Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Robust Variable and Interaction Selection for Logistic Regression and General Index Models

Yang Li & Jun S. Liu

To cite this article: Yang Li & Jun S. Liu (2017): Robust Variable and Interaction Selection for Logistic Regression and General Index Models, Journal of the American Statistical Association, DOI: 10.1080/01621459.2017.1401541

To link to this article: http://dx.doi.org/10.1080/01621459.2017.1401541

Robust Variable and Interaction Selection for Logistic Regression and General Index Models

Yang Li, Jun S. Liu *

October 8, 2017

Abstract

Under the logistic regression framework, we propose a forward-backward method, SODA, for variable selection with both main and quadratic interaction terms. In the forward stage, SODA adds in predictors that have significant overall effects, whereas in the backward stage SODA removes unimportant terms to optimize the extended Bayesian Information Criterion (EBIC). Compared with existing methods for variable selection in quadratic discriminant analysis, SODA can deal with high-dimensional data in which the number of predictors is much larger than the sample size and does not require the joint normality assumption on predictors, leading to much enhanced robustness. We further extend SODA to conduct variable selection and model fitting for general index models. Compared with existing variable selection methods based on the Sliced Inverse Regression (SIR) (Li, 1991), SODA requires neither linearity nor constant variance condition and is thus more robust. Our theoretical analysis establishes the variable-selection consistency of SODA under high-dimensional settings, and our simulation studies as well as real-data applications demonstrate superior performances of SODA in dealing with non-Gaussian design matrices in both logistic and general index models.

Keywords: Classification; Forward screening; High-dimensional; Quadriatic discriminant analysis; Semi-parametric; Stepwise selection

*Yang Li is Sr. Market Scientist, Vatic Labs LLC, New York, NY 10036 (Email: yli@vaticlabs.com). Jun S Liu is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138; and is also co-Director for the Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China (E-mail: jliu@stat.harvard.edu). The research was supported in part by the NSF grants DMS-1208771 and DMS-1613035, and the NIH grant R01 GM113242-01.

1 Introduction

Classification, also known as "supervised learning," is a fundamental building block of statistical machine learning. Applications of classification methods include, for example, cancer diagnosis (Tibshirani et al., 2002), text categorization (Joachims, 1998), computer vision (Phillips, 1998), protein interaction predictions (Chowdhary et al., 2009), etc. Well-known classification methods include logistic regression, naive Bayes classifier, K-nearest-neighbors, support vector machines (Boser et al., 1992), random forests (Breiman, 2001), neural networks, etc. As important early players in statistical classification, linear and quadratic discriminant analysis (LDA and QDA) (Anderson, 1958) are widely used. Compared with LDA, QDA is able to exploit interaction effects of predictors.

With rapid technological advances in data collection, it has become common in many statistical learning problems that the number of predictors is much larger than the number of observations, which is also known as the "large p small n" problem. For example, in gene expression microarray analysis, usually n is in hundreds of samples, whereas p is in thousands of genes (Efron, 2010). In a typical genome-wide association study, n is in thousands and p (the number of SNP markers) is from thousands to millions (Waldmann et al., 2013). Vanilla LDA or QDA are infeasible when $p > n$ since the sample covariance matrices are singular and non-invertible. Even in low-dimensional scenarios, including many irrelevant predictors can significantly impair the classification accuracy.

A number of variable selection methods have been developed for high-dimensional classification problems, of which many focused on imposing regularization on the LDA model for both sparsity of model and stability of parameter estimation. For example, Witten and Tibshirani (2011) proposed to use fused Lasso to penalize discriminant vectors in Fisher's discriminant problem. Cai and Liu (2011) proposed to directly estimate the discriminant

function, i.e., the product of the precision matrix and the difference between two mean vectors, through a constrained L_1 minimization. Han et al. (2013) relaxed the normal assumption of LDA to entertain Gaussian Copula models. For more examples of recent development of high-dimensional LDA, see Guo et al. (2007); Fan and Fan (2008); Clemmensen et al. (2011); Shao et al. (2011); Mai et al. (2012); Fan et al. (2013).

Aforementioned methods work for LDA models, which have only linear main effects. In many applications, interaction effects may be significant and scientifically interesting. However, in moderate to high dimensional situations, including too many noise variables in QDA models can lead to an over-fitting problem more severe than that of LDA models, resulting in a much impaired prediction accuracy. In recent years, there has been a significant surge of interest in methods for detecting interaction effects in regression or classification problems (Simon and Tibshirani, 2012; Bien et al., 2013; Jiang and Liu, 2014; Fan et al., 2015). We use the term "interaction" to refer to all second-order effects, including both two-way interactions X_iX_j with $i \neq j$ and quadratic terms X_i^2.

To motivate later developments, we consider a two-class Gaussian classification problem with both linear and interaction effects involving three active predictors. The oracle Bayes rule is to classify an observation to class 1 if $Q(\mathbf{X}) > 0$, and to class 0 otherwise, where

$$Q(\mathbf{X}) = 1.627 + X_1 - 0.6X_1^2 - 0.6X_3^2 - 0.7X_1X_2 - 0.7X_2X_3. \quad (1)$$

We simulated 100 independent datasets, each having 100 observations in every class. Figure 1 shows the scatterplot of (X_1, X_2) for one simulated dataset. For each simulated dataset, we applied LDA, logistic regression, and QDA to train classifiers, and the classification accuracy was estimated by using 1000 additional testing samples generated from the Oracle model. As shown in Table 1, both LDA and logistic regression with only linear terms had poor prediction powers, whereas QDA improved the classification accuracy dramatically.

We further tested the classification accuracy of QDA when k additional noise predictors were included ($k = 1, \ldots, 50$), each being drawn independently from $\mathcal{N}(0, 1)$. Figure 1 shows that the classification error rate of QDA increased dramatically as the number of noise predictors increased, demonstrating the necessity of developing methods capable of selecting both main and interaction terms efficiently.

