(1) - \eta_\alpha(0) + \sum_{j\neq\alpha}\{\eta_{\alpha j}(1, x_j) - \eta_{\alpha j}(0, x_j)\} \\
&= 2\eta_\alpha(1) + 2\sum_{j\neq\alpha}\eta_{\alpha j}(1, x_j) \\
&\triangleq \varsigma^\dagger + \sum_{j\neq\alpha}\eta_j^\dagger(x_j)
\end{aligned}$$

where the third equation comes from the SS ANOVA model (29), and fourth equation is the consequence of the side conditions: $\eta_\alpha(1)+\eta_\alpha(0) = 0$ and $\eta_{\alpha j}(1, x_j)+\eta_{\alpha j}(0, x_j) = 0$ for $j \neq \alpha$. Therefore, $\eta_{\alpha j}(x_\alpha, x_j) = 0$ (i.e. there is no edge between x_α and x_j) is equivalent to $\eta_j^\dagger(x_j) = 0$. Consequently, we can consider a logistic regression model with the following model space for the function ν,

$$\mathcal{M}_\alpha^\dagger = \{1\} \oplus \left\{\bigoplus_{j\neq\alpha}\mathcal{H}_{(j)}^\dagger\right\}, \tag{33}$$

where $\eta_j^\dagger \in \mathcal{H}_{(j)}^\dagger$.

We write the conditional density $f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}}) = \exp\{x_\alpha\nu(\boldsymbol{x}) - \log(1 + e^{\nu(\boldsymbol{x})})\}$. Then, the penalized log-likelihood function

$$-\frac{1}{n}\sum_{i=1}^n\left\{x_{\alpha,i}\nu(\boldsymbol{x}_i) - \log(1 + e^{\nu(\boldsymbol{x}_i)})\right\} + \frac{\lambda_1}{2}\sum_{j\neq\alpha}||\eta_j^\dagger||. \tag{34}$$

Similar to Zhang et al. (2011), instead of (34), we will minimize the following equivalent but more convenient form

$$-\frac{1}{n}\sum_{i=1}^n\left\{x_{\alpha,i}\nu(\boldsymbol{x}_i) - \log(1 + e^{\nu(\boldsymbol{x}_i)})\right\} + \frac{\lambda_1}{2}\sum_{j\neq\alpha}\theta_j^{-1}||\eta_j^\dagger||^2 + \lambda_2\sum_{j\neq\alpha}\theta_j. \tag{35}$$

We approximate the solution as described in Section 3. Denote the approximate solution as $\hat{\nu}(\boldsymbol{x}) = d + \sum_{u=1}^q c_u\left\{\sum_{j\neq\alpha}\theta_j\xi_{ju}(x_j)\right\} = d + \boldsymbol{\xi}^T(\boldsymbol{x})\boldsymbol{c}$, where $d \in \mathbb{R}$, $\xi_{ju}(x_j) = R_j(\tilde{x}_{u,j}, x_j)$,

and $\boldsymbol{\xi}(\boldsymbol{x}) = (\xi_1(\boldsymbol{x}), \cdots, \xi_q(\boldsymbol{x}))^T$. Then, (35) reduces to

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{x_{\alpha,i}(d+\boldsymbol{\xi}_i^T\boldsymbol{c}) - \log(1+e^{d+\boldsymbol{\xi}_i^T\boldsymbol{c}})\right\} + \frac{\lambda_1}{2}\boldsymbol{c}^T Q\boldsymbol{c} + \lambda_2 \mathbf{1}^T\boldsymbol{\theta}, \tag{36}$$

subject to $\boldsymbol{\theta} \geq 0$, where $\boldsymbol{\xi}_i = \boldsymbol{\xi}(\boldsymbol{x}_i)$, $Q = \left\{\sum_{j\neq\alpha}\theta_j R_j(\tilde{x}_{u,j}, \tilde{x}_{v,j})\right\}_{u,v=1}^{q}$ are defined similarly as in Section 3, $\mathbf{1}$ is a $p-1$ vector with all 1's, and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_{\alpha-1}, \theta_{\alpha+1}, \cdots, \theta_p)^T$.

We need to compute $\boldsymbol{c}$, d, and $\boldsymbol{\theta}$ as minimizers of (36). Again, as the Algorithm in Section 3, we estimate $\boldsymbol{c}$, d, and $\boldsymbol{\theta}$ alternatively. With fixed $\boldsymbol{\theta}$, dropping the last term which is independent of $\boldsymbol{c}$ and d, (36) has the same form as (5.1) in Gu (2013). Therefore, we update $\boldsymbol{c}$ and d using the Newton-Raphson procedure with λ_1 selected by the generalized approximate cross-validation (GACV) method (Gu, 2013).

With fixed $\boldsymbol{c}$ and d, we rewrite $\hat{\nu}(\boldsymbol{x}) = d + \boldsymbol{\psi}^T(\boldsymbol{x})\boldsymbol{\theta}$, where $\boldsymbol{\psi}^T(\boldsymbol{x}) = (\psi_1^\dagger, \cdots, \psi_{\alpha-1}^\dagger, \psi_{\alpha+1}^\dagger, \cdots, \psi_p^\dagger)$ and $\psi_j^\dagger = \sum_{u=1}^{q} c_u \xi_{ju}(x_j)$. Plugging $\hat{\nu}(\boldsymbol{x}_i)$ into (35) and keeping terms involving $\boldsymbol{\theta}$ only, we have

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{x_{\alpha,i}\boldsymbol{\psi}_i^T\boldsymbol{\theta} - \log(1+e^{d+\boldsymbol{\psi}_i^T\boldsymbol{\theta}})\right\} + \frac{\lambda_1}{2}\boldsymbol{c}^T Q\boldsymbol{c} + \lambda_2 \mathbf{1}^T\boldsymbol{\theta}, \tag{37}$$

where $\boldsymbol{\psi}_i = \boldsymbol{\psi}(\boldsymbol{x}_i)$. Furthermore, minimizing (37) is equivalent to minimizing

$$A(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\left\{x_{\alpha,i}\boldsymbol{\psi}_i^T\boldsymbol{\theta} - \log(1+e^{d+\boldsymbol{\psi}_i^T\boldsymbol{\theta}})\right\} + \frac{\lambda_1}{2}\boldsymbol{c}^T Q\boldsymbol{c}, \tag{38}$$

subject to $\boldsymbol{\theta} \geq 0$ and $\mathbf{1}^T\boldsymbol{\theta} \leq M$ for some constant M. It is easy to see that the Hessian matrix of $A(\boldsymbol{\theta})$ is semi-definite, and consequently, $A(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$. We solve (38) iteratively using quadratic programming. Denote the current estimate of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}$, $\hat{\nu}(\boldsymbol{x}) = d + \boldsymbol{\psi}^T(\boldsymbol{x})\hat{\boldsymbol{\theta}}$, and $\hat{\nu}_i = \hat{\nu}(\boldsymbol{x}_i) = d + \boldsymbol{\psi}_i^T\hat{\boldsymbol{\theta}}$. We update $\boldsymbol{\theta}$ by minimizing the following second-order Taylor approximation of $A(\boldsymbol{\theta})$ (some constants independent of $\boldsymbol{\theta}$ have been removed):

