Statistica Sinica Preprint No: SS-2020-0463	
Title	Nonparametric Interaction Selection
Manuscript ID	SS-2020-0463
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0463
Complete List of Authors	Yushen Dong and Yichao Wu
Corresponding Author	Yichao Wu
E-mail	yichaowu@uic.edu
Notice: Accepted version subject to English editing.	

Nonparametric Interaction Selection

Yushen Dong and Yichao Wu

University of Illinois at Chicago

Abstract:

We consider the nonparametric two-way interaction model and propose a method to select important main effect and interaction effect terms simultaneously. Our method is based on backfitting local constant smoothing. Interaction selection is achieved by solving a constrained optimization problem to identify which main effect and interaction effect terms favor an infinity smoothing bandwidth. We establish selection consistency for the proposed method. Simulation examples and a real data example are used to illustrate its competitive finite-sample performance.

Key words and phrases: Additive model, backfitting, local constant smoothing, variable selection.

1. Introduction

The readily available high dimensional data due to technology advance has motivated the extremely active research area of variable selection. There have been a lot of methods proposed in the literature for variable selection.

In this paper, we will target at a special kind of variable selection, namely interaction selection. More explicitly, we study how predictor variables contribute to the response via pairwise interaction and how to select important pairwise interaction.

We consider the nonparametric regression of a univariate response Y on multivariate predictors $\mathbf{X} = (X_1, X_2, \ldots, X_d)^T$ with $X_j \in \Omega_j \subset R$, $j = 1, 2, \ldots, d$. The additive model $Y = \alpha + \sum_{j=1}^{d} m_j(X_j) + \epsilon$ is a simplification of the fully nonparametric regression model $Y = m(\mathbf{X}) + \epsilon$ by assuming predictors' effects to be additive. Yet this additivity assumption may not be reasonable in many real applications. Note that the fully nonparametric regression model can be decomposed as $Y = \alpha + \sum_{j=1}^{d} m_j(X_j) + \sum_{1\le j<k\le d} m_{jk}(X_j, X_k) + \sum_{1\le j<k<l\le d} m_{jkl}(X_j, X_k, X_l) + \ldots + m_{12\ldots d}(X_1, X_2, \ldots, X_d) + \epsilon$ by separating interaction effects at different orders. In this sense, the additive model is essentially an approximation of the fully nonparametric regression model by ignoring all interaction effects.

In this paper, we focus on the nonparametric two-way interaction model

$$Y = \alpha + \sum_{j=1}^{d} m_j(X_j) + \sum_{1\le j<k\le d} m_{jk}(X_j, X_k) + \epsilon \tag{1.1}$$

and propose a new method to select important main effect and interaction effect terms simultaneously. However the main idea can be easily extended to more general cases with higher-order interactions. Model (1.1)

is not identifiable itself. Additional identifiability conditions are required. There are different ways to formulate its identifiability conditions. To facilitate the implementation of our proposed nonparametric interaction selection method, we adopt the following fixed-point identifiability conditions (Gustafson 2000):

$$m_j(x_{j,0}) = 0, j = 1, \ldots, d; \tag{1.2}$$

$$m_{jk}(x_{j,0}, \cdot) = 0, m_{jk}(\cdot, x_{k,0}) = 0 \text{ and } m_{jk}(x_{j,0}, x_{k,0}) = 0, 1 \le j < k \le d, \tag{1.3}$$

where $x_{j,0}$ is any fixed point in the domain Ω_j of X_j, $j = 1, 2, \ldots, d$. Our goal is to estimate the sets of important main and interaction effects denoted by $\mathcal{M} = \{j : m_j(\cdot) \neq 0\}$ and $\mathcal{I} = \{(j,k) : m_{jk}(\cdot,\cdot) \neq 0\}$, respectively.

In the literature, there are many attempts to perform parametric interaction selection. Parametric two-way interaction model essentially assumes further $m_j(X_j) = \beta_j X_j$ and $m_{jk}(X_j, X_K) = \beta_{jk} X_j X_k$ in the above nonparametric two-way interaction model model (1.1). This parametric two-way interaction model is also called quadratic regression model. Zhao et al. (2009) proposed a composite absolute penalties family and demonsrated that their method can perform parametric interaction selection for the parametric two-way interaction model. Yuan et al. (2009) proposed a structured variable selection and estimation procedure for the parametric

two-way interaction model. Choi et al. (2010) propose a parametric interaction selection method under strong heredity assumption. Here strong heredity requires $j \in \mathcal{M}$ and $k \in \mathcal{M}$ as long as $(j,k) \in \mathcal{I}$. In comparison, weak heredity requires $j \in \mathcal{M}$ or $k \in \mathcal{M}$ or both if $(j,k) \in \mathcal{I}$. Bien et al. (2013) proposed a lasso for hierarchical interactions. Hao and Zhang (2014) and Niu et al. (2018) studied interaction screening. Hao et al. (2018) proposed a new regularization method, regularization algorithm under marginality principle (RAMP), to perform parametric interaction selection. Other related methods include Kong et al. (2017) and Wang et al. (2020) among many others.

Our focus is on nonparametric interaction selection. Lin and Zhang (2006) proposed a component selection and smoothing operator based on smoothing spline ANOVA and can be used to fit the above nonparametric two-way interaction model and perform interaction selection. Radchenko and James (2010) proposed a method, variable selection using adaptive non-linear interaction structures in high dimension (VANISH), for model (1.1). Radchenko and James (2010)'s method represents each main effect and interaction effect term using a preselected set of univariate and bivariate, respectively, orthonormal basis functions. In particular, the bivariate orthonormal basis functions is chosen to be the tensor products of

the univariate basis functions in their implementation. This leads to some challenges in approximating complex interaction effect component function.

In this paper we propose a new nonparametric interaction selection method in the framework of coupling backfitting with local constant smoothing. The essential idea is that if an infinity smoothing bandwidth is used in the local constant smoothing for each main effect or interaction effect component function, the corresponding component function estimate will be a constant function implying that it is unimportant for the prediction of the response variable. Since we are backfitting local constant smoothing, our method is much more flexible in fitting any complex interaction component function and can overcome the aforementioned limitation of using tensor products of univariate basis functions to approximate bivariate interaction component functions. In addition, our algorithm does not need strong or weak heredity assumption. Yet it is possible to incorporate strong or weak heredity if such an information is available as discussed as the end of the paper.

For nonparametric variable selection, Wu and Stefanski (2015) studied the additive model

$$Y = \alpha + \sum_{j=1}^{d} m_j(X_j) + \epsilon$$

without interaction and proposed a structure recovery scheme towards poly-

nomial modeling. It is capable of identifying unimportant predictors, linear predictors, quadratic predictors, etc. White et al. (2017) proposed a variable selection method for the fully nonparametric model

$$Y = m(X_1, X_2, \ldots, X_d) + \epsilon.$$

The nonparametric two-way interaction model (1.1), which sits between the additive model and the fully nonparametric model, is the focus of the current paper. The proposed method can estimate the sets of important main effects and two-way interaction effects. In addition, it can be readily extended to models with high-order interactions. With this new contribution, we now have a full spectrum of nonparametric variable selection methods.