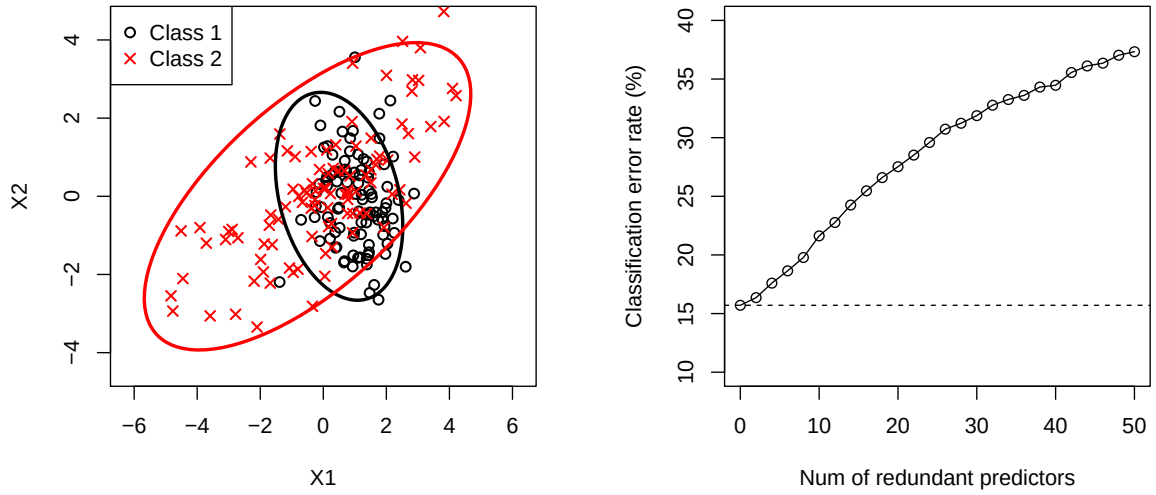

Figure 1: A two-class classification problem, where samples for the two classes were drawn from $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Omega}_1^{-1})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Omega}_2^{-1})$, respectively. We set $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2 = (0.5, 0, 0)$, $\boldsymbol{\Omega}_1 = \mathbf{I}_3 - \boldsymbol{\Omega}$, and $\boldsymbol{\Omega}_2 = \mathbf{I}_3 + \boldsymbol{\Omega}$, where $\boldsymbol{\Omega}$ has entries $\omega_{22} = 1$, $\omega_{11} = \omega_{33} = -0.60$, $\omega_{12} = \omega_{23} = -0.35$, and $\omega_{13} = 0$. **Left**: Scatterplot of (X_1, X_2) overlaid with corresponding theoretical contours; **Right**: QDA classification error rate versus the number of noise predictors included.

A direct application of Lasso-logistic regression with all second-order terms is prohibitive for moderately large p (e.g., $p \geq 1000$). To cope with this difficulty, Fan et al. (2015) proposed innovated interaction screening (IIS), which, after transforming each original p-

Method	LDA	Logistic regression	QDA	QDA+50
Test error %	34.81 (1.47)	34.88 (1.38)	15.65 (0.84)	37.33 (1.78)

Table 1: Means (standard deviations) of testing error rates for different classification methods over 100 replications. "QDA+50" means the QDA model with 50 additional noise variables.

dimensional predictor vector by multiplying it with the estimated precision matrix for its class, first reduces the number of predictors to a smaller order of p and then identifies both important main effects and interactions using the elastic net penalty (Zou and Hastie, 2005). The performance of the resulting method, IIS-SQDA, relies heavily on the estimation of the $p \times p$ precision matrix, which is usually a hard problem under high-dimensional settings. Murphy et al. (2010), Zhang and Wang (2011), and Maugis et al. (2011) proposed stepwise procedures for QDA variable selection. These methods were shown to be consistent under the multivariate Gaussian assumption on the design matrix. In practice, however, performances of these methods can be much compromised when the normality assumption is violated, especially when predictors follow heavier-tailed distributions or when they are correlated in non-linear manners (see Section 4).

In order to gain robustness and computational efficiency, we propose the method Stepwise cOnditional likelihood variable selection for Discriminant Analysis (SODA) under the logistic regression framework. It starts with forward steps to select predictors with significant overall effects, and finishes with backward elimination steps for further pinning down the exact model. The criterion used for both forward and backward steps is the extended Bayesian information criterion (EBIC) (Chen and Chen, 2008). Although stepwise variable selection methods have been widely known and used for regression problems, stepwise

selection of interaction terms has been rare. Available methods typically consider adding interaction terms only among those predictors that have been selected for their main effects. In comparison, in each forward addition step, SODA evaluates the overall contribution of a predictor including both its main effects and its interactions with selected predictors. Under some regularity conditions, we establish the screening consistency of the forward step and the individual term selection consistency of the backward step of SODA under high-dimensional settings.

An interesting and useful extension of SODA is for variable selection in general index models (Li, 1991; Cook, 2007; Jiang and Liu, 2014). These models assume that the response Y (either discrete or continuous) depends on a subspace of $\mathbf{X}$ through an unknown (nonlinear) link function. The most popular method for estimating the subspace is the sliced inverse regression (SIR) method (Li, 1991), which estimates $\mathrm{Cov}[\mathbb{E}(\mathbf{X} \mid Y)]$ by first slicing (discretizing) the response variable Y. We note that after the discretization of Y, we may also employ SODA for variable selection and model fitting. We call this extension the Sliced SODA (S-SODA). Compared with SIR and its variations (see Jiang and Liu (2014) for more references), S-SODA does not require the linearity condition and enjoys much improved robustness without much sacrifice in sensitivity.

The rest of the article is organized as follows. SODA and S-SODA are presented in full detail in Section 2. Theoretical properties of SODA are studied in Section 3. Simulation results are shown in Section 4 to compare performances of SODA and S-SODA with those of other methods. Applications of SODA and S-SODA to real data examples are demonstrated in Section 5. A short discussion in Section 6 concludes the article. Detailed theoretical proofs and additional empirical results are provided in the Supplementary Materials.

2 Models and Methods

2.1 Quadratic logistic regression and extended BIC

We consider the K-class classification problem. Let $Y \in \{1, \ldots, K\}$ denote the class label, let $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ be a vector of p predictors, and let $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ denote n independent observations on $(\mathbf{X}, Y)$. When p is large, usually only a small proportion of predictors have predictive power on Y. Let $\mathcal{P}$ denote the set of relevant predictors, and let $\mathcal{P}^c = \{1, \ldots, p\} \setminus \mathcal{P}$ be noise ones. That is,

$$\Pr\left(Y \mid \mathbf{X}_{\mathcal{P}}, \mathbf{X}_{\mathcal{P}^c}\right) = \Pr\left(Y \mid \mathbf{X}_{\mathcal{P}}\right).$$

We consider the following logistic model:

$$\Pr\left(Y = k \mid \mathbf{X}, \boldsymbol{\theta}\right) = \frac{\exp\left[\delta_k\left(\mathbf{X} \mid \boldsymbol{\theta}\right)\right]}{1 + \sum_{l=1}^{K-1} \exp\left[\delta_l\left(\mathbf{X} \mid \boldsymbol{\theta}\right)\right]}, \quad k = 1, \ldots, K, \tag{2}$$

where $\delta_k\left(\mathbf{X} \mid \boldsymbol{\theta}\right)$ is the discriminant function for class k and $\boldsymbol{\theta}$ denotes the vector of parameters. Choosing class K as the baseline class so that $\delta_K\left(\mathbf{X} \mid \boldsymbol{\theta}\right) = 0$, we assume that

$$\delta_k\left(\mathbf{X} \mid \boldsymbol{\theta}\right) = \alpha_k + \boldsymbol{\beta}_k^T \mathbf{X} + \mathbf{X}^T \mathbf{A}_k \mathbf{X}, \quad \text{for } k = 1, \ldots, K-1. \tag{3}$$