$$\frac{1}{2}\boldsymbol{\theta}^T H_A(\tilde{\boldsymbol{\theta}})\boldsymbol{\theta} + \boldsymbol{\theta}^T\left\{G_A(\tilde{\boldsymbol{\theta}}) - H_A(\tilde{\boldsymbol{\theta}})\tilde{\boldsymbol{\theta}}\right\} \tag{39}$$

subject to $\boldsymbol{\theta} \geq 0$ and $\mathbf{1}^T\boldsymbol{\theta} \leq M$ for some constant M, where $G_A(\tilde{\boldsymbol{\theta}}) = -n^{-1}\sum_{i=1}^{n}\{x_{\alpha,i}\boldsymbol{\psi}_i - \boldsymbol{\psi}_i e^{\hat{\nu}_i}/(1+e^{\hat{\nu}_i})\} + \lambda_1\boldsymbol{q}/2$ is the gradient, $H_A(\tilde{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T e^{\hat{\nu}_i}/(1+e^{\hat{\nu}_i})^2$ is the Hessian, $\boldsymbol{q} = (\boldsymbol{c}^T Q_1\boldsymbol{c}, \cdots, \boldsymbol{c}^T Q_{\alpha-1}\boldsymbol{c}, \cdots, \boldsymbol{c}^T Q_{\alpha+1}\boldsymbol{c}, \cdots \boldsymbol{c}^T Q_p\boldsymbol{c})^T$, and $Q_j = \left\{R_j(\tilde{x}_{u,j}, \tilde{x}_{v,j})\right\}_{u,v=1}^{q}$ for $j = 1, \cdots, p$ and $j \neq \alpha$. We select the tuning parameter M by the K-fold cross-validation or the BIC method. We skip the implementation detail for the above algorithm since it is similar to that for the pseudo log-likelihood approach described in the next section.

Appendix B. Algorithm Implementation

In this section, we provide details about the implementation of the proposed algorithm using existing R packages. Specifically, we implement the Newton-Raphson procedure in the algorithm using a modification of the `sscden1` function in the `gss` package (Gu et al., 2014) and quadratic programming using the R function `solve.QP` in the `quadprog` package (Turlach and Weingessel, 2007).

B.1 Implementation of the Newton-Raphson Method

Given the current value of $\boldsymbol{\theta}_2$, we update $\boldsymbol{c}$ and $\boldsymbol{d}$ by minimizing (13) using the Newton-Raphson method. We implement by modifying the function `sscden1` in the `gss` package since (13) has the same form as (10.31) in Gu (2013) with different penalties and certain smoothing parameters being fixed. By definition, $\mathcal{H}_{(\alpha k)} = \mathcal{H}_{(\alpha)} \otimes \mathcal{H}_{(k)} = (\mathcal{H}^0_{(\alpha)} \oplus \mathcal{H}^1_{(\alpha)}) \otimes (\mathcal{H}^0_{(k)} \oplus \mathcal{H}^1_{(k)}) = (\mathcal{H}^0_{(\alpha)} \otimes \mathcal{H}^0_{(k)}) \oplus (\mathcal{H}^0_{(\alpha)} \otimes \mathcal{H}^1_{(k)}) \oplus (\mathcal{H}^1_{(\alpha)} \otimes \mathcal{H}^0_{(k)}) \oplus (\mathcal{H}^1_{(\alpha)} \otimes \mathcal{H}^1_{(k)}) = \mathcal{H}^{(0)}_{(\alpha k)} \oplus \mathcal{H}^{(1)}_{(\alpha k)} \oplus \mathcal{H}^{(2)}_{(\alpha k)} \oplus \mathcal{H}^{(3)}_{(\alpha k)}$ where $\mathcal{H}^{(0)}_{(\alpha k)} = \mathcal{H}^0_{(\alpha)} \otimes \mathcal{H}^0_{(k)}$, $\mathcal{H}^{(1)}_{(\alpha k)} = \mathcal{H}^0_{(\alpha)} \otimes \mathcal{H}^1_{(k)}$, $\mathcal{H}^{(2)}_{(\alpha k)} = \mathcal{H}^1_{(\alpha)} \otimes \mathcal{H}^0_{(k)}$, and $\mathcal{H}^{(3)}_{(\alpha k)} = \mathcal{H}^1_{(\alpha)} \otimes \mathcal{H}^1_{(k)}$. For density estimation, the penalized log-likelihood method in Gu (2013) does not penalize functions in the parametric component space $\mathcal{H}^0_{(\alpha k)}$ and has different smoothing parameters for components in the nonparametric component spaces $\mathcal{H}^{(1)}_{(\alpha k)}$, $\mathcal{H}^{(2)}_{(\alpha k)}$, and $\mathcal{H}^{(3)}_{(\alpha k)}$. Our goal is edge detection by detecting nonzero interactions. Therefore, we penalize the combined interaction $\eta_{\alpha k} \in \mathcal{H}_{(\alpha k)}$ as a whole with a smoothing parameter $\theta_{\alpha k}$ for $k = 1, \cdots, p$ and $k \neq \alpha$. The interaction $\eta_{\alpha k}$ collects parametric and nonparametric interaction components in $\mathcal{H}^{(0)}_{(\alpha k)}$, $\mathcal{H}^{(1)}_{(\alpha k)}$, $\mathcal{H}^{(2)}_{(\alpha k)}$, and $\mathcal{H}^{(3)}_{(\alpha k)}$. Note that $\boldsymbol{\theta}_2 = (\theta_{\alpha 1}, \cdots, \theta_{\alpha(\alpha-1)}, \theta_{\alpha(\alpha+1)}, \cdots, \theta_{\alpha p})^T$ is fixed at this step. We modified the function `sscden1` to solve (13) with smoothing parameters λ_1 and $\boldsymbol{\theta}_1$ estimated by the approximated cross-validation method.