The rest of the paper is organized as follows. Section 2 presents the basic backfitting local constant smoothing procedure for the two-way interaction model. Our new nonparametric interaction selection method is introduced in Section 3. Some implementation issues are discussed in Section 4 with a toy example to illustrate how it works. Selection consistency is established in Section 5. Simulation examples in Section 6 and a real data example in Section 7 are used to demonstrate of the proposed method's competitive finite-sample performance. Section 8 gives some discusion on how to incorporate strong or weak heredity information and possible future extensions.

2. Backfitting estimation of the two-way interaction model

Backfitting is a commonly-used technique for the estimation of the additive model (Hastie and Tibshirani 1990). It can also be used to fit the two-way interaction model (1.1) with both main and interaction effect terms. Backfitting algorithm is an iterative algorithm. In each iteration, it sequentially updates the estimate of one model component at a time. Each updating requires a univariate or bivariate smoothing, depending on whether we are updating a main effect or an interaction effect term. For the purpose of selecting important main and interaction effect terms, we will couple backfitting with local constant smoothing (Fan and Gijbels 1996).

2.1 Univariate local constant smoothing

Univariate local constant smoothing is used to update the estimate of the main effect terms. To estimate a univariate regression function $g(t) = E(Z|T = t)$ from a random sample $\{(T_i, Z_i) : i = 1, \ldots, n\}$, the univariate local constant smoothing approximates $g(t)$ by a constant a. A weighted least squares approach is used to estimate a with weights specified by a kernel function $K(\cdot)$ and a smoothing bandwidth $h > 0$. More specifically, the univariate local constant smoothing estimate $\hat{g}(t)$ of $g(t)$ at any t is given by $\hat{a}$, the optimizer of $\hat{a} = \arg\min_a \sum_{i=1}^{n} \{Z_i - a\}^2 K(\frac{T_i - t}{h})$. We denote such

a univariate local constant smoothing by $S_{K,h}$.

2.2 Bivariate local constant smoothing

Bivariate local constant smoothing is used to estimate the interaction effect terms. It is based on exactly the same idea as the univariate local constant smoothing but is used for the case with two predictors. Suppose we estimate a bivariate regression function $g(s,t) = E(Z|S = s, T = t)$ from a random sample $\{(S_i, T_i, Z_i) : i = 1, \ldots, n\}$. The bivariate local constant smoothing estimate $\hat{g}(s,t)$ of $g(s,t)$ at any s and t is given by $\hat{c}$, the optimizer of

$$\hat{c} = \arg\min_c \sum_{i=1}^{n} \{Z_i - c\}^2 K(\frac{S_i - s}{h}) K(\frac{T_i - t}{h}).$$

Note that potentially different smoothing bandwidths can be used for S and T. Yet for simplicity, we will use a same smoothing bandwidth. Denote this bivariate local constant smoothing by $S2_{K,h}$.

2.3 Backfitting algorithm

With the above univariate and bivariate local constant smoothings in place, we are ready to present the backfitting algorithm for the two-way interaction model (1.1). The backfitting algorithm is an iterative algorithm. The essential idea is to update the estimate of a single main or interaction effect term at every step while keeping estimates of all other terms

fixed. The detailed backfitting algorithm for the two-way interaction model (1.1) is given in Algorithm 1 with given smoothing bandwidths $h_j > 0$ and $\tilde{h}_{jk} > 0$ for main and interaction effect terms, respectively. Denote the estimates at the convergence by $\widehat{\alpha}^{BF}(\mathbf{h}, \tilde{\mathbf{h}})$, $\widehat{m}_j^{BF}(\cdot; \mathbf{h}, \tilde{\mathbf{h}})$, and $\widehat{m}_{jk}^{BF}(\cdot, \cdot; \mathbf{h}, \tilde{\mathbf{h}})$ with $\mathbf{h} = (h_1, h_2, \ldots, h_d)^T$ and $\tilde{\mathbf{h}} = (\tilde{h}_{1,2}, \tilde{h}_{1,3}, \ldots, \tilde{h}_{(d-1),d})^T$. Note here that we use two notations $\tilde{h}_{jk}$ and $\tilde{h}_{j,k}$ interchangeably to avoid potential confusion. Similarly $m_{jk}(\cdot, \cdot)$ (resp. $\tilde{\lambda}_{jk}$ and $\widehat{\tilde{\lambda}}_{jk}$ to be defined) is same as $m_{j,k}(\cdot, \cdot)$ (resp. $\tilde{\lambda}_{j,k}$ and $\widehat{\tilde{\lambda}}_{j,k}$).

In Algorithm 1, it is important to update interaction effect terms before updating main effect terms in each iteration for the following reason. After applying a bivariate local constant smoothing to update the estimate of an interaction effect term in Step 2(a), the updated estimate of the interaction effect term may not satisfy the identifiability condition (1.3). To ensure the identifiability condition, a follow-up updating Step 2(b) is necessary and this follow-up updating in the interaction effect term will change the estimates of the corresponding two main effect terms. This changing may lead to suboptimal estimates of the main effect terms. It can be automatically fixed by the upcoming updating of the mean effect terms in Step 3.

Algorithm 1: Backfitting algorithm for the two-way interaction model (1.1)

Step 1: Initialize by setting $\widehat{\alpha} = n^{-1}\sum_{i=1}^{n} Y_i$, $\widehat{m}_j(\cdot) \equiv 0$ for $j = 1, \ldots, d$ and $\widehat{m}_{jk}(\cdot) \equiv 0$ for $1 \leq j < k \leq d$.

Step 2: For $j = 1, \ldots, d-1$; $k = j+1, \ldots, d$:

(a) apply the bivariate local constant smoother $S2_{K,\tilde{h}_{jk}}$ to

$$\left[\left\{(X_{ij}, X_{ik}), Y_i - \widehat{\alpha} - \sum_{l=1}^{d} \widehat{m}_l(X_{il}) - \sum_{s<t\,:\,(s,t)\neq(j,k)} \widehat{m}_{st}(X_{is}, X_{it})\right\}; i = 1, \ldots, n\right]$$

and set the estimated function to be the updated estimate $\widehat{m}_{jk}(\cdot,\cdot)$ of $m_{jk}(\cdot,\cdot)$.

(b) update $\widehat{\alpha} \leftarrow \widehat{\alpha} + \widehat{m}_{jk}(x_{j,0}, x_{k,0})$,

$\widehat{m}_j(\cdot) \leftarrow \widehat{m}_j(\cdot) + \widehat{m}_{jk}(\cdot, x_{k,0}) - \widehat{m}_{jk}(x_{j,0}, x_{k,0})$,

$\widehat{m}_k(\cdot) \leftarrow \widehat{m}_k(\cdot) + \widehat{m}_{jk}(x_{j,0}, \cdot) - \widehat{m}_{jk}(x_{j,0}, x_{k,0})$ and

$\widehat{m}_{jk}(\cdot,\cdot) \leftarrow \widehat{m}_{jk}(\cdot,\cdot) - \widehat{m}_{jk}(x_{j,0}, \cdot) - \widehat{m}_{jk}(\cdot, x_{k,0}) + \widehat{m}_{jk}(x_{j,0}, x_{k,0})$

to implement the identifiability conditions (1.2) and (1.3).