Since $\mathbf{X}$ is conditioned on, we do not need to model the distribution of $\mathbf{X}_{\mathcal{P}}$ or $\mathbf{X}_{\mathcal{P}^c}$, which is both convenient and robust for variable selection. Special cases of this model include:

- Multinomial logistic regression (with $\mathbf{A}_k = \mathbf{0}$ for all k)
- LDA and QDA, where $p\left(\mathbf{X}_{\mathcal{P}} \mid Y\right)$ is multivariate normal distribution
- Discriminant analyses where $p\left(\mathbf{X}_{\mathcal{P}} \mid Y\right)$ is in the multivariate exponential family,

$$p\left(\mathbf{x}_{\mathcal{P}} \mid Y = k, \boldsymbol{\eta}\right) = h\left(\mathbf{x}_{\mathcal{P}}\right) g\left(\boldsymbol{\eta}_k\right) \exp\left(\boldsymbol{\eta}_k^T \mathbf{x}_{\mathcal{P}}\right).$$

To see the connection between QDA and model (2), it is noted that for QDA models,

$$\begin{aligned}\alpha_k &= \log(\pi_k/\pi_K) - \frac{1}{2}\left(\log|\boldsymbol{\Sigma}_k| - \log|\boldsymbol{\Sigma}_K| + \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \boldsymbol{\mu}_K^T\boldsymbol{\Sigma}_K^{-1}\boldsymbol{\mu}_K\right),\\ \boldsymbol{\beta}_k^T &= \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\mu}_K^T\boldsymbol{\Sigma}_K^{-1},\\ \mathbf{A}_k &= -\frac{1}{2}\left(\boldsymbol{\Sigma}_k^{-1} - \boldsymbol{\Sigma}_K^{-1}\right), \quad \text{for } k = 1,\ldots,K-1.\end{aligned}$$

Let $\mathcal{M}$ and $\mathcal{I}$ denote subsets of main effects and interaction pairs, respectively, and let $\mathcal{M}_0$ and $\mathcal{I}_0$ denote the corresponding true sets defined as

$$\mathcal{M}_0 = \{j : \exists\, k \text{ s.t. } \beta_{k,j} \neq 0\} \quad \text{and} \quad \mathcal{I}_0 = \{(i,j) : \exists k \text{ s.t. } A_{k,i,j} \neq 0\},$$

with k indicating the class label. Let $\mathcal{A} = \mathcal{M}_0 \cup \mathcal{I}_0$ denote the true set of all effects, and let $\mathcal{S} = \mathcal{M} \cup \mathcal{I}$. The true set of relevant predictors $\mathcal{P}$ can be derived from $\mathcal{A}$ as

$$\mathcal{P} = \mathcal{M}_0 \cup \{j : \exists\, i \text{ s.t. } (i,j) \in \mathcal{I}_0\}.$$

Our main objective is to infer $\mathcal{A}$, with a special interest in terms in $\mathcal{I}_0$.

Let $\boldsymbol{\theta}_{\mathcal{S}}$ denote the collection of all coefficients in model (3), whose 0's correspond to terms not in $\mathcal{S}$, and let $\boldsymbol{\theta}_{k,\mathcal{S}}$ denote the corresponding coefficients for class k. For a dataset $\{(\mathbf{x}_i, y_i) : i = 1,\ldots,n\}$, the log-likelihood for $\boldsymbol{\theta}_{\mathcal{S}}$ is denoted as $l_n(\boldsymbol{\theta}_{\mathcal{S}})$. Let $\mathbf{Z} \equiv (1, \mathbf{X}, \mathbf{X}\bigotimes\mathbf{X})$ be the augmented version of $\mathbf{X}$, containing intercept 1, main effects, and all interaction terms of $\mathbf{X}$. Let $\mathbf{z}_i$ be the i-th observation of $\mathbf{Z}$. Then $l_n(\boldsymbol{\theta}_{\mathcal{S}})$ takes the form of a logistic regression model in $\mathbf{Z}$:

$$l_n(\boldsymbol{\theta}_{\mathcal{S}}) = \sum_{i=1}^{n}\left\{\boldsymbol{\theta}_{y_i,\mathcal{S}}^T\mathbf{z}_i - \log\left(1 + \sum_{l=1}^{K-1}\exp\left(\boldsymbol{\theta}_{l,\mathcal{S}}^T\mathbf{z}_i\right)\right)\right\}.$$

Let $\tilde{\boldsymbol{\theta}}_{\mathcal{S}}$ denote the MLE of $\boldsymbol{\theta}_{\mathcal{S}}$. By Lemma 2 in the appendix, with high probability $l_n(\boldsymbol{\theta}_{\mathcal{S}})$ is convex and $\tilde{\boldsymbol{\theta}}_{\mathcal{S}}$ can be obtained by the Newton-Raphson algorithm. Let $\boldsymbol{\theta}_0$ denote the true parameter vector. Theorem 1 illustrates the consistency of $\tilde{\boldsymbol{\theta}}_{\mathcal{S}}$ for any reasonable set $\mathcal{S}$.

Theorem 1. *Under conditions C1 ∼ C4 in Section 3, as* $n \to \infty$,

$$\max_{\mathcal{S} \supset \mathcal{A},\, |\mathcal{S}| \leq Q} \left\| \tilde{\boldsymbol{\theta}}_{\mathcal{S}} - \boldsymbol{\theta}_0 \right\|_2 = O_p\left(n^{-1/2+\xi}\right), \tag{4}$$

for any constants $0 < \xi < 1/2$ *and* $Q \geq |\mathcal{A}|$ *independent of* n.

In high-dimensional settings, the Bayesian information criterion (BIC) (Schwarz, 1978) is too liberal and tends to over-select (Broman and Speed, 2002). Chen and Chen (2008) proposed extended BIC (EBIC) and showed it to be consistent for linear regression models under high-dimensional settings. The EBIC for the set $\mathcal{S}$ of predictors is defined as

$$\mathrm{EBIC}_\gamma(\mathcal{S}) = -2\, l_n\left(\tilde{\boldsymbol{\theta}}_{\mathcal{S}}\right) + |\mathcal{S}| \log n + 2\gamma\ |\mathcal{S}| \log p, \tag{5}$$

where $|\mathcal{S}|$ is the size of set $\mathcal{S}$, and γ is a tuning parameter. The selection of γ depends on the relative order of n and p, and some heuristics are discussed in Section 2.5. Let $\tilde{\mathcal{S}}_{\mathrm{EBIC}}$ be the selected set of predictors minimizing the EBIC, and let Q be any positive constant greater than constant p_0 in condition (C1) in section 3. Then,

$$\tilde{\mathcal{S}}_{\mathrm{EBIC}} = \underset{\mathcal{S}:\, |\mathcal{S}| \leq Q}{\arg\min}\ \mathrm{EBIC}_\gamma(\mathcal{S}), \tag{6}$$

where $|\mathcal{S}|$ denotes the size of set $\mathcal{S}$. The asymptotic property of $\tilde{\mathcal{S}}_{\mathrm{EBIC}}$ is shown by the following theorem.