B.2 Implementation of Quadratic Programming

Denote the current estimate of $\boldsymbol{\theta}_2$ as $\tilde{\boldsymbol{\theta}}_2$ and $\tilde{g}(\boldsymbol{x}) = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{d} + \boldsymbol{\psi}_1^T(\boldsymbol{x})\boldsymbol{\theta}_1 + \boldsymbol{\psi}_2^T(\boldsymbol{x})\tilde{\boldsymbol{\theta}}_2$. Define $\mu_{\tilde{g}}(h) = \sum_{i=1}^n e^{-\tilde{g}(\boldsymbol{x}_i)} h(\boldsymbol{x}_i) / \sum_{i=1}^n e^{-\tilde{g}(\boldsymbol{x}_i)}$, $V_{\tilde{g}}(h_1, h_2) = \mu_{\tilde{g}}(h_1 h_2) - \mu_{\tilde{g}}(h_1)\mu_{\tilde{g}}(h_2)$ for any functions h, h_1, and h_2. We update $\boldsymbol{\theta}_2$ by minimizing the following second-order Taylor approximation of $A_2(\boldsymbol{\theta}_2)$ with some constants independent of $\boldsymbol{\theta}_2$ have been removed:

$$\frac{1}{2}\boldsymbol{\theta}_2^T H_A(\tilde{\boldsymbol{\theta}}_2)\boldsymbol{\theta}_2 + \boldsymbol{\theta}_2^T \left\{ G_A(\tilde{\boldsymbol{\theta}}_2) - H_A(\tilde{\boldsymbol{\theta}}_2)\tilde{\boldsymbol{\theta}}_2 \right\} \tag{40}$$

subject to $\boldsymbol{\theta}_2 \geq 0$ and $\boldsymbol{w}^T\boldsymbol{\theta}_2 \leq M$ for some constant M, where $G_A(\tilde{\boldsymbol{\theta}}_2) = -\mu_{\tilde{g}}(\boldsymbol{\psi}_2) + \boldsymbol{b}_{\boldsymbol{\psi}_2} + \lambda_1 \boldsymbol{q}_2/2$ is the gradient, $H_A(\tilde{\boldsymbol{\theta}}_2) = V_{\tilde{g}}(\boldsymbol{\psi}_2, \boldsymbol{\psi}_2^T)$ is the Hessian, $\boldsymbol{q}_2 = (w_{\alpha 1}^{-1}\boldsymbol{c}^T Q_{\alpha 1}\boldsymbol{c}, \cdots, w_{\alpha(\alpha-1)}^{-1}\boldsymbol{c}^T Q_{\alpha(\alpha-1)}\boldsymbol{c}, w_{\alpha(\alpha+1)}^{-1}\boldsymbol{c}^T Q_{\alpha(\alpha+1)}\boldsymbol{c}, \cdots, w_{\alpha p}^{-1}\boldsymbol{c}^T Q_{\alpha p}\boldsymbol{c})^T$, and $Q_{\alpha k} = \left\{ R_{\alpha k}((\tilde{x}_{u,\alpha}, \tilde{x}_{u,k}), (\tilde{x}_{v,\alpha}, \tilde{x}_{v,k})) \right\}_{u,v=1}^q$ for $k = 1, \cdots, p$ and $k \neq \alpha$.

We use the R function `solve.QP` to solve (40). We estimate the tuning parameter M by minimizing a K-fold cross-validation or BIC score defined as follows. Let $I_1, \cdots, I_K$ be K randomly partitioned subsamples of the original data, $n_j = |I_j|$, and $n_{(-j)} = n - n_j$. Denote

$g_M^{(-j)}$ as the estimate without observations in the subset I_j which minimizes the following function with respect to $\boldsymbol{\theta}_2$:

$$\log\left\{\frac{1}{n_{(-j)}}\sum_{i\notin I_j} e^{-g(\boldsymbol{x}_i^\alpha)}\right\} + \frac{1}{n_{(-j)}}\sum_{i\notin I_j}\int_{\mathcal{X}_\alpha} g(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + \lambda_1\sum_{k\neq\alpha} w_{\alpha k}\theta_{\alpha k}^{-1}||\eta_{\alpha k}||^2 \qquad (41)$$

subject to $\theta_{\alpha k}\geq 0$ for $k\neq\alpha$ and $\boldsymbol{w}^T\boldsymbol{\theta}_2\leq M$. The K-fold cross-validation estimate of M is the minimizer of the following score:

$$\mathrm{CV}(M) = \log\left\{\frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j} e^{-g_M^{(-j)}(\boldsymbol{x}_i^\alpha)}\right\} + \frac{1}{n}\sum_{j=1}^{K}\sum_{i\in I_j}\int_{\mathcal{X}_\alpha} g_M^{(-j)}(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha. \qquad (42)$$

The BIC estimate of M is the minimizer of the following score:

$$\mathrm{BIC}(M) = \log\left\{\frac{1}{n}\sum_{i=1}^{n} e^{-g_M(\boldsymbol{x}_i^\alpha)}\right\} + \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha} g_M(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + \log(nk_n), \qquad (43)$$

where g_M expresses the dependence of the estimate on M explicitly, and k_n is the number of nonzero elements in the estimate of $\boldsymbol{\theta}_2$. We applied the K-fold cross-validation method in all simulations with $K=5$. We applied the BIC method in real data examples to get sparser graphs.

B.3 Initial Values and Convergence Criterion

To get a good initial value $\boldsymbol{\theta}_{2,0}$, we first estimate the conditional density $f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})$ with $\tau_1\sum_{k\neq\alpha} w_{\alpha k}||\eta_{\alpha k}||$ in (7) being replaced by $(\lambda_1/2)\sum_{k\neq\alpha}\theta_{\alpha k}^{-1}||\eta_{\alpha k}||^2$. We modified the `sscden` function in the `gss` package to estimate the conditional density and denote the estimate of $\eta_{\alpha k}$ as $\check{\eta}_{\alpha k}$. Since $\theta_{\alpha k}=0$ in $\boldsymbol{\theta}_2$ iff $\eta_{\alpha k}=0$, the magnitude of $\check{\eta}_{\alpha k}$ provides one way to initialize $\theta_{\alpha k}$. Specifically, we set $\theta_{\alpha k}^0 = \{\sum_{i=1}^n \check{\eta}_{\alpha k}^2(\boldsymbol{x}_i)\}^{1/2}$.

The convergence criterion in the algorithm is $||\boldsymbol{\theta}_2-\tilde{\boldsymbol{\theta}}_2||_2/(||\tilde{\boldsymbol{\theta}}_2||_2+10^{-6})\leq\varepsilon$ or the number of zeros in $\boldsymbol{\theta}_2$ stops increasing for fixed number of steps, where $\boldsymbol{\theta}_2$ and $\tilde{\boldsymbol{\theta}}_2$ are the updated and previous estimates, respectively, $||\cdot||_2$ is the Euclidean norm, and ε a threshold. We set $\varepsilon=0.001$ in simulation and real data examples.