Step 3: for $j = 1, \ldots, d$:

(a) apply the univariate local constant smoother S_{K,h_j} to

$$\left[\left\{X_{ij}, Y_i - \widehat{\alpha} - \sum_{l\neq j} \widehat{m}_l(X_{il}) - \sum_{1\leq s<t\leq d} \widehat{m}_{st}(X_{is}, X_{it})\right\}; i = 1, \ldots, n\right]$$

and set the estimated function to be the updated estimate $\widehat{m}_j(\cdot)$ of $m_j(\cdot)$.

(b) update $\widehat{\alpha} \leftarrow \widehat{\alpha} + \widehat{m}_j(x_{j,0})$, $\widehat{m}_j(\cdot) \leftarrow \widehat{m}_j(\cdot) - \widehat{m}_j(x_{j,0})$ to implement the identifiability conditions (1.3).

Step 4: Update $\widehat{\alpha} \leftarrow \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{d} \widehat{m}_j(X_{ij}) - \sum_{1\leq s<t\leq d} \widehat{m}_{st}(X_{is}, X_{it})\right)$.

3. Main and interaction effect selection

It was noted in Wu and Stefanski (2015) that when $h_j = \infty$, the univariate local constant smoothing in Step 3(a) approximates $m_j(\cdot)$ by a constant leading to a constant function estimate. The follow-up updating Step 3(b) will shift the constant function estimate to a zero function $\hat{m}_j(\cdot) = 0$ to satisfy the identifiability condition (1.2). As a result, an infinity smoothing bandwidth in the backfitting algorithm leads to the corresponding predictor's main effect being estimated to be unimportant. Based on this finding, Wu and Stefanski (2015) proposed a variable selection method for the additive model.

By the same token, if $\tilde{h}_{jk} = \infty$ in Algorithm 1, the bivariate local constant smoothing in Step 2(a) leads to a bivariate constant function estimate. The follow-up updating Step 2(b) shifts it to zero function estimate $\hat{m}_{jk}(\cdot, \cdot) = 0$ exactly in the same way. Corresponding interpretation is that the interaction effect between X_j and X_k is estimated to be unimportant.

According to these findings, the selection of important main effect and interaction effect terms for the two-way interaction model (1.1) boils down to the identification of which main effect and interaction effect terms favor an infinity smoothing bandwidth in Algorithm 1. Based on this, we now propose a new method to perform main effect and interaction effect selection

simultaneously for the two-way interaction model (1.1).

It is not easy to estimate an infinity. We convert the estimation of an infinity to the estimation of a zero by reparemetrizing $\lambda_j = 1/h_j$ and $\tilde{\lambda}_{jk} = 1/\tilde{h}_{jk}$ as in Wu and Stefanski (2015) and White et al. (2017). Denote $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_d)^T$ and $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_{12}, \tilde{\lambda}_{13}, \ldots, \tilde{\lambda}_{(d-1)d})^T$ to be the vectors of inverse smoothing bandwidths for main and interaction effect terms, respectively. For a vector $\boldsymbol{\lambda}$, we denote $\boldsymbol{\lambda}^{-1} = (1/\lambda_1, 1/\lambda_2, \ldots, 1/\lambda_d)^T$.

Following Wu and Stefanski (2015) and White et al. (2017), we propose to estimate the favored smoothing bandwidth for each main effect or interaction effect term by solving a constrained optimization problem

$$\begin{aligned} \min_{\boldsymbol{\lambda},\tilde{\boldsymbol{\lambda}}} \quad & \sum_{i=1}^{n} \Bigg\{ Y_i - \widehat{\alpha}^{BF}(\boldsymbol{\lambda}^{-1}, \tilde{\boldsymbol{\lambda}}^{-1}) - \sum_{j=1}^{d} \widehat{m}_j^{BF}(X_{ij}; \boldsymbol{\lambda}^{-1}, \tilde{\boldsymbol{\lambda}}^{-1}) \\ & \qquad - \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} \widehat{m}_{jk}^{BF}(X_{ij}, X_{ik}; \boldsymbol{\lambda}^{-1}, \tilde{\boldsymbol{\lambda}}^{-1}) \Bigg\}^2 \qquad (3.4) \\ \text{subject to} \quad & \lambda_j \geq 0, j = 1, \ldots, d; \\ & \tilde{\lambda}_{jk} \geq 0, 1 \leq j < k \leq d; \\ & \sum_{j=1}^{d} \lambda_j + \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} \tilde{\lambda}_{jk} = \tau, \end{aligned}$$

where $\tau \geq 0$ is a regularization parameter to be tuned. Denote the optimizer by $\widehat{\boldsymbol{\lambda}} \equiv \widehat{\boldsymbol{\lambda}}(\tau) = (\widehat{\lambda}_1(\tau), \widehat{\lambda}_2(\tau), \ldots, \widehat{\lambda}_d(\tau))^T$ and

$$\widehat{\tilde{\boldsymbol{\lambda}}} \equiv \widehat{\tilde{\boldsymbol{\lambda}}}(\tau) = (\widehat{\tilde{\lambda}}_{1,2}(\tau), \widehat{\tilde{\lambda}}_{1,3}(\tau), \ldots, \widehat{\tilde{\lambda}}_{(d-1)d}(\tau))^T.$$

For an appropriately tuned τ, some components of $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\widetilde{\boldsymbol{\lambda}}}$ will be exactly zero. Then the estimated set of important main and interaction effects are given by $\widehat{\mathcal{M}}(\tau) = \{j : \widehat{\lambda}_j(\tau) > 0\}$ and $\widehat{\mathcal{I}}(\tau) = \{(j,k) : \widehat{\widetilde{\lambda}}_{jk}(\tau) > 0\}$, respectively. To match our asymptotic consistency to be developed in Section 5, we can possibly use alternative definition $\widehat{\mathcal{M}}(\tau) = \{j : \widehat{\lambda}_j(\tau) > \varepsilon\}$ and $\widehat{\mathcal{I}}(\tau) = \{(j,k) : \widehat{\widetilde{\lambda}}_{jk}(\tau) > \varepsilon\}$, respectively, for some small $\varepsilon > 0$. For example, ε can be chosen to be twice the convergence tolerance adopted in the forthcoming modified coordinate descent algorithm. Yet based on our limited numerical experience, we have observed that these two definitions are always giving us the same selection result. This is due to the lasso-type constraint.

4. Implementation issues and a toy example

4.1 Modified coordinate descent algorithm

Convexity is a highly desired property in optimization. However due to the complicated backfitting algorithm coupled with univariate and bivariate local constant smoothing, the objective function of the optimization problem (3.4) is not convex. We borrow the modified coordinate descent algorithm (Wu and Stefanski 2015) to solve (3.4) for any given $\tau > 0$. We skip the details to save space.