Theorem 2. (EBIC criterion consistency) *Under conditions C1 ∼ C4 in Section 3,* $\tilde{\mathcal{S}}_{EBIC}$ *is a consistent estimator of* $\mathcal{A}$,

$$\Pr\left(\tilde{\mathcal{S}}_{EBIC} = \mathcal{A}\right) \to 1, \ \text{as } n \to \infty,$$

for any $\gamma > 2 - 1/(2\kappa)$, *where* $p \leq n^\kappa$.

By treating our model as a logistic regression on $(\mathbf{Z}, Y)$, Theorem 2 follows directly from the asymptotic consistency of EBIC for generalized linear models (GLM), which was proved in Chen and Chen (2012) and Foygel and Drton (2011) in both fixed and random design contexts. We thus omit its proof. Different from Chen and Chen (2012) and Foygel and Drton (2011), here we require $\gamma > 2 - 1/(2\kappa)$ instead of $\gamma > 1 - 1/(2\kappa)$ to penalize additional model flexibility caused by the inclusion of interaction terms.

2.2 Stepwise variable and interaction selection

In practice it is infeasible to enumerate all possible $\mathcal{S}$ to find the one that minimizes the EBIC. For a closely related generalized linear model variable selection problem, Chen and Chen (2012) and Foygel and Drton (2011) used Lasso (Tibshirani, 1996) to obtain a solution path of predictor sets, and chose along the path the set with the lowest EBIC. However, this method also fails numerically under the high-dimensional setting for QDA, in which there are $O(p^2)$ candidate interaction terms. Furthermore, Lasso's variable selection consistency for logistic regression requires the incoherence condition (Ravikumar et al., 2010), which can be easily violated due to correlations among interaction terms and their corresponding main effect terms. The IIS procedure (Fan et al., 2015) requires the estimation of $p \times p$ precision matrix, which is by itself a difficult problem. If the related and unrelated predictors are moderately correlated, IIS's marginal screening strategy has difficulties in filtering out noise predictors. We propose here the stepwise procedure SODA, which consists of three stages: (1) preliminary forward main effect selection; (2) forward variable selection (considering both main and interaction effects); and (3) backward elimination.

1. **Preliminary main effect selection**: This step is the same as that in the standard stepwise regression method. Let $\mathcal{M}_t$ denote the selected set of main effects at step t.

SODA starts with $\mathcal{M}_1 = \emptyset$ and iterates the operations below until termination.

(a) For each predictor $j \notin \mathcal{M}_t$, create a new candidate set $\mathcal{M}_{t,j} = \mathcal{M}_t \cup \{j\}$.

(b) Find the predictor j with lowest $\text{EBIC}_\gamma(\mathcal{M}_{t,j})$. If $\text{EBIC}_\gamma(\mathcal{M}_{t,j}) < \text{EBIC}_\gamma(\mathcal{M}_t)$, continue with $\mathcal{M}_{t+1} = \mathcal{M}_{t,j}$, otherwise terminate with $\tilde{\mathcal{M}}_F$ and go to 2.

2. **Forward variable addition (both main and interaction effects):** Let $\mathcal{C}_t$ denote the selected set of predictors at step t, and let $\mathcal{S}_t = \tilde{\mathcal{M}}_F \cup \mathcal{C}_t \cup (\mathcal{C}_t \times \mathcal{C}_t)$ denote the set of terms induced by $\mathcal{C}_t$. SODA starts with $\mathcal{C}_1 = \emptyset$ and iterates the operations below until termination.

(a) For each $j \notin \mathcal{C}_t$, create a candidate set $\mathcal{C}_{t,j} = \mathcal{C}_t \cup \{j\}$ and let $\mathcal{S}_{t,j} = \tilde{\mathcal{M}}_F \cup \mathcal{C}_{t,j} \cup (\mathcal{C}_{t,j} \times \mathcal{C}_{t,j})$.

(b) Find the predictor j with lowest $\text{EBIC}_\gamma(\mathcal{S}_{t,j})$. If $\text{EBIC}_\gamma(\mathcal{S}_{t,j}) < \text{EBIC}_\gamma(\mathcal{S}_t)$, continue with $\mathcal{C}_{t+1} = \mathcal{C}_{t,j}$, otherwise terminate with $\tilde{\mathcal{C}}_F$ and go to 3.

3. **Backward elimination**: Let $\mathcal{S}_t$ denote the selected set of individual terms at step t of backward stage. SODA starts with $\mathcal{S}_1 = \tilde{\mathcal{M}}_F \cup \tilde{\mathcal{C}}_F \cup \left(\tilde{\mathcal{C}}_F \times \tilde{\mathcal{C}}_F\right)$ and iterate the operations below until termination.

(a) For each main or interaction term $j \in \mathcal{S}_t$ (e.g. $j = 1$ or $j = (1, 2)$), create a candidate set $\mathcal{S}_{t,j} = \mathcal{S}_t \setminus \{j\}$.

(b) Find term j with lowest $\text{EBIC}_\gamma(\mathcal{S}_{t,j})$. If $\text{EBIC}_\gamma(\mathcal{S}_{t,j}) < \text{EBIC}_\gamma(\mathcal{S}_t)$, remove term j, otherwise terminate and retain set $\tilde{\mathcal{S}} = \mathcal{S}_t$.

Stepwise methods had been primary tools for conducting variable selection in regression problems long before the recent development of Lasso-type methods. The forward

stepwise procedure has also been considered for variable screening for linear regressions in high-dimensional settings (Wasserman and Roeder, 2009; Wang, 2009). When considering interactions, a typical approach is to examine only those among the variables that have been deemed significant due to their main effects. However, Stage 2 of SODA for forward variable addition is different. After the preliminary selection of Stage 1, in Stage 2 SODA keeps track of a new set of variables $\mathcal{C}_t$, of which all main and quadratic terms are considered in the model. In other words, at each step SODA evaluates the EBIC for the overall effect of adding one predictor instead of one individual term. Thus, choosing one variable to add in the forward variable selection stage is of order $O(p)$, and the whole stage is of order $O(ps)$, where s is the number of truly relevant predictors. A naive method that searches through all individual terms is of order $O\left(p^2s^2\right)$. Another important feature of SODA is that each backward step only eliminates one individual term instead of all terms related to one predictor. In other words, SODA selects individual main and interaction effect terms without any nesting requirements.