Appendix C. Proofs

Proof of Proposition 1: We show the equivalence between minimization problems (7) and (9). First,

$$\begin{aligned}
&\min_{\eta\in\mathcal{M}_\alpha}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\eta(\boldsymbol{x}_i)}+\int_{\mathcal{X}_\alpha}\eta(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha\right\}+\frac{\lambda_1}{2}\sum_{j=1}^{p}\theta_j^{-1}||P_j\eta_j||^2+\tau_1\sum_{k\neq\alpha}w_{\alpha k}||\eta_{\alpha k}||\right\}\\
&=\min_{g\in\mathcal{G},\varsigma\in\mathbb{R}}\left\{\frac{1}{n}\sum_{i=1}^{n}\left\{e^{-g(\boldsymbol{x}_i)-\varsigma}+\int_{\mathcal{X}_\alpha}(g(\boldsymbol{x}_i^\alpha)+\varsigma)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha\right\}+\frac{\lambda_1}{2}\sum_{j=1}^{p}\theta_j^{-1}||P_j\eta_j||^2\right.\\
&\quad\left.+\tau_1\sum_{k\neq\alpha}w_{\alpha k}||\eta_{\alpha k}||\right\}. \qquad (44)
\end{aligned}$$

Setting the derivative of (44) with respect to ς to zero, we get $e^{\varsigma} = n^{-1}\sum_{i=1}^{n} e^{-g(\boldsymbol{x}_i)}$. Plugging it back to (44), we have the following profiled penalized pseudo log-likelihood

$$\min_{g\in\mathcal{G}}\Bigg\{1+\frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha} g(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha+\log\Big\{\frac{1}{n}\sum_{i=1}^{n}e^{-g(\boldsymbol{x}_i)}\Big\}+\frac{\lambda_1}{2}\sum_{j=1}^{p}\theta_j^{-1}||P_j\eta_j||^2$$
$$+\tau_1\sum_{k\neq\alpha}w_{\alpha k}||\eta_{\alpha k}||\Bigg\}.$$

□

Proof of Proposition 2: Set $\lambda_2=\tau_1^2/2\lambda_1$. Denote the functional in (9) as $B_1(g)$ and the functional in (10) as $B_2(\boldsymbol{\theta}_2, g)$. For any $\theta_{\alpha k}\geq 0$ and $g\in\mathcal{G}$, we have $\lambda_1\theta_{\alpha k}^{-1}||\eta_{\alpha k}||^2/2+\lambda_2\theta_{\alpha k}\geq \sqrt{2}\lambda_1^{1/2}\lambda_2^{1/2}||\eta_{\alpha k}||=\tau_1||\eta_{\alpha k}||$, and the equality holds if and only if $\theta_{\alpha k}=\lambda_1^{1/2}\lambda_2^{-1/2}||\eta_{\alpha k}||/\sqrt{2}$. Therefore, $B_2(\boldsymbol{\theta}_2,g)\geq B_1(g)$ for any $\theta_{\alpha k}\geq 0$ and $g\in\mathcal{G}$, and the equality holds if and only if $\theta_{\alpha k}=\lambda_1^{1/2}\lambda_2^{-1/2}||\eta_{\alpha k}||/\sqrt{2}$ for $\alpha\neq k$. The equivalence between (9) and (10) follows. □

Proof of Proposition 3: The population version of the pseudo log-likelihood $l_\alpha = n^{-1}\sum_{i=1}^{n}\Big\{e^{-\eta(\boldsymbol{x}_i)}+\int_{\mathcal{X}_\alpha}\eta(\boldsymbol{x}_i^\alpha)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha\Big\}$ in (7) is

$$\begin{aligned} l(\eta) &= E[e^{-\eta(\boldsymbol{x})}]+\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(x_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}\eta(\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x}\\ &= \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha} e^{-\eta(\boldsymbol{x})}f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})d\boldsymbol{x}+\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}\eta(\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x}, \end{aligned} \tag{45}$$

where $f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})$ is the density of $\boldsymbol{X}_{\setminus\{\alpha\}}$ on $\mathcal{X}_{\setminus\{\alpha\}}=\mathcal{X}_1\times\cdots\times\mathcal{X}_{\alpha-1}\times\mathcal{X}_{\alpha+1}\times\cdots\times\mathcal{X}_p$. The first and second-order Fréchet derivatives of $l(\eta)$ are

$$\begin{aligned} Dl(\eta)h_1 =& -\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha} e^{-\eta(\boldsymbol{x})}h_1(\boldsymbol{x})f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})d\boldsymbol{x}\\ &+\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha} h_1(\boldsymbol{x})\rho(\boldsymbol{x})d\boldsymbol{x}\\ D^2l(f)h_1h_2 =& \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha} e^{-\eta(\boldsymbol{x})}h_1(\boldsymbol{x})h_2(\boldsymbol{x})f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})d\boldsymbol{x}, \end{aligned}$$

where D denotes Fréchet derivative operator.

We set $Dl(\eta)h_1=\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(x_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}h_1(\boldsymbol{x})[\rho(\boldsymbol{x})-e^{-\eta(\boldsymbol{x})}f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})]d\boldsymbol{x}=0$ for all $h_1\in\mathcal{M}_\alpha$. Then, we have $\rho(\boldsymbol{x})-e^{-\eta(\boldsymbol{x})}f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})=0$, which implies that $f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})=e^{\eta(\boldsymbol{x})}\rho(\boldsymbol{x})$. In addition, $D^2l(\eta)hh=\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}e^{-\eta(\boldsymbol{x})}h^2(\boldsymbol{x})f(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})d\boldsymbol{x}>0$ for any nonzero $h\in\mathcal{M}_\alpha$, and consequently, l is strictly convex. Therefore, if $\hat{\eta}$ is the solution to (7), the estimate of the conditional density equals $e^{\hat{\eta}(\boldsymbol{x})}\rho(\boldsymbol{x})$.

We note that $\hat{\eta}=\hat{\varsigma}+\hat{g}$ where $\hat{\eta}$ is the solution to (7), $\hat{\varsigma}$ is given in Proposition 1, and $\hat{g}$ is the solution to (10). Since $\hat{\varsigma}$ is a constant independent of x_α, then the estimate of the conditional density $\hat{f}(x_\alpha|\boldsymbol{x}_{\setminus\{\alpha\}})$ is proportional to $e^{\hat{g}(\boldsymbol{x})}\rho(\boldsymbol{x})$. □

Proof of Convexity of $A_2(\boldsymbol{\theta}_2)$: We show that the Hessian matrix $H_A(\boldsymbol{\theta}_2)$ of $A_2(\boldsymbol{\theta}_2)$ is positive semi-definite. For any vector $\boldsymbol{\nu} \neq \mathbf{0}$, let $s_i = e^{-g(\boldsymbol{x}_i)}$ and $t_i = \boldsymbol{\nu}^T \boldsymbol{\psi}_2(\boldsymbol{x}_i)$, we have

$$\boldsymbol{\nu}^T H_A(\boldsymbol{\theta}_2)\boldsymbol{\nu} = \frac{\left(\sum_{i=1}^n s_i t_i^2\right)\left(\sum_{i=1}^n s_i\right) - \left(\sum_{i=1}^n t_i s_i\right)^2}{\left(\sum_{i=1}^n s_i\right)^2} \geq 0, \tag{46}$$

by the Cauchy-Schwartz inequality. □

In the remainder of the Appendix, we first introduce three lemmas and then provide proofs of Theorem 1, Theorem 2, and Corollary 1.