4.2 Tuning

AIC, BIC, and cross validation can be used to tune the hyperparameter τ in the constrained optimization problem (3.4). For AIC and BIC, sum of squared errors and degrees of freedom are needed. The sum of squared errors can be simply calculated by

$$\sum_{i=1}^{n}\left\{Y_i-\widehat{\alpha}^{BF}(\widehat{\boldsymbol{\lambda}}^{-1},\widehat{\tilde{\boldsymbol{\lambda}}}^{-1})-\sum_{j=1}^{d}\widehat{m}_j^{BF}(X_{ij};\widehat{\boldsymbol{\lambda}}^{-1},\widehat{\tilde{\boldsymbol{\lambda}}}^{-1})-\sum_{j=1}^{d-1}\sum_{k=j+1}^{d}\widehat{m}_{jk}^{BF}(X_{ij},X_{ik};\widehat{\boldsymbol{\lambda}}^{-1},\widehat{\tilde{\boldsymbol{\lambda}}}^{-1})\right\}^2.$$

Note that the univariate and bivariate local constant smoothings are linear smoothers (Fan and Gijbels 1996). The trace of the corresponding smoothing matrix can be used to gauge the degrees of freedom for the backfitting estimate of each model component of the two-way interaction model (1.1).

In particular, the degrees of freedom for main effect estimate $\widehat{m}_j^{BF}(\cdot;\mathbf{h},\tilde{\mathbf{h}})$ is given by $\mathrm{tr}\left(\boldsymbol{S_j}-\mathbf{1}\left(\mathbf{s}_j(x_{j,0})\right)^T\right)$. Here $\mathbf{1}$ is a column vector of ones of an appropriate length and in the current context is of length n, and $\mathbf{S}_j=(\mathbf{s}_j(x_{1j}),\mathbf{s}_j(x_{2j}),\ldots,\mathbf{s}_j(x_{nj}))^T$ is the smoothing matrix of the local constant smoothing with

$$\mathbf{s}_j(x_j)=\left(K(\frac{X_{1j}-x_j}{h_j}),K(\frac{X_{2j}-x_j}{h_j}),\ldots,K(\frac{X_{nj}-x_j}{h_j})\right)^T/\sum_{i=1}^{n}K(\frac{X_{ij}-x_j}{h_j}).$$

Note that the first and second terms of $\mathrm{tr}(\boldsymbol{S_j}-\mathbf{1}\left(\mathbf{s}_j(x_{j,0})\right)^T)$ correspond to Step 3(a) and 3(b), respectively. Since $\mathrm{tr}(\mathbf{1}\left(\mathbf{s}_j(x_{j,0})\right)^T)=\mathrm{tr}(\left(\mathbf{s}_j(x_{j,0})\right)^T\mathbf{1})=1$, we have $\mathrm{tr}(\boldsymbol{S_j}-\mathbf{1}\left(\mathbf{s}_j(x_{j,0})\right)^T)=\mathrm{tr}(\boldsymbol{S_j})-1$ as in Wu and Stefanski (2015).

It becomes more involved for the interaction effect term estimate $\widehat{m}_{jk}(\cdot,\cdot;\mathbf{h},\tilde{\mathbf{h}})$. Here are the details. Denote

$$\tilde{\mathbf{s}}_{jk}(x_j,x_k)=\frac{1}{\sum_{i=1}^{n}K(\frac{X_{ij}-x_j}{\tilde{h}_{jk}})K(\frac{X_{ik}-x_k}{\tilde{h}_{jk}})}\begin{pmatrix}K(\frac{X_{1j}-x_j}{\tilde{h}_{jk}})K(\frac{X_{1k}-x_k}{\tilde{h}_{jk}})\\K(\frac{X_{2j}-x_j}{\tilde{h}_{jk}})K(\frac{X_{2k}-x_k}{\tilde{h}_{jk}})\\\vdots\\K(\frac{X_{nj}-x_j}{\tilde{h}_{jk}})K(\frac{X_{nk}-x_k}{\tilde{h}_{jk}})\end{pmatrix}.$$

Then $\tilde{\mathbf{S}}_{jk}=(\tilde{\mathbf{s}}_{jk}(x_{1j},x_{1k}),\tilde{\mathbf{s}}_{jk}(x_{2j},x_{2k}),\ldots,\tilde{\mathbf{s}}_{jk}(x_{nj},x_{nk}))^T$ is the smoothing matrix for the bivariate local constant smoothing in Step 2(a) of Algorithm (1). For the follow-up updating Step 2(b), we similarly denote $\tilde{\mathbf{S}}_{j0k}=(\tilde{\mathbf{s}}_{jk}(x_{j,0},x_{1k}),\tilde{\mathbf{s}}_{jk}(x_{j,0},x_{2k}),\ldots,\tilde{\mathbf{s}}_{jk}(x_{j,0},x_{nk}))^T$ and

$$\tilde{\mathbf{S}}_{jk0}=(\tilde{\mathbf{s}}_{jk}(x_{1j},x_{k,0}),\tilde{\mathbf{s}}_{jk}(x_{2j},x_{k,0}),\ldots,\tilde{\mathbf{s}}_{jk}(x_{nj},x_{k,0}))^T.$$

Then the degrees of freedom of the interaction effect estimate $\widehat{m}_{jk}(\cdot,\cdot;\mathbf{h},\tilde{\mathbf{h}})$ is given by

$$\mathrm{tr}\left\{\tilde{\mathbf{S}}_{jk}-\tilde{\mathbf{S}}_{j0k}-\tilde{\mathbf{S}}_{jk0}+\mathbf{1}\left(\tilde{\mathbf{s}}_{jk}(x_{j,0},x_{k,0})\right)^T\right\},$$

where the last three terms are due to the follow-up updating Step 2(b) to make the interaction effect estimate $\widehat{m}_{jk}(\cdot,\cdot;\mathbf{h},\tilde{\mathbf{h}})$ satisfy the identifiability condition (1.3).

Following Buja et al. (1989) and putting all these together, the total degrees of freedom for the backfitting estimate for the two-way interaction

model (1.1) is given by

$$1+\sum_{j=1}^{d}(\mathrm{tr}(\mathbf{S}_j)-1)+\sum_{1\le j<k\le d}\left[\mathrm{tr}\left\{\tilde{\mathbf{S}}_{jk}-\tilde{\mathbf{S}}_{j0k}-\tilde{\mathbf{S}}_{jk0}\right\}+1\right]$$

by noting similarly that $\mathrm{tr}(\mathbf{1}\,(\tilde{\mathbf{s}}_{jk}(x_{j,0},x_{k,0}))^T)=\mathrm{tr}((\tilde{\mathbf{s}}_{jk}(x_{j,0},x_{k,0}))^T\,\mathbf{1})=1$. Here the very first term 1 is the degrees of freedom to account for the intercept term estimated in Step 4 of Algorithm (1).

In our forthcoming numerical examples, we will use a BIC criterion to tune the regularization parameter τ.

4.3 Refitting

With the tuned optimal $\widehat{\tau}$, the final estimated set of main and interaction effects are given by $\widehat{\mathcal{M}}(\widehat{\tau})$ and $\widehat{\mathcal{I}}(\widehat{\tau})$, respectively. If we want to estimate the overall regression function $m(\mathbf{x})=\alpha+\sum_{j=1}^{d}m_j(x_j)+\sum_{1\le j<k\le d}m_{jk}(x_j,x_k)$ as well, a refitting step may be necessary to improve performance. Note that in the nonparametric main and interaction effect estimation method proposed above, we need to couple the backfitting algorithm with local constant smoothing to perform selection. But it is well known that the local constant smoothing is suboptimal if one cares about estimating the regression function (Fan and Gijbels 1996). In particular, Fan and Gijbels (1996) showed theoretically that the local linear smoothing can do much better than the local constant smoothing in terms of reducing smoothing

bias while estimating the regression function. Consequently a refitting step can be adopted to improve performance in terms of estimating the overall regression function $m(\mathbf{x})$.