Our theory shows that the forward variable addition step is sufficient for SODA to achieve the screening consistency. However, the number of parameters and the EBIC penalization in this forward step increases quadratically with the cardinality of $\mathcal{C}_t$. Therefore it can be hard to add predictors with only weak main effects. To optimize the empirical performance, we include the preliminary main effect selection stage to identify predictors with only weak main effects.

2.3 Sliced SODA (S-SODA) for general index models

In his seminal work on nonlinear dimension reduction, Li (1991) proposed a semi-parametric index model of the form

$$Y = f\left(\boldsymbol{\beta}_1^T\mathbf{X}, \boldsymbol{\beta}_2^T\mathbf{X}, \ldots, \boldsymbol{\beta}_d^T\mathbf{X}, \varepsilon\right), \tag{7}$$

where f is an unknown function and ε is random error independent of $\mathbf{X}$, and the sliced inverse regression (SIR) method to estimate the sufficient dimension reduction (SDR) space spanned by the directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$. SIR starts by discretizing the response variable Y into H intervals (aka "slices") and estimating $\mathrm{Cov}\left(\mathbb{E}\left(\mathbf{X} \mid Y\right)\right)$ via:

$$\widehat{\mathrm{Cov}}\left(\mathbb{E}\left(\mathbf{X} \mid Y\right)\right) = \frac{1}{n}\sum_{h=1}^{H} n_h \left(\mathbf{x}_h - \bar{\mathbf{x}}\right)\left(\mathbf{x}_h - \bar{\mathbf{x}}\right)^T.$$

It then uses the first d eigenvectors of $\left(\widehat{\mathrm{Cov}}\left(\mathbf{X}\right)\right)^{-1}\widehat{\mathrm{Cov}}\left(\mathbb{E}\left(\mathbf{X} \mid Y\right)\right)$ (corresponding to the top d eigenvalues) to estimate the SDR directions when d is known. In practice, however, d is often unknown and needs to be estimated. Since the estimation of SDR does not automatically lead to variable selection, several methods have been developed to do simultaneous dimension reduction and variable selection for index models. For example, Li et al. (2005) designed a backward subset selection method, and Li (2007) developed the sparse SIR (SSIR) algorithm to obtain shrinkage estimates of the SDR directions under L_1 norm. Motivated by stepwise regression for linear models, Zhong et al. (2012) proposed a forward stepwise procedure called correlation pursuit (COP) for index models. Lin et al. (2015) found the necessary and sufficient condition for SIR to be consistent in high-dimensional settings and introduced a diagonal thresholding method, DT-SIR, for variable selection. Lin et al. (2016) proposed a new formulation of the SIR estimation and a direct application of Lasso for variable selection with index models.

Szretter and Yohai (2009) showed that the SIR procedure is equivalent to estimating the d-dimensional subspace by the maximum likelihood method under the LDA model for $\mathbf{X}$ and the discretized Y, i.e., $[\mathbf{X} \mid s(Y) = h] \sim \mathcal{N}(\boldsymbol{\mu}_h, \Sigma)$, for $h = 1, \ldots, H$. In high dimensions when many irrelevant predictors are present, we let $\mathbf{X}_{\mathcal{P}}$ denote the set of predictors that have non-zero coefficients in $\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_d^T$. Then, we can formulate the variable selection problem as

$$\begin{aligned} \mathbf{X}_{\mathcal{P}} \mid s(Y) = h &\sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}), \\ \mathbf{X}_{\mathcal{P}^c} \mid \mathbf{X}_{\mathcal{P}}, s(Y) = h &\sim \mathcal{N}\left(\boldsymbol{a} + \boldsymbol{B}^T \mathbf{X}_{\mathcal{P}}, \boldsymbol{\Sigma}_0\right), \end{aligned} \quad h = 1, \ldots, H. \tag{8}$$

Thus, those variable selection methods reviewed in Section 1 for LDA models can also be applied to do variable selection for general index models.

The aforementioned SIR-based methods consider primarily the information from the first conditional moment, $\mathbb{E}(\mathbf{X} \mid Y)$, and tend to miss important variables with second-order effects. In order to overcome this problem, Jiang and Liu (2014) proposed SIRI, which utilizes both the first and the second conditional moments to select variables. SIRI augments the model in (8) with slice-specific covariance matrix Σ_h and makes it a QDA model:

$$\begin{aligned} \mathbf{X}_{\mathcal{P}} \mid s(Y) = h &\sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \\ \mathbf{X}_{\mathcal{P}^c} \mid \mathbf{X}_{\mathcal{P}}, s(Y) = h &\sim \mathcal{N}\left(\boldsymbol{a} + \boldsymbol{B}^T \mathbf{X}_{\mathcal{P}}, \boldsymbol{\Sigma}_0\right), \end{aligned} \quad h = 1, \ldots, H. \tag{9}$$

Jiang and Liu (2014) showed that SIRI is a consistent variable selection procedure for model (9), and also for a more general class of models satisfying the following linearity and constant variance conditions. In fact, all the aforementioned methods require either or both of the conditions imposed on the irrelevant variables $\mathbf{X}_{\mathcal{P}^c}$.

Linearity condition: $\mathbb{E}(\mathbf{X}_{\mathcal{P}^c} \mid \mathbf{X}_{\mathcal{P}})$ is linear in $\mathbf{X}_{\mathcal{P}}$.

Constant variance condition: $\mathrm{Cov}(\mathbf{X}_{\mathcal{P}^c} \mid \mathbf{X}_{\mathcal{P}})$ is a constant.

When the linearity and constant variance conditions approximately hold, SIRI and other SIR-based methods usually enjoy excellent empirical performances. However, when either condition is violated, performances of these methods deteriorate rapidly. Note that under the QDA model (9), $\Pr(s(Y) = h \mid \mathbf{X})$ follows a multi-category logistic regression model as in equation (2) and $\Pr(s(Y) \mid \mathbf{X}) = \Pr(s(Y) \mid \mathbf{X}_{\mathcal{P}})$, which suggests that one can select $\mathcal{P}$ using SODA. We thus propose the procedure sliced-SODA (S-SODA), which starts with the same procedure as SIR by sorting the samples in the ascending order of y_i, and partitioning them into H equal slices (in terms of ranks). It then applies SODA to data $\{(s_i, \mathbf{x}_i)\}_{i=1}^n$, where s_i denote the slice index for y_i. S-SODA finally outputs the main and interaction terms selected by SODA, as well as the set $\hat{\mathcal{P}}$ that contains the variables involved in these selected terms. S-SODA uses a logistic regression model with quadratic terms to capture the relationship between the discretized Y and covariates $\mathbf{X}$, which is implied by SIRI's working model (9) but without the conditional Gaussian assumption. Although in practice S-SODA worked well for all index models we tested, we are only able to prove its variable selection consistency for the logistic regression model (2) for $[s(Y) \mid \mathbf{X}_{\mathcal{P}}]$ under some technical assumptions in Section 3. Compared with SIRI (Jiang and Liu, 2014), S-SODA is much more robust, being effective for a wide class of models without requiring the linearity and constant variance conditions.