Lemma 1 *Assume* $J^*(g_0) < \infty$. *Under Conditions 1–3, as* $\lambda_1 \to 0$ *and* $n \to \infty$,

$$(V^* + \lambda_1 J^*)(\tilde{g} - g_0) = O_p(n^{-1}\lambda_1^{-1/r} + \lambda_1).$$

Proof: By the Fourier series expansions of $\tilde{g}$ and g_0, we have

$$V^*(\tilde{g} - g_0) = \sum_v (\tilde{a}_v - a_{v,0})^2 = \sum_v \frac{\kappa_v^2 - 2\kappa_v\lambda_1\gamma_v a_{v,0} + \lambda_1^2\gamma_v^2 a_{v,0}^2}{(1+\lambda_1\gamma_v)^2},$$

$$\lambda_1 J^*(\tilde{g} - g_0) = \sum_v \lambda_1\gamma_v(\tilde{a}_v - a_{v,0})^2 = \sum_v \lambda_1\gamma_v \frac{\kappa_v^2 - 2\kappa_v\lambda_1\gamma_v a_{v,0} + \lambda_1^2\gamma_v^2 a_{v,0}^2}{(1+\lambda_1\gamma_v)^2}.$$

Since $E(\kappa_v) = 0$ and $E(\kappa_v^2) \leq c_1/n$, we have

$$E[V^*(\tilde{g} - g_0)] \leq \frac{c_1}{n}\sum_v \frac{1}{(1+\lambda_1\gamma_v)^2} + \lambda_1 \sum_v \frac{\lambda_1\gamma_v}{(1+\lambda_1\gamma_v)^2}\gamma_v a_{v,0}^2,$$

$$E[\lambda_1 J^*(\tilde{g} - g_0)] \leq \frac{c_1}{n}\sum_v \frac{\lambda_1\gamma_v}{(1+\lambda_1\gamma_v)^2} + \lambda_1 \sum_v \frac{(\lambda_1\gamma_v)^2}{(1+\lambda_1\gamma_v)^2}\gamma_v a_{v,0}^2. \tag{47}$$

Following similar arguments in the proof of Lemma 9.1 in Gu (2013), we have

$$\sum_v \frac{\lambda_1\gamma_v}{(1+\lambda_1\gamma_v)^2} = O(\lambda_1^{-1/r}), \quad \sum_v \frac{1}{(1+\lambda_1\gamma_v)^2} = O(\lambda_1^{-1/r}), \quad \sum_v \frac{1}{1+\lambda_1\gamma_v} = O(\lambda_1^{-1/r}).$$

The lemma follows from (47) and the fact that $\sum_v \gamma_v a_{v,0}^2 = J^*(g_0) < \infty$. □

As in Gu (2013), when g_0 is "supersmooth" in the sense that $\sum_v \gamma_v^l a_{v,0}^2 < \infty$ for some $1 < l \leq 2$, which is assumed in Theorem 1, the rates can be improved to $O(n^{-1}\lambda_1^{-1/r} + \lambda_1^l)$.

Now we want to bound the approximation error $\hat{g} - \tilde{g}$. Define

$$\begin{aligned} A_{h_1,h_2}(\tau) =& \frac{1}{n}\sum_{i=1}^n e^{-(h_1+\tau h_2)(\boldsymbol{X}_i)} + \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{X}_\alpha} (h_1+\tau h_2)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + \frac{\lambda_1}{2}J^*(h_1+\tau h_2) \\ &+ \lambda_2 \sum_{k\neq\alpha}\theta_{\alpha k}, \\ B_{h_1,h_2}(\tau) =& \frac{1}{n}\sum_{i=1}^n -e^{-g_0(\boldsymbol{X}_i)}(h_1+\tau h_2)(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^n \int_{\mathcal{X}_\alpha} (h_1+\tau h_2)\rho(\boldsymbol{x}_i^\alpha)dx_\alpha \\ &+ \frac{1}{2}V^*(h_1+\tau h_2 - g_0) + \frac{\lambda_1}{2}J^*(h_1+\tau h_2). \end{aligned}$$

Taking derivatives of A_{h_1,h_2} and B_{h_1,h_2} with respect to τ and evaluating them at $\tau = 0$, we obtain

$$\dot{A}_{h_1,h_2}(0) = -\frac{1}{n}\sum_{i=1}^{n} e^{-h_1(\boldsymbol{X}_i)}h_2(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha} h_2\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + \lambda_1 J^*(h_1, h_2), \tag{48}$$

$$\dot{B}_{h_1,h_2}(0) = -\frac{1}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)}h_2(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha} h_2\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + V^*(h_1 - g_0, h_2) + \lambda_1 J^*(h_1, h_2). \tag{49}$$

Setting $h_1 = \hat{g}$ and $h_2 = \hat{g} - \tilde{g}$ in (48), we have

$$-\frac{1}{n}\sum_{i=1}^{n} e^{-\hat{g}(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha} (\hat{g} - \tilde{g})\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + \lambda_1 J^*(\hat{g}, \hat{g} - \tilde{g}) = 0. \tag{50}$$

Setting $h_1 = \tilde{g}$ and $h_2 = \hat{g} - \tilde{g}$ in (49), we have

$$\begin{aligned} &-\frac{1}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})(\boldsymbol{X}_i) + \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{X}_\alpha} (\hat{g} - \tilde{g})\rho(\boldsymbol{x}_i^\alpha)dx_\alpha + V^*(\tilde{g} - g_0, \hat{g} - \tilde{g}) \\ &+ \lambda_1 J^*(\tilde{g}, \hat{g} - \tilde{g}) = 0. \end{aligned} \tag{51}$$

Subtracting (51) from (50), we obtain

$$\begin{aligned} &\lambda_1 J^*(\hat{g} - \tilde{g}) - \frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\hat{g}(\boldsymbol{X}_i)} - e^{-\tilde{g}(\boldsymbol{X}_i)}\right\}(\hat{g} - \tilde{g})(\boldsymbol{X}_i) \\ =&\frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\tilde{g}(\boldsymbol{X}_i)} - e^{-g_0(\boldsymbol{X}_i)}\right\}(\hat{g} - \tilde{g})(\boldsymbol{X}_i) + V^*(\hat{g} - \tilde{g}, \tilde{g} - g_0). \end{aligned} \tag{52}$$