For the selected final model

$$Y = \alpha + \sum_{j \in \widehat{\mathcal{M}}(\widehat{\tau})} m_j(X_j) + \sum_{(j,k) \in \widehat{\mathcal{I}}(\widehat{\tau})} m_{jk}(X_j, X_k) + \epsilon,$$

we couple the backfitting algorithm with univariate (resp. bivariate) local linear smoothing for updating the main (resp. interaction) effect terms to obtain a final estimate of the overall regression function. An optimization problem similar to (3.4) can be used to determine optimal smoothing bandwidths for each term in conjunction with AIC to tune the corresponding regularization parameter since local linear smoothing is also a linear smoother (Fan and Gijbels 1996).

4.4 A toy example

To get a better idea how our proposed selection method works, we now illustrate with a toy example. A random sample of size $n = 200$ is generated from the following model with five predictors in total, two important main effect terms and three important interaction effect terms

$$Y = m_1(X_1) + m_2(X_2) + m_{1,2}(X_1, X_2) + m_{1,3}(X_1, X_3) + m_{4,5}(X_4, X_5) + \varepsilon,$$

where $m_1(t) = m_2(t) = 2\sin(\pi t)$, $m_{1,2}(s,t) = m_{1,3}(s,t) = m_{4,5}(s,t) = 2\sin(\pi st)$, $X_1, \ldots, X_5 \overset{iid}{\sim} Unif(-1,1)$, and independent $\varepsilon \sim N(0,1)$. The identifiability conditions (1.2) and (1.3) are satisfied with $x_{j,0} = 0$, $j = 1, 2, \ldots, 5$. Note that there are 5 main effect terms and 10 interaction effect terms in total. We apply our proposed main and interaction effect selection algorithm. The solution path in Figure 1 plots $\widehat{\lambda}_j(\tau)$ and $\widehat{\widetilde{\lambda}}_{jk}(\tau)$ versus the tuning parameter τ for main and interaction effects. Note that we only plot up to $\tau = 30$ for best visual effect.

In the beginning with $\tau = 0$, all optimal inverse smoothing bandwidths are zero since τ is the summation of all inverse smoothing bandwidth. As τ gradually increases, $\widehat{\lambda}_1(\tau)$, $\widehat{\lambda}_2(\tau)$, $\widehat{\widetilde{\lambda}}_{1,3}(\tau)$, $\widehat{\widetilde{\lambda}}_{1,2}(\tau)$, and $\widehat{\widetilde{\lambda}}_{4,5}(\tau)$ corresponding to important main and interaction effect terms sequentially depart from zero before any unimportant terms component does. Note until $\tau = 25$, the optimal inverse smoothing bandwidth corresponding to one unimportant term becomes nonzero. Therefore, our proposed method can perform main and interaction effect selection perfectly as long as τ is tuned in a large interval $[9, 24]$. After rescaling appropriately, we overlay the BIC in Figure 1, denoted by the thin black dotted line. It shows that the BIC tuning leads to a perfect main and interaction effect selection.

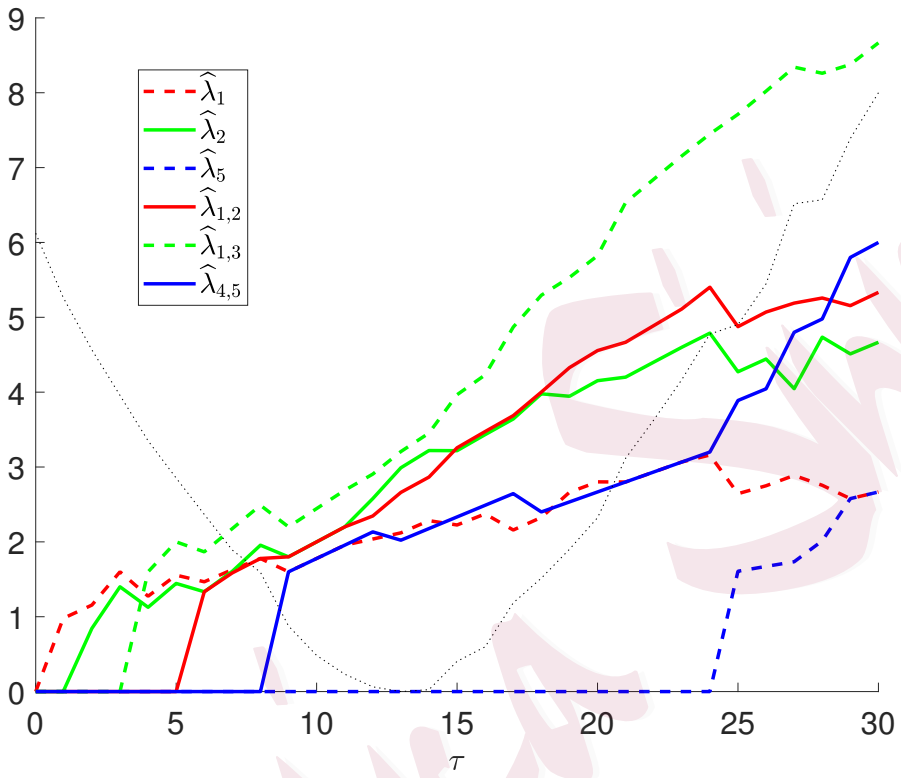

Figure 1: Solution path for a toy example.

5. Consistency

To establish selection consistency for the proposed nonparametric main and interaction effect selection method, we have proved the following asymptotic results for the optimizer of (3.4).

Theorem 1. Under Conditions 1-5 in Appendix, if $\tau \to \infty$ and $\tau^4/n \to 0$ as $n \to \infty$, the optimizer of (3.4) satisfies $\widehat{h}_j(\tau) \xrightarrow{p} \infty$ and $\widehat{h}_{j'}(\tau) \xrightarrow{p} 0$

for any $j \in \mathcal{M}$ and $j' \notin \mathcal{M}$, and $\widehat{\tilde{h}}_{jk}(\tau) \xrightarrow{p} \infty$ and $\widehat{\tilde{h}}_{j'k'}(\tau) \xrightarrow{p} 0$ for any $(j,k) \in \mathcal{I}$ and $(j',k') \notin \mathcal{I}$.

Theorem 1 implies selection consistency straightforwardly. Namely $P(\widehat{\mathcal{M}} = \mathcal{M}, \widehat{\mathcal{I}} = \mathcal{I}) \to 1$ as $n \to \infty$.

6. Simulation studies

Predictors in our simulation examples are generated in two steps. We first generate multivariate Gaussian $(Z_1, Z_2, \ldots, Z_d)^T$ with $E(Z_j) = 0$ and $\text{cov}(Z_j, Z_k) = \rho^{|j-k|}$ for $1 \le j, k \le d$. Here ρ controls the correlation among predictors and we will consider $\rho = 0.6$ in all of our simulation examples. Our predictors are generated by applying transformation $X_j = 2\Phi(U_j) - 1$ with $\Phi(\cdot)$ being the cumulative distribution function of standard normal distribution so that marginally $X_j \sim Unif(-1, 1)$ for $j = 1, 2, \ldots, d$. In our simulation studies, we fix $x_{j,0} = 0$ for $j = 1, 2, \ldots, d$ in the identifiability conditions (1.2) and (1.3). The dimension of predictors d is either 10 or 20 for all simulation examples.