2.4 Post-selection prediction for continuous response

S-SODA conducts variable selection for semi-parametric model (7) without knowing the true functional form of the link function. After variable selection, it is of interest to predict the response variable $\tilde{y}$ for a new observation $\tilde{\mathbf{x}}$ of the predictor vector. Suppose our training data consist of n independent observations $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$. Let $\tilde{\mathcal{S}}$ denote the selected set of terms by S-SODA, and let $\tilde{\mathcal{P}}$ denote the set of predictors with any term in $\tilde{\mathcal{S}}$, which is the

S-SODA estimate of $\mathcal{P}$. Let $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_H)$, $\hat{\boldsymbol{\Sigma}} = \left(\hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\Sigma}}_H\right)$, where $\hat{\boldsymbol{\mu}}_h$ and $\hat{\boldsymbol{\Sigma}}_h$ are respectively the sample mean vector and covariance matrix of $\mathbf{X}_{\tilde{\mathcal{P}}}$ in slice h. Note that $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are MLEs of parameters in model $[\mathbf{X}_{\mathcal{P}} \mid s(Y) = h] \sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$. Inverting the model by the Bayes rule, we have

$$\Pr(s(Y) = h \mid \mathbf{X}_{\tilde{\mathcal{P}}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{\mathcal{N}(\mathbf{X}_{\tilde{\mathcal{P}}} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}{\sum_{l=1}^{H} \mathcal{N}(\mathbf{X}_{\tilde{\mathcal{P}}} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}, \quad h = 1, \ldots, H.$$

The conditional expectation $\mathbb{E}[Y \mid \mathbf{X}_{\tilde{\mathcal{P}}}]$ is a reasonable prediction of Y given $\mathbf{X}_{\tilde{\mathcal{P}}}$:

$$\begin{aligned} \mathbb{E}[Y \mid \mathbf{X}_{\tilde{\mathcal{P}}}] &= \sum_{h=1}^{H} \mathbb{E}[Y \mid s(Y) = h, \mathbf{X}_{\tilde{\mathcal{P}}}] \Pr(s(Y) = h \mid \mathbf{X}_{\tilde{\mathcal{P}}}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \sum_{h=1}^{H} \frac{\mathbb{E}[Y \mid s(Y) = h, \mathbf{X}_{\tilde{\mathcal{P}}}] \cdot \mathcal{N}(\mathbf{X}_{\tilde{\mathcal{P}}} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)}{\sum_{l=1}^{H} \mathcal{N}(\mathbf{X}_{\tilde{\mathcal{P}}} \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \end{aligned}$$

Replacing the unknown parameters in $\mathbb{E}[Y \mid \mathbf{X}_{\tilde{\mathcal{P}}}]$ by their estimates, we have the prediction

$$\hat{Y} = \hat{\mathbb{E}}[Y \mid \mathbf{X}_{\tilde{\mathcal{P}}}] = \sum_{h=1}^{H} \frac{\hat{M}_h \cdot \mathcal{N}\left(\mathbf{X}_{\tilde{\mathcal{P}}} \mid \hat{\boldsymbol{\mu}}_h, \hat{\Sigma}_h\right)}{\sum_{l=1}^{H} \mathcal{N}\left(\mathbf{X}_{\tilde{\mathcal{P}}} \mid \hat{\boldsymbol{\mu}}_l, \hat{\Sigma}_l\right)}, \tag{10}$$

where $\hat{M}_h$ is the sample mean of the response Y in slice h. $\hat{M}_h$ can be considered as the zero-th order approximation to $\mathbb{E}[Y \mid s(Y) = h, \mathbf{X}_{\tilde{\mathcal{P}}}]$, in the sense that $\hat{M}_h$ is independent of $\mathbf{X}_{\tilde{\mathcal{P}}}$. A more sophisticated model is to consider the first-order approximation that models $\mathbb{E}[Y \mid s(Y) = h, \mathbf{X}_{\tilde{\mathcal{P}}}]$ as a linear combination of $\mathbf{X}_{\tilde{\mathcal{P}}}$ in each slice.

2.5 Implementation issues of SODA

Sections 3 characterizes asymptotic properties of the EBIC and SODA, and provides some theoretical insights for choosing the tuning parameter γ of EBIC. However, these asymptotic results are not directly usable . In practice, we propose to use a 10-fold cross-validation

(CV) procedure for selecting γ from the set $\{0, 0.5, 1.0\}$. For simulation studies and real data analyses in Sections 4 and 5, we fix $\gamma = 0.5$ as suggested in Chen and Chen (2012) in order to make SODA more easily comparable with Lasso-EBIC they studied.

The forward variable addition stage terminates if EBICs of all candidate models are larger than the EBIC of the current model. Therefore, the screening depth of the forward stage is determined by the EBIC. In Theorem 3, we show that this procedure is asymptotically screening consistent; namely, the truly relevant terms will be all included by the end of the forward stage. Nevertheless, SODA is not sensitive to adding more terms in the forward stage since those unrelated terms will be eventually eliminated in the backward stage. Missing one relevant term is usually more harmful than including one noise term. Therefore, to optimize the empirical performance, we let SODA continue the forward variable addition for p_f steps after the step that fails to decrease EBIC (default $p_f = 3$).

3 Theoretical properties of SODA

To gain some theoretical insights for SODA, we assume the following regularity conditions:

(C1) The divergence speed of p is bounded above by $p \leq n^{\kappa}$ for some $\kappa > 0$, and the size of the true predictor set $\mathcal{P}$ is bounded as $|\mathcal{P}| \leq p_0$ for a fixed integer p_0.

(C2) Magnitudes of true coefficients in $\boldsymbol{\theta}_{\mathcal{A}}$ are bounded above and below by constants, namely there exist positive constants $\theta_{\max} > \theta_{\min} > 0$ such that

$$\theta_{\min} \leq \min\{|\theta_j| : j \in \mathcal{A}\} \leq \max\{|\theta_j| : j \in \mathcal{A}\} \leq \theta_{\max}.$$

(C3) The interaction effect terms $X_{j_1}X_{j_2}, 1 \leq j_1 \leq j_2 \leq p$ are sub-exponential, i.e., there are positive constants C_1 and C_2 such that,

$$\Pr\left(|X_{j_1}X_{j_2} - \mathbb{E}\left[X_{j_1}X_{j_2}\right]| > t\right) \leq C_1 \exp\left(-C_2 t\right) \qquad \text{for all } t > 0.$$

Note that X_j^2 being sub-exponential implies that X_j is sub-Gaussian (Vershynin, 2012), and consequently sub-exponential.