Applying the mean value theorem, we have $e^{-\hat{g}(\boldsymbol{X}_i)} - e^{-\tilde{g}(\boldsymbol{X}_i)} = -e^{-(\tilde{g}+\tau_i(\hat{g}-\tilde{g}))(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})(\boldsymbol{X}_i)$ where $\tau_i \in [0, 1]$. Since $\hat{g}$ and $\tilde{g}$ belong to B_0 which is a convex set around g_0, under Condition 4, there exists a $b_0^{(i)} \in (c_2, c_3)$ such that $-e^{-(\tilde{g}+\tau_i(\hat{g}-\tilde{g}))(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})(\boldsymbol{X}_i) = -b_0^{(i)}e^{-g_0(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})(\boldsymbol{X}_i)$. Then

$$\begin{aligned} -\frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\hat{g}(\boldsymbol{X}_i)} - e^{-\tilde{g}(\boldsymbol{X}_i)}\right\}(\hat{g} - \tilde{g})(\boldsymbol{X}_i) =&\frac{1}{n}\sum_{i=1}^{n} b_0^{(i)}e^{-g_0(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})^2(\boldsymbol{X}_i) \\ \geq&\frac{c_2}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})^2(\boldsymbol{X}_i). \end{aligned} \tag{53}$$

By the same argument, there exists a $c_0^{(i)} \in (c_2, c_3)$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\left\{e^{-\tilde{g}(\boldsymbol{X}_i)} - e^{-g_0(\boldsymbol{X}_i)}\right\}(\hat{g} - \tilde{g})(\boldsymbol{X}_i) = -\frac{1}{n}\sum_{i=1}^{n} c_0^{(i)}e^{-g_0(\boldsymbol{X}_i)}(\hat{g} - \tilde{g})(\boldsymbol{X}_i)(\tilde{g} - g_0)(\boldsymbol{X}_i). \tag{54}$$

Lemma 2 *Under Conditions 1, 2, and 5, suppose h_1 and h_2 are functions satisfying $J^*(h_1) < \infty$ and $J^*(h_2) < \infty$, as $\lambda_1 \to 0$ and $n\lambda_1^{2/r} \to \infty$, one has*

$$\Big|\frac{1}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)} h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i) - V^*(h_1,h_2)\Big| = o_p\Big(\{(V^*+\lambda_1 J^*)(h_1)(V^*+\lambda_1 J^*)(h_2)\}^{1/2}\Big). \tag{55}$$

Proof: Since $J^*(h_1) < \infty$ and $J^*(h_2) < \infty$, then h_1 and h_2 can be expressed as Fourier series $h_1 = \sum_v h_{1,v}\zeta_v$ and $h_2 = \sum_v h_{2,v}\zeta_v$. Let

$$U_i = \zeta_v(\boldsymbol{X}_i)\zeta_u(\boldsymbol{X}_i)e^{-g_0(\boldsymbol{X}_i)} - \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) \int_{\mathcal{X}_\alpha} \zeta_v(\boldsymbol{x})\zeta_u(\boldsymbol{x})\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}}.$$

Note that U_i are i.i.d. random variables with $\mathrm{E}(U_i) = 0$. Then under Condition 5, we have

$$E\left(\frac{1}{n}\sum_{i=1}^{n} U_i\right)^2 = \frac{1}{n}\mathrm{Var}\left(\zeta_v(\boldsymbol{X}_1)\zeta_u(\boldsymbol{X}_1)e^{-g_0(\boldsymbol{X}_1)}\right) < \frac{c_4}{n}.$$

Furthermore,

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)} h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i) - V^*(h_1,h_2)\Big| \\
=&\Big|\sum_v\sum_u h_{1,v}h_{2,u}\frac{1}{n}\sum_{i=1}^{n} U_i\Big| \\
\le&\left\{\sum_v\sum_u \frac{1}{1+\lambda_1\gamma_v}\frac{1}{1+\lambda_1\gamma_u}\left(\frac{1}{n}\sum_{i=1}^{n} U_i\right)^2\right\}^{1/2}\left\{\sum_v\sum_u (1+\lambda_1\gamma_v)(1+\lambda_1\gamma_u)h_{1,v}^2h_{2,u}^2\right\}^{1/2} \\
=&O_p(n^{-1/2}\lambda_1^{-1/r})\{(V^*+\lambda_1 J^*)(h_1)(V^*+\lambda_1 J^*)(h_2)\}^{1/2} \\
=&o_p\Big(\{(V^*+\lambda_1 J^*)(h_1)(V^*+\lambda_1 J^*)(h_2)\}^{1/2}\Big),
\end{aligned}$$

where the second equality holds because of the fact $\sum_v \frac{1}{1+\lambda_1\gamma_v} = O(\lambda_1^{-1/r})$ and the strong law of large numbers. □

Lemma 3 *Under Conditions 1, 2, and 5, as $\lambda_1 \to 0$ and $n\lambda_1^{2/r} \to \infty$, then*

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)} h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^{n} c_0^{(i)} e^{-g_0(\boldsymbol{X}_i)} h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)\Big| \\
&\qquad\qquad \le 2c_0\{(V^*+\lambda_1 J^*)(h_1)(V^*+\lambda_1 J^*)(h_2)\}^{1/2}
\end{aligned} \tag{56}$$

holds with probability 1, where $c_0 = max\{|c_2 - 1|, |c_3 - 1|\}$.

Proof: Note that

$$
\begin{aligned}
&\mathrm{E}|e^{-g_0(\boldsymbol{X}_i)}h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)| \\
=&\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) \int_{\mathcal{X}_\alpha} |h_1(\boldsymbol{x})h_2(\boldsymbol{x})|\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}} \\
\leq&\Big\{\big(\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) \int_{\mathcal{X}_\alpha} h_1^2(\boldsymbol{x})\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}}\big)\big(\int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) \int_{\mathcal{X}_\alpha} h_2^2(\boldsymbol{x})\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}}\big)\Big\}^{1/2} \\
=&\{V^*(h_1)V^*(h_2)\}^{1/2} \\
\leq&\{(V^*+\lambda_1 J^*)(h_1)(V^*+\lambda_1 J^*)(h_2)\}^{1/2},
\end{aligned}
$$

where the first inequality follows the Cauchy-Schwartz inequality. Then, we have

$$
\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^n e^{-g_0(\boldsymbol{X}_i)}h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^n c_0^{(i)} e^{-g_0(\boldsymbol{X}_i)}h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)\Big| \\
=&\Big|\frac{1}{n}\sum_{i=1}^n (1-c_0^{(i)}) e^{-g_0(\boldsymbol{X}_i)}h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)\Big| \\
\leq&\frac{1}{n}\sum_{i=1}^n |(1-c_0^{(i)})||e^{-g_0(\boldsymbol{X}_i)}h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)| \\
\leq&c_0\frac{1}{n}\sum_{i=1}^n |e^{-g_0(\boldsymbol{X}_i)}h_1(\boldsymbol{X}_i)h_2(\boldsymbol{X}_i)| \\
\leq&2c_0\{(V^*+\lambda_1 J^*)(h_1)(V^*+\lambda_1 J^*)(h_2)\}^{1/2},
\end{aligned}
$$