We compare our proposed method with two existing methods: regularization algorithm under marginal principle (RAMP) method (Hao et al. 2018) and variable selection using adaptive non-linear interaction structures in high dimensions (VANISH) method (Radchenko and James 2010). We

evaluate performance of different methods in terms of two criteria: identification of important main and interaction effects, and integrated squared error (ISE) of each estimate of the overall regression function $m(\cdot)$, defined as $ISE(\widehat{m}) = E_{\mathbf{X}}(m(\mathbf{X}) - \widehat{m}(\mathbf{X}))^2$, where $\widehat{m}(\cdot)$ denotes an estimate of $m(\cdot)$. The expectation $E_{\mathbf{X}}$ is replaced by an empirical expectation based on a big independent test set.

Note that the VANISH method is designed for a nonlinear two-way interaction model with strong heredity and it requires an extra validation set to tune its regularization parameter. Here the strong heredity requires that if an interaction effect term is important, the corresponding two main effect terms must be important. The RAMP method is designed for quadratic regression, essentially an extended linear model with interaction effect terms added, and uses EBIC for tuning. In this sense, the RAMP is a linear method for main and interaction effect selection. It requires either strong or weak heredity. The weak heredity assumption requires that if an interaction effect term is important, at least one of corresponding two main effect terms is important. So to provide a fair comparison and a thorough investigation of our proposed method's finite-sample performance, we consider both linear and nonlinear two-way interaction model with and without strong heredity. In total we consider four simulation examples. For the models with strong

heredity, strong heredity version of RAMP is used while for the models without strong heredity, weak heredity version of RAMP is used. For all these four examples, VANISH uses Fourier basis.

6.1 Models with strong heredity

First we consider models with strong heredity.

Example 1. Linear two-way interaction model with strong heredity

Data are generated from model

$$Y = 2.1X_1 + 2.1X_2 + 2.1X_3 + 2.1X_4 + 3.7X_1X_2 + 3.7X_1X_3 + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$ is independent of predictors. In this model, there are four important main effect terms and two important interaction effect terms. Training sets of size 200 are used. An independent test set of size 1000 is generated to evaluate the ISE for each final estimate of the overall regression function. Strong heredity version of RAMP is used for Examples 1 and 2. Since the tuning of VANISH requires a separate tuning set, we generate an independent tuning set of size being the same as the training sets specifically for the tuning of VANISH for all simulation examples. In this sense, VANISH uses more data than the other two methods being compared.

Results over 100 repetitions are summarized in the first block of Table (1). For all three methods, M and NM columns are the average number of

selected true and false main effect terms; I and NI columns are the average number of selected true and false interaction effect terms; CM is the number of times recovering exactly the correct model (selecting all important terms and getting rid of all unimportant terms) among 100 repetitions; ISE is the integrated squared error defined above. For our new method, there are two extra columns OISE (Oracle ISE), which reports the ISE corresponding to the oracle model with only true important main and interaction effect terms, and PC (Path Consistency), which is the number of times the solution path contain at least one exactly correct model. The OISE is essentially obtained by applying the refitting step of Section 4.3 with true sets of important main and interaction effect terms. It serves as a benchmark how good our method can achieve. Numbers in parentheses are the corresponding standard errors.

With regards to main and interaction effect selection, our proposed method performs perfectly: selecting all important main and interaction effect terms perfectly and excluding all unimportant terms. RAMP also performs well as a linear method: it selects all important main and interaction effect terms but mistakenly include a few unimportant interactions. In comparison, VANISH has trouble to select all important terms, resulting in smaller number of correct models. RAMP has the smallest ISE since the true model is a linear two-way interaction model. ISE of our proposed

method is much smaller than that of VANISH.

Example 2. Nonlinear two-way interaction model with strong heredity

Data are generated from the model

$$Y = m_1(X_1)+m_2(X_2)+m_3(X_3)+m_4(X_4)+m_{(1,2)}(X_1, X_2)+m_{(1,3)}(X_1, X_3)+\varepsilon,$$

where $m_1(X_1) = 2.1\exp(X_1)$, $m_2(X_2) = 2.1\exp(X_2)$, $m_3(X_3) = 1.9\cos(X_3\pi)$, $m_4(X_4) = 1.9\cos(X_4\pi)$, $m_{(1,2)}(X_1, X_2) = 1.9\cos((X_1-X_2)\pi)$ and $m_{(1,3)}(X_1, X_3) = 6.8|X_1X_3| \cdot I_{\{X<0\}}(X_1X_3)$. The sample size of training data is 250 and all other settings are same as the linear case.

The second block of Table (1) summarizes the corresponding simulation results exactly in the same way. In terms of main and interaction effect selection, our proposed method still performs perfectly. RAMP misses some important terms, especially when the shape of nonlinear function is far away from linear, and adds some unimportant terms. VANISH also has trouble to select some important main effect and interaction effect terms. As a result, both RAMP and VANISH have low numbers of correct models. Our proposed method has a significantly smaller ISE compared to the other two methods. Overall, our proposed method outperforms RAMP and VANISH in this nonlinear case.

6.2 Models without strong heredity

Although strong heredity assumption is commonly used, weak heredity and no heredity constraints are possible in practice. Next we consider some more general models without strong heredity.

Example 3. Linear two-way interaction model without strong heredity

$$Y = 2.5X_1 + 2.5X_2 + 4X_1X_2 + 4X_1X_3 + 4X_4X_5 + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$. In this model, there are two important main terms and three important interaction terms. Three different cases of important interaction terms are considered to evaluate the performance of different methods: interaction term (X_1, X_2) with both corresponding main effects being important; interaction term (X_1, X_3) with one of the corresponding main effects being important; interaction term (X_4, X_5) with none of the corresponding main effects being important. Training data sets of size 150 and an independent test set of size 1000 are used. The third block of Table (1) shows the simulation results over 100 repetitions.

In terms of main and interaction effect selection, our proposed method still performs well: selecting important main and interaction effect terms perfectly (except missing one interaction term for one repetition) and unimportant terms at very low frequency. In comparison, both RAMP and

VANISH suffer a little. RAMP on average selects several unimportant interaction terms while VANISH fails to select some important main and interaction effects. Note that both RAMP and VANISH have either weak or strong heredity requirement. The interaction term (X_4, X_5) does not satisfy either weak or strong heredity and thus cannot be chosen. RAMP tend to add some unimportant terms into the model to make up for it, resulting in a smaller number of correct models. Our proposed method has the smallest ISE. Although here is a linear model, our proposed method has better selection performance and leads to a smaller ISE than the linear method RAMP does.

Example 4. Nonlinear two-way interaction model without strong heredity

$$Y = m_1(X_1)+m_2(X_2)+m_{(1,2)}(X_1, X_2)+m_{(1,3)}(X_1, X_3)+m_{(4,5)}(X_4, X_5)+\varepsilon, \tag{6.5}$$

where $m_i(X_i) = 1.9\cos(X_i\pi)$ and $m_{(j,k)}(X_j, X_k) = 2.1sin(X_jX_k\pi)$. The sample size of training data is 250 and all other settings are the same as in Example 3. The fourth block of Table (1) summarizes the corresponding simulation results over 100 repetitions.