(C4) Let $\mathrm{Cov}(\mathbf{Z})$ denote the covariance matrix of $\mathbf{Z}$, where $\mathbf{Z} \equiv (1, \mathbf{X}, \mathbf{X} \bigotimes \mathbf{X})$. There exist constants $0 < \tau_1 < \tau_2 < \infty$ such that

$$\tau_1 \leq \lambda_{\min}(\mathrm{Cov}(\mathbf{Z})) < \lambda_{\max}(\mathrm{Cov}(\mathbf{Z})) \leq \tau_2,$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a matrix. This is a commonly imposed requirement used to ensure identifiability.

We show that the forward variable addition stage (Stage 2 of SODA) is already screening consistent. To proceed, we need to define the following concept to study the stepwise detectability of true predictors in $\mathcal{P}$. Let $\boldsymbol{\theta}_{\mathcal{S}}^*$ denote the population version of the risk minimizer,

$$\boldsymbol{\theta}_{\mathcal{S}}^* = \arg\min_{\boldsymbol{\theta}_{\mathcal{S}}} \mathbb{E}\left[-\log p\left(Y \mid \mathbf{X}, \boldsymbol{\theta}_{\mathcal{S}}\right)\right],$$

where the expectation is over the joint distribution of $(Y, \mathbf{X})$. Let vector $\boldsymbol{\theta}_{\mathcal{S}}^{j*}$ be parameters in $\boldsymbol{\theta}_{\mathcal{S}}^*$ associated with predictor X_j. The stepwise detectable condition is necessary for the screening consistency of the forward variable addition stage.

Definition 1. (Stepwise detectable condition) A set of predictors $\mathcal{C}_1$ is stepwise detectable given $\mathcal{C}_2$ if $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$, and for any set $\mathcal{C}$ satisfying $\mathcal{C} \supset \mathcal{C}_2$ and $\mathcal{C} \not\supset \mathcal{C}_1$, there exist constants $\theta_{\max} > \theta_{\min} > 0$, such that

$$\theta_{\min} \leq \max_{j \in \mathcal{C}^c \cap \mathcal{C}_1} \left\| \boldsymbol{\theta}_{\mathcal{S}_{\mathcal{C}\cup\{j\}}}^{j*} \right\|_\infty \leq \theta_{\max},$$

where $\mathcal{S}_{\mathcal{C}\cup\{j\}} = \mathcal{M}^j \cup \mathcal{I}^j$ with $\mathcal{M}^j = \mathcal{C} \cup \{j\}$ and $\mathcal{I}^j = \mathcal{M}^j \times \mathcal{M}^j$, and $\|\cdot\|_\infty$ denotes the L_∞ norm. Let $\mathcal{T}_m = \left\{j : \text{predictor } j \text{ is stepwise detectable given } \cup_{i=0}^{m-1} \mathcal{T}_{i-1}\right\}$ and $\mathcal{T}_0 = \emptyset$. The set of true predictors $\mathcal{P}$ is said to be stepwise detectable if $j \in \cup_{i=1}^{\infty} \mathcal{T}_i$ for all $j \in \mathcal{P}$.

In other words, if the current selection $\mathcal{C}$ contains $\mathcal{C}_2$, then there always exist detectable predictors conditioning on currently selected variables until we include all the predictors indexed by $\mathcal{C}_1$. A true predictor $j \in \mathcal{P}$ is not stepwise detectable either because it perfectly correlates with some other terms, or its effects can only be detected conditioning on some other stepwise undetectable terms.

We give an example to illustrate scenarios when the stepwise detectable condition may or may not hold. Suppose there are two true jointly normal relevant predictors X_1 and X_2 with means μ_1 and μ_2, and there is only one interaction term X_1X_2 in model (2), i.e. $\mathcal{A} = \{(1,2)\}$. $\mathcal{P}$ is not stepwise detectable if both $\mu_1 = 0$ and $\mu_2 = 0$. Starting from empty set $\emptyset$, the forward procedure will not add X_1 or X_2 into the model, because there is no main effect for X_1 and X_2 and the interaction term X_1X_2 does not correlate with marginal terms X_1 and X_2 ($\mathrm{Cov}(X_1, X_1X_2) = 0$ and $\mathrm{Cov}(X_2, X_1X_2) = 0$). However, if either $\mu_1 \neq 0$ or $\mu_2 \neq 0$, $\mathcal{P} = \{1, 2\}$ is stepwise detectable.

Let $\tilde{\mathcal{S}}_F = \tilde{\mathcal{M}}_F \cup \tilde{\mathcal{C}}_F \cup \left(\tilde{\mathcal{C}}_F \times \tilde{\mathcal{C}}_F\right)$ denote the selected set of terms at the end of the forward variable addition stage. It is unrealistic to require $\tilde{\mathcal{S}}_F = \mathcal{A}$. However, it should be demanded that $\tilde{\mathcal{S}}_F \supseteq \mathcal{A}$, i.e. $\tilde{S}_F$ contains all relevant terms. We define the forward stage to be screening consistent if $\Pr\left(\tilde{\mathcal{S}}_F \supseteq \mathcal{A}\right) \to 1$. We also do not want the size of $\tilde{\mathcal{S}}_F$ to be too large, otherwise forward variable addition loses its purpose. The screening consistency of forward stage is established by the following theorem.

Theorem 3. (Forward stage screening consistency) *If conditions C1 $\sim$ C4 hold, and all predictors in $\mathcal{P}$ are stepwise detectable, then the forward variable addition stage finishes in finite number of steps and is screening consistent. In particular, as $n \to \infty$,*

$$\Pr\left(\left|\tilde{\mathcal{C}}_F\right| \leq Q\right) \to 1, \quad \textit{and} \quad \Pr\left(\tilde{\mathcal{C}}_F \supseteq \mathcal{P}\right) \to 1,$$

where $Q = \left\lceil 8\lambda_1^{-1}\theta_{\min}^{-2} \log K \right\rceil$, λ_1 is a positive constant defined in Lemma 2 in appendix, K

is the number of classes, and $\theta_{\min}$ *is a positive constant defined in condition C2.*

In other words, asymptotically $\tilde{\mathcal{C}}_F$ contains all predictors in $\mathcal{P}$, which implies $\tilde{\mathcal{S}}_F \supseteq \mathcal{A}$, and the forward stage stops in finite number of steps. We show in the following theorem two uniform bounds guaranteeing that all unrelated terms will be eliminated and all related terms will be kept in the backward stage.