where the last inequality holds due to the strong law of large numbers. □

Proof of Theorem 1: Note that $\mathrm{E}\{e^{-g_0(\boldsymbol{X}_i)}(\hat{g}-\tilde{g})^2(\boldsymbol{X}_i)\} = \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) \int_{\mathcal{X}_\alpha} (\hat{g}-\tilde{g})^2(\boldsymbol{x})\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}} = V^*(\hat{g}-\tilde{g})$. Substituting (53) into the left-hand side of (52), we have

$$
\begin{aligned}
&\lambda_1 J^*(\hat{g}-\tilde{g}) - \frac{1}{n}\sum_{i=1}^n \left\{e^{-\hat{g}(\boldsymbol{X}_i)} - e^{-\tilde{g}(\boldsymbol{X}_i)}\right\}(\hat{g}-\tilde{g})(\boldsymbol{X}_i) \\
\geq&\frac{c_2}{n}\sum_{i=1}^n e^{-g_0(\boldsymbol{X}_i)}(\hat{g}-\tilde{g})^2(\boldsymbol{X}_i) + \lambda_1 J^*(\hat{g}-\tilde{g}) \\
\geq&\frac{c_2}{2}V^*(\hat{g}-\tilde{g}) + \lambda_1 J^*(\hat{g}-\tilde{g}), \qquad (57)
\end{aligned}
$$

where the last equality holds due to the strong law of large numbers. Substituting (55) and (56) into the right-hand side of (52) and letting $h_1 = \hat{g} - \tilde{g}$, $h_2 = \tilde{g} - g_0$, we have

$$\begin{aligned}
&\Big|\frac{1}{n}\sum_{i=1}^{n}\Big\{e^{-\tilde{g}(\boldsymbol{X}_i)} - e^{-g_0(\boldsymbol{X}_i)}\Big\}(\hat{g}-\tilde{g})(\boldsymbol{X}_i) + V^*(\hat{g}-\tilde{g}, \tilde{g}-g_0)\Big| \\
\leq&\Big|V^*(\hat{g}-\tilde{g}, \tilde{g}-g_0) - \frac{1}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)}(\hat{g}-\tilde{g})(\boldsymbol{X}_i)(\tilde{g}-g_0)(\boldsymbol{X}_i)\Big| \\
&+\Big|\frac{1}{n}\sum_{i=1}^{n} e^{-g_0(\boldsymbol{X}_i)}(\hat{g}-\tilde{g})(\boldsymbol{X}_i)(\tilde{g}-g_0)(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^{n} c_0^{(i)} e^{-g_0(\boldsymbol{X}_i)}(\hat{g}-\tilde{g})(\boldsymbol{X}_i)(\tilde{g}-g_0)(\boldsymbol{X}_i)\Big| \\
\leq&(o_p(1)+2c_0)\{(V^*+\lambda_1 J^*)(\hat{g}-\tilde{g})(V^*+\lambda_1 J^*)(\tilde{g}-g_0)\}^{1/2}, \qquad (58)
\end{aligned}$$

where the first inequality follows (54) and the second inequality follows Lemma 2 and 3. Combining (52), (57), and (58), we obtain

$$(\frac{c_2}{2}V^* + \lambda_1 J^*)(\hat{g}-\tilde{g}) \leq (o_p(1)+2c_0)\{(V^*+\lambda_1 J^*)(\hat{g}-\tilde{g})(V^*+\lambda_1 J^*)(\tilde{g}-g_0)\}^{1/2}. \quad (59)$$

Combining (59) with Lemma 1, as $\lambda_1 \to 0$ and $n\lambda_1^{2/r} \to \infty$, we have $(V^*+\lambda_1 J^*)(\hat{g}-\tilde{g}) = O_p(n^{-1}\lambda_1^{-1/r} + \lambda_1^l)$ and Theorem 1 holds. □

Proof of Theorem 2: We know

$$\sum_{k\neq\alpha}||\eta_{\alpha k}(x_\alpha, x_k)||^2 \leq \left\{\sum_{k\neq\alpha}||\eta_{\alpha k}(x_j, x_k)||\right\}^2 \leq (p-1)\sum_{k\neq\alpha}||\eta_{\alpha k}(x_\alpha, x_k)||^2. \qquad (60)$$

Therefore, there exists some constant $C \in [1, \sqrt{p-1}]$ such that $C\{\sum_{k\neq\alpha}||\eta_{\alpha k}(x_\alpha, x_k)||^2\}^{1/2} = \sum_{k\neq\alpha}||\eta_{\alpha k}(x_\alpha, x_k)||$. Since $\sum_{k\neq\alpha}\theta_{\alpha k}$ is bounded by M, we can scale λ_1 and λ_2 such that $\theta_{\alpha k} \leq 1$. Since $J_2^*(g) = \sum_{k\neq\alpha}\theta_{\alpha k}^{-1}||\eta_{\alpha k}(x_\alpha, x_k)||^2 = \boldsymbol{c}^T(\sum_{k\neq\alpha}\theta_{\alpha k}Q_{\alpha k})\boldsymbol{c}$, $\sum_{k\neq\alpha}||\eta_{\alpha k}(x_\alpha, x_k)||^2 = \boldsymbol{c}^T(\sum_{k\neq\alpha}\theta_{\alpha k}^2 Q_{\alpha k})\boldsymbol{c}$, we have $J_2^2(g) = C^2\sum_{k\neq\alpha}||\eta_{\alpha k}(x_\alpha, x_k)||^2 \leq C^2 J_2^*(g)$ and consequently $J_2 \leq C(J^*)^{1/2}$. Furthermore, since $V_2^2(g^{(2)}) = \int_{\mathcal{X}_{\setminus\{\alpha\}}} f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}})\int_{\mathcal{X}_\alpha}\left\{g^{(2)}(\boldsymbol{x})\right\}^2\rho(\boldsymbol{x})dx_\alpha d\boldsymbol{x}_{\setminus\{\alpha\}} = V^*(g^{(2)})$, we have $V_2(g^{(2)}) = [V^*(g^{(2)})]^{1/2}$. Therefore,

$$(V_2+\lambda_1 J_2)(g^{(2)}) = ((V^*)^{1/2} + C\sqrt{\lambda_1}(\lambda_1 J^*)^{1/2})(g^{(2)}) \leq (1+C^2\lambda_1)^{1/2}(V^*+\lambda_1 J^*)^{1/2}(g^{(2)})$$

by the Cauchy-Schwarz inequality. Finally,

$$\begin{aligned}
(V+\lambda_1 J)(\hat{g}-g^0) &= (V_1+\lambda_1 J_1)(\hat{g}^{(1)}-g_0^{(1)}) + (V_2+\lambda_1 J_2)(\hat{g}^{(2)}-g_0^{(2)}) \\
&\leq (V^*+\lambda_1 J^*)(\hat{g}^{(1)}-g_0^{(1)}) + (1+C^2\lambda_1)^{1/2}(V^*+\lambda_1 J^*)^{1/2}(\hat{g}^{(2)}-g_0^{(2)}) \\
&= O_p(n^{-1}\lambda_1^{-1/r} + \lambda_1^l) + O(n^{-1/2}\lambda_1^{-1/2r} + \lambda_1^{l/2}) \\
&= O_p(n^{-1/2}\lambda_1^{-1/2r} + \lambda_1^{l/2}). \qquad (61)
\end{aligned}$$