The performance comparison is very similar to Example 3. In this case, our method achieves perfect main effect and interaction effect selection while RAMP and VANISH have some challenge.

It is observed that path consistency (PC) of our new method is always 100 out of 100 repetitions for all four simulation examples, even though the correct model (CM) is not equal to 100 for Example 3. This could be potentially improved by looking into an alternative tuning method.

During the review process, one reviewer inquired about the computational speed. On a MacBook equipped with Intel Core i5 @2.3GHz, on average it takes 1.37 and 23.81 minutes to solve the optimization problem (3.4) for $p = 10$ and $p = 20$, respectively, in Example 3; it takes 3.60 and 59.81 minutes to solve the optimization problem (3.4) for $p = 10$ and $p = 20$, respectively, in Example 4.

Table 1: Performance comparison for the simulation examples.

Example	d	New method								RAMP						VANISH					
		M	NM	I	NI	CM	PC	ISE	OISE	M	NM	I	NI	CM	ISE	M	NM	I	NI	CM	ISE
1	10	4.00	.00	2.00	.00	100	100	0.15(0.01)	0.15(0.01)	4.00	.00	2.00	0.04	96	0.04(0.01)	3.50	0.00	1.89	.04	44	4.96(0.14)
	20	4.00	.00	2.00	.00	100	100	0.15(0.01)	0.15(0.01)	4.00	.02	2.00	0.04	94	0.04(0.01)	3.64	0.00	1.93	.07	56	4.91(0.16)
2	10	4.00	.00	2.00	.00	100	100	0.70(0.01)	0.70(0.01)	2.62	.79	1.23	3.08	0	3.89(0.16)	3.89	0.00	1.38	.91	8	3.15(0.08)
	20	4.00	.00	2.00	.00	100	100	0.68(0.01)	0.68(0.01)	2.36	.56	0.94	1.99	0	4.78(0.17)	3.89	0.00	1.24	.80	2	3.02(0.08)
3	10	2.00	.05	2.99	.00	94	100	0.40(0.02)	0.38(0.02)	2.00	.50	2.49	2.74	17	1.12(0.11)	2.00	0.05	1.02	.02	0	8.79(0.16)
	20	2.00	.07	3.00	.00	93	100	0.42(0.02)	0.43(0.02)	2.00	.44	2.25	2.02	7	1.86(0.11)	1.98	0.02	0.96	.03	0	8.52(0.28)
4	10	2.00	.00	3.00	.00	100	100	0.98(0.03)	0.98(0.03)	0.34	.80	1.26	3.58	0	6.00(0.23)	2.00	1.20	2.04	.92	0	3.17(0.06)
	20	2.00	.00	3.00	.00	100	100	1.06(0.03)	1.06(0.03)	0.11	.39	0.53	1.41	0	6.95(0.14)	2.00	1.15	2.02	.65	0	2.89(0.05)

7. A real data example

We apply our proposed method to analyze a real data, the Real Estate Valuation data reported in Yeh and Hsu (2018). The data set includes 414 properties' information during the period of June 2012 to May 2013 from Xindian districts in Taipei City. The response is the residential housing price per unit area and there are six predictors: X_1=transaction date, X_2=house age, X_3=distance to the nearest MRT (Taipei Mass Rapid Transit) station, X_4=number of convenience stores, X_5=latitude, X_6=longitude. There is no missing value in this dataset.

We randomly split data into a training set of size $n = 210$ and a test set of size $\tilde{n} = 204$. We repeat with 30 random repetitions. We still compare our method with RAMP and VANISH in terms of the number of selected main and interaction term. Yet the ISE is replaced by the mean squared prediction error (MSPE) over the corresponding test set, namely $MSPE = \frac{1}{\tilde{n}} \sum_{i=1}^{\tilde{n}} (\tilde{Y}_i - \hat{\tilde{Y}}_i)^2$, where $\tilde{Y}_i$ and $\hat{\tilde{Y}}_i$ are the observed response and predicted response for the ith observation in the test set for each repetition. RAMP uses its weak heredity version and EBIC for tuning; VANISH uses a 10-fold cross-validation for tuning; and our proposed method uses BIC for tuning.

Table 2: Performance comparison for the real data example.

	New method	RAMP	VANISH
Main term size	3.1(0.2)	2.9(0.2)	1.0(0.0)
Interaction term size	0.3(0.1)	5.1(0.6)	0.0(0.0)
MSPE	76.03(2.58)	75.73(2.04)	170.85(3.71)

Table (2) summarizes the result for three methods over 30 repetitions. VANISH only selects the third predictor in all repetitions and the MSPE is larger than the other two methods. Our method is comparable with RAMP in terms of MSPE. But at the same time, the model selected by our proposed method is more parsimonious and easier to interpret since our method in general selects a model with much fewer number of terms, especially for the interaction. Our method has a good balance between the model complexity and prediction performance.

For a random repetition, our proposed model selects the main effect of X_3 and the interaction effect of X_2 and X_3. The estimated main effect component function $\widehat{m}_3(X_3)$ and interaction effect component function $\widehat{m}_{2,3}(X_2, X_3)$ are plotted plotted in Figures 2 and 3, respectively. It obviously shows that X_2 (house age) and X_3 (distance to the nearest MRT station) does show interaction effect that won't be able to be explained by

the additive model without interaction.

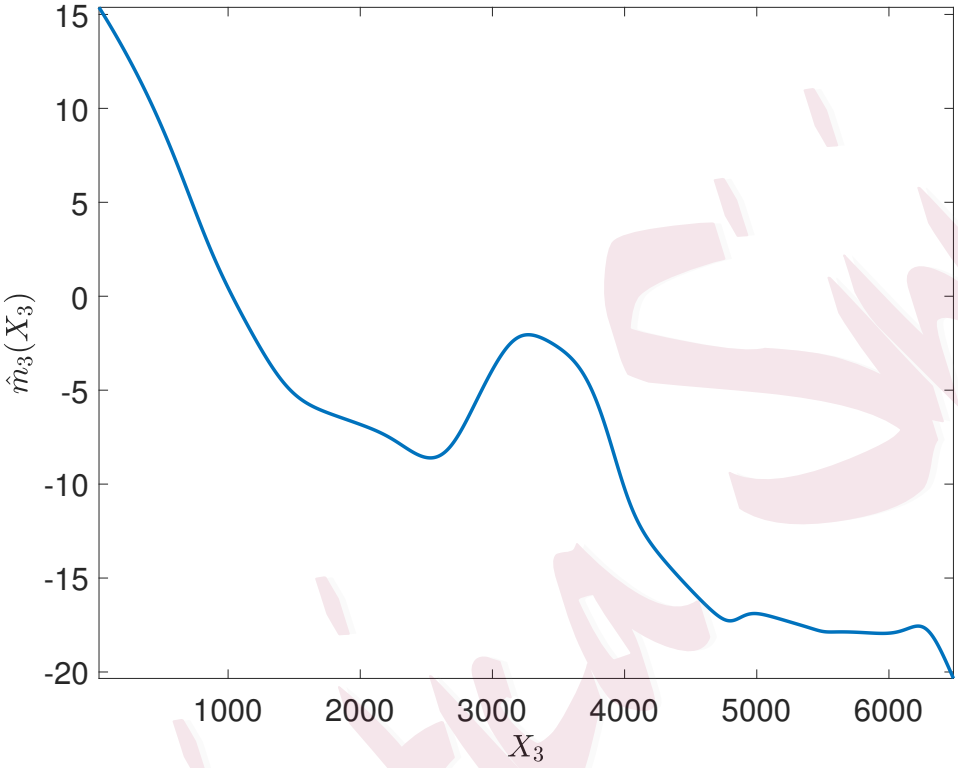

Figure 2: Plot of fitted main effect component function of X_3.

8. Discussion

During the review process, one referee pointed out that it will be desirable to provide a version to achieve strong or weak heredity. In fact, this is possible. To achieve weak heredity, we can minimize (3.4) subject to constraints

$$\lambda_j \geq 0, \quad j = 1, 2, \ldots, d$$

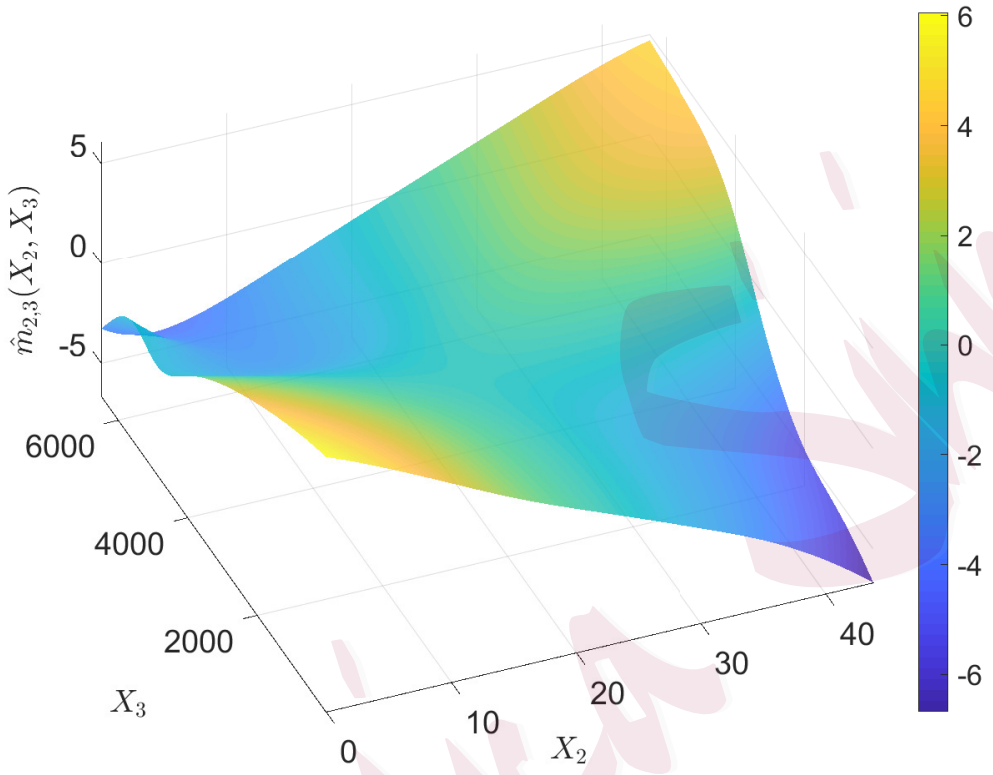

Figure 3: Plot of fitted interaction effect component function of X_2 and X_3.

$$\sum_{j=1}^{d} \lambda_j = \tau$$
$$\tilde{\lambda}_{jk} = \lambda_j + \lambda_k, \quad 1 \le j < k \le d.$$

For strong heredity, we can minimize (3.4) subject to constraints

$$\delta_j \ge 0, \quad j = 1, 2, \ldots, d,$$

$$\lambda_j = \sum_{k \neq j} \tilde{\lambda}_{jk} + \delta_j, \quad j = 1, 2, \ldots, d,$$
$$\tilde{\lambda}_{jk} \geq 0, \quad 1 \leq j < k \leq d,$$
$$\sum_{j=1}^{d} \delta_j + \sum_{j=1}^{d-1} \sum_{k=j+1}^{d} \tilde{\lambda}_{jk} = \tau.$$

We admit that our algorithm is not a fast one by nature. Our algorithm involves two layers of iterations: (modified) coordinate descent and backfitting alorithms, in addition to the local constant smoothing. Yet it is still manageable for a moderate dimensionality. For high dimensional case, we are working on an interaction screening procedure by extending the sure independence screening for nonparametric regression (Feng et al. 2018). The selection consistency in Section 5 was established for the case with a fixed dimensionality. It will be of great interest to extend it to the case with a diverging dimensionality.

Supplementary Materials

Contain the brief description of the online supplementary materials.

Acknowledgements

The authors would like to thank two anonymous referees, an Associate Editor, and co-Editor Prof. Yazhen Wang for their constructive comments, which has led us to a substantially improved version of the paper. The research is partially supported by NSF grants DMS-1821171 and CCF-

1934915.

References

Bien, J., J. Taylor, and R. Tibshirani (2013). A lasso for hierarchical interactions. *The Annals of Statistics 41*(3), 1111–1141.

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models. *Ann. Statist. 17*(2), 453–510.

Choi, N. H., W. Li, and J. Zhu (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association 105*(489), 354–364.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.

Feng, Y., Y. Wu, and L. A. Stefanski (2018). Nonparametric independence screening via favored smoothing bandwidth. *Journal of Statistical Planning and Inference 197*, 1–14.

Gustafson, P. (2000). Bayesian regression modeling with interactions and smooth effects. *Journal of the American Statistical Association 95*(451), 795–806.

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

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

REFERENCES

Hastie, T. and R. Tibshirani (1990). *Generalized additive models*. Wiley Online Library.

Kong, Y., D. Li, Y. Fan, and J. Lv (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics 45*(2), 897–922.

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

Niu, Y. S., N. Hao, and H. H. Zhang (2018). Interaction screening by partial correlation. *Statistics and Its Interface 11*, 317–325.

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

Wang, C., B. Jiang, and L. Zhu (2020). Penalized interaction estimation for ultrahigh dimensional quadratic regression. *Statistica Sinica*. in press.

White, K. R., L. A. Stefanski, and Y. Wu (2017). Variable selection in kernel regression using measurement error selection likelihoods. *Journal of the American Statistical Association 112*, 1587–1597.

Wu, Y. and L. A. Stefanski (2015). Automatic structure recovery for additive models. *Biometrika 102*(2), 381–395.

Yeh, I.-C. and T.-K. Hsu (2018). Building real estate valuation models with comparative approach through case-based reasoning. *Appl. Soft Comput. 65*, 260–271.

REFERENCES

Yuan, M., V. R. Joseph, and H. Zou (2009). Structured variable selection and estimation. *The Annals of Applied Statistics 3*(4), 1738–1757.

Zhao, P., G. Rocha, and B. Yu (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics 37*, 3468–3497.

University of Illinois at Chicago

E-mail: ydong37@uic.edu

University of Illinois at Chicago

E-mail: yichaowu@uic.edu