□

Proof of Corollary 1: By the definition of $V(\cdot)$, $V(\hat{g}-g^0) = V_1(\hat{g}^{(1)}-g_0^{(1)}) + V_2(\hat{g}^{(2)}-g_0^{(2)}) = V^*(\hat{g}^{(1)}-g_0^{(1)}) + [V^*(\hat{g}^{(2)}-g_0^{(2)})]^{1/2}$. Following (61),

$$[V^*(\hat{g}^{(2)}-g_0^{(2)})]^{1/2} = O_p(n^{-1/2}\lambda_1^{-1/2r} + \lambda_1^{l/2}).$$

Following Lin et al. (2000), under the condition $0 < c_5 < \rho(\boldsymbol{x}) < c_6$ and $0 < c_7 < f_{\setminus\{\alpha\}}(\boldsymbol{x}_{\setminus\{\alpha\}}) < c_8$ for some positive constants c_5, c_6, c_7, and c_8, $[V^*(g)]^{1/2}$ is equivalent to the L_2 norm. Specifically, $V^*(g) \sim ||g||_2^2 = \sum_{j=1}^{p} ||\eta_j||_2^2 + \sum_{k\neq\alpha} ||\eta_{\alpha k}(x_\alpha, x_k)||_2^2$, $V^*(g^{(1)}) \sim \sum_{j=1}^{p} ||\eta_j||_2^2$, and $V^*(g^{(2)}) \sim \sum_{k\neq\alpha} ||\eta_{\alpha k}(x_\alpha, x_k)||_2^2$, where $\sim$ means equivalence. By definition, $V(g^{(2)}) = [V^*(g^{(2)})]^{1/2} \sim (\sum_{k\neq\alpha} ||\eta_{\alpha k}(x_\alpha, x_k)||_2^2)^{1/2}$. Consequently, two-way interactions under L_2 norm have the same convergence rate as $[V^*(g^{(2)})]^{1/2}$,

$$||\hat{\eta}_{\alpha k} - \eta_{0\alpha k}||_2 = O_p(n^{-1/2}\lambda_1^{-1/2r} + \lambda_1^{l/2}), \quad k \neq \alpha,\ k = 1, \cdots, p.$$

□

References

Rajiv Agarwal, Joyce L. Davis, and Linda Smith. Serum albumin is strongly associated with erythropoietin sensitivity in hemodialysis patients. *Clin J Am Soc Nephrol.*, 3(1): 98–104, 2008.

Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. The multiple quantile graphical model. *arXiv preprint arXiv:1607.00515*, 2016.

Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Shizhe Chen, Daniela M Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2015.

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

Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. *arXiv preprint arXiv:1011.6640*, 2010.

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

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

Chong Gu and Ping Ma. Nonparametric regression with cross-classified responses. *Canadian Journal of Statistics*, 39(4):591–609, 2011.

Chong Gu, Yongho Jeon, and Yi Lin. Nonparametric density estimation in high-dimensions. *Statistica Sinica*, pages 1131–1153, 2013.

Chong Gu et al. Smoothing spline anova models: R package gss. *Journal of Statistical Software*, 58(5):1–25, 2014.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

Holger Höfling and Robert Tibshirani. Estimation of sparse binary pairwise markov networks using pseudo-likelihoods. *Journal of Machine Learning Research*, 10(4), 2009.

Cho-Jui Hsieh, Inderjit S Dhillon, Pradeep K Ravikumar, and Mátyás A Sustik. Sparse inverse covariance matrix estimation using quadratic approximation. In *Advances in neural information processing systems*, pages 2330–2338, 2011.

Cho-Jui Hsieh, Mátyás A Sustik, Inderjit S Dhillon, Pradeep Ravikumar, et al. Quic: quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.*, 15(1):2911–2947, 2014.

Yongho Jeon and Yi Lin. An effective method for high-dimensional log-density anova estimation, with application to nonparametric graphical model building. *Statistica Sinica*, pages 353–374, 2006.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Eduardo Lacson, John Rogus, Ming Teng, Michael Lazarus, and Raymond Hakim. The association of race with erythropoietin dose in patients on long-term hemodialysis. *American Journal of Kidney Diseases*, 52(6):1104–1114, 2008.

John Lafferty, Han Liu, Larry Wasserman, et al. Sparse nonparametric graphical models. *Statistical Science*, 27(4):519–537, 2012.

Ginette Lafit, Francis Tuerlinckx, Inez Myin-Germeys, and Eva Ceulemans. A partial correlation screening approach for controlling the false positive rate in sparse gaussian graphical models. *Scientific Reports*, 9(1):1–24, 2019.

Jason D Lee and Trevor J Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

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

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

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

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

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.

Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

Jian Shi, Anna Liu, and Yuedong Wang. Spline density estimation and inference with model-based penalties. *Journal of Nonparametric Statistics*, 31:596–611, 2019.

Arun Suggala, Mladen Kolar, and Pradeep K Ravikumar. The expxorcist: nonparametric graphical models via conditional exponential densities. In *Advances in neural information processing systems*, pages 4446–4456, 2017.

Berwin A Turlach and Andreas Weingessel. quadprog: Functions to solve quadratic programming problems. *CRAN-Package quadprog*, 2007.

Peter Kehinde Uduagbamen, John Omotola Ogunkoya, Igwebuike Chukwuyerem Nwogbe, Solomon Olubunmi Eigbe, and Oluwamayowa Ruth Timothy. Ultrafiltration volume: Surrogate marker of the extraction ratio, determinants, clinical correlates and relationship with the dialysis dose. *J Clin Nephrol Ren Care*, 7:068, 2021.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

Yuedong Wang. *Smoothing splines: methods and applications*. CRC Press, 2011.

Hans F Weinberger. *Variational methods for eigenvalue approximation*. SIAM, 1974.

Anja Wille, Philip Zimmermann, Eva Vranová, Andreas Fürholz, Oliver Laule, Stefan Bleuler, Lars Hennig, Amela Prelić, Peter von Rohr, Lothar Thiele, et al. Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome biology*, 5(11):1–13, 2004.

Jiahui Yu, Jian Shi, Anna Liu, and Yuedong Wang. Smoothing spline semiparametric density models. *Journal of the American Statistical Association*, 2020.

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

Hao Helen Zhang, Guang Cheng, and Yufeng Liu. Linear or nonlinear? automatic structure discovery for partially linear models. *Journal of the American Statistical Association*, 106 (495):1099–1112, 2011.