Theorem 4. (Uniform bound of EBIC in backward stage) *Fix any positive constant* $Q > 0$. *Under conditions C1* $\sim$ *C4, as* $n \to \infty$,

$$\Pr\left(\max_{\mathcal{S}\supsetneq\mathcal{A}:|\mathcal{S}|\leq Q}\min_{j\in\mathcal{S}\setminus\mathcal{A}}\left\{EBIC_\gamma\left(\mathcal{S}\setminus\{j\}\right)-EBIC_\gamma\left(\mathcal{S}\right)\right\}<0\right)\to 1, \tag{11}$$

and

$$\Pr\left(\min_{\mathcal{S}\supset\mathcal{A}:|\mathcal{S}|\leq Q}\min_{j\in\mathcal{A}}\left\{EBIC_\gamma\left(\mathcal{S}\setminus\{j\}\right)-EBIC_\gamma\left(\mathcal{S}\right)\right\}<0\right)\to 0, \tag{12}$$

for any constant $\gamma > Q - |\mathcal{A}| - (2\kappa)^{-1}$.

Eq (11) implies that if $\mathcal{S} \supsetneq \mathcal{A}$ and $|\mathcal{S}| \leq Q$, there will be at least one unrelated term $j \in \mathcal{S} \cap \mathcal{A}^c$ such that removing j from $\mathcal{S}$ leads to lower EBIC. Eq (12) implies that if $\mathcal{S} \supset \mathcal{A}$ and $|\mathcal{S}| \leq Q$, there is no related term $j \in \mathcal{A}$ such that removing j from $\mathcal{S}$ leads to a lower EBIC. In summary, as $n \to \infty$, with probability tending to 1, no related term will be eliminated and all unrelated terms will be eliminated in the backward stage until $\tilde{\mathcal{S}} = \mathcal{A}$. Theorem 4 requires candidate sets to have finite size ($|\mathcal{S}| \leq Q$), which is proved by Theorem 3 to hold asymptotically for the starting set of the backward stage $\tilde{\mathcal{S}}_F$. Hence, the combination of Theorems 3 and 4 establishes the model selection consistency of SODA. Proofs of the theorems are in the on-line Supplementary Materials.

4 Simulation results

4.1 Logistic regression with interactions

We first compare performances of several methods on variable and interaction selections under the logistic regression framework. Besides SODA, we consider the backward procedure in Zhang and Wang (2011) (denoted as ZW), the forward-backward method in Murphy et al. (2010) (denoted as MDR), hierNet in Bien et al. (2013) and IIS-SQDA in Fan et al. (2015). Both ZW and MDR assume joint normality between $\mathbf{X}_{\mathcal{P}}$ and $\mathbf{X}_{\mathcal{P}^c}$. The method hierNet is a Lasso-like procedure to detect multiplicative interactions between predictors under hierarchical constraints. For hierNet, we select the regularization parameter with the lowest CV error. We have also reported in the Supplementary Materials a comparison between SODA and Lasso-logistic for variable selections when the underlying logistic regression model has only linear main effects, and found that SODA was competitive with Lasso in all cases we tested and out-performed Lasso significantly when the "incoherence" condition (Ravikumar et al., 2010) was violated.

We first considered four simulation settings in Examples 1.1∼1.4 for the toy classification example introduced in Section 1 (see also Figure 1 for more details), and further examined two more simulation scenarios (Examples 1.5 and 1.6) in which the interaction effects and main effects are from different predictors. For Examples 1.1∼1.4, there are two classes ($K = 2$) and p predictors, among which X_1, X_2 and X_3 are relevant ones, i.e., $\mathcal{P} = \{1, 2, 3\}$, and other $p - 3$ predictors are irrelevant but correlated with the 3 relevant ones. Their oracle Bayes classification rule is to label an observation class 1 if $Q(\mathbf{x}) > 0$, and 0 otherwise, where

$$Q(\mathbf{x}) = 1.627 + X_1 - 0.6X_1^2 - 0.6X_3^2 - 0.7X_1X_2 - 0.7X_2X_3,$$

indicating that $\mathcal{A} = \{1, (1,1), (3,3), (1,2), (2,3)\}$, representing one linear effect (X_1) and four interaction effects (X_1^2, X_3^2, X_1X_2, and X_2X_3) without the hierarchy restriction. The first setting follows the multivariate normal model while the other three do not. Examples 1.1∼1.3 are of moderate dimension with p=50, and Example 1.4 simulates a high-dimensional scenario with p=1000.

For each simulation setting, we generated 100 datasets with 10 different sample sizes for each class, ranging linearly in log-scale from 100 to 1000. For SODA, hierNet, and IIS-SQDA, the set of selected predictors is defined as the union of all predictors appearing in the selected linear and interaction terms. We calculated the average number of false negatives and false positives for variable selection (VFN and VFP), main effect term selection (MFN and MFP), and interaction term selection (IFN and IFP).

To benchmark the classification accuracy, we also include the full model of LDA and QDA with all predictors, and the Oracle model that contains exactly the five true terms. The average classification test error rate (TE) of models is estimated by applying the trained model to 10,000 additional observations simulated from true models. OVerall results for Examples 1.1-1.4 are shown in Figure 2. The performances of IIS-SQDA, hierNet and SODA on individual term selections are shown in Table 2.

Example 1.1. *Multivariate Gaussian.* Irrelevant predictors were simulated as linear combinations of relevant ones as follows:

$$X_j = b_{j,0} + b_{j,1}X_l + b_{j,2}X_k + \varepsilon_j,\ \ j = 3, \ldots, 50,$$

where X_k and X_l were randomly selected from $\{X_1, X_2, X_3\}$, coefficients $b_{j,0}$, $b_{j,1}$ and $b_{j,2}$ were drawn from uniform distribution $U[-1,1]$, and $\varepsilon_j \sim \mathcal{N}(0,2)$.

As shown in Figure 2, for this example ZW, MDR, and SODA were all able to detect all relevant predictors as n increases, with both VFN and VFP being very low. They achieved

almost the Oracle classification accuracy. In contrast, IIS-SQDA and hierNet selected too many false positives, which resulted in high test error rates, and the number VFP+VFN increased with n. This strange phenomenon has also been observed by other researchers (Fan et al., 2015; Yu and Feng, 2014). As shown in Table 2, SODA selected individual terms nearly perfectly. HierNet is based on Lasso and IIS-SQDA uses elastic net. The variable selection consistency of Lasso and elastic net require the *Irrepresentable Condition* (Zhao and Yu, 2006) and the *Elastic Irrepresentable Condition* (Jia and Yu, 2010), which may not hold here. Moreover, it was observed that the cross-validation is too liberal for Lasso, leading to a large number of false positives (Yu and Feng, 2014). As expected, LDA and QDA without variable selection performed the worst.

Example 1.2. *Non-Gaussian irrelevant predictors.* Irrelevant variables were simulated to be quadratically dependent of relevant ones:

$$X_j = b_{j,0} + b_{j,1}X_k + b_{j,2}X_l + b_{j,3}X_k^2 + b_{j,4}X_l^2 + \varepsilon_j,\ j = 3, \ldots, 50, \tag{13}$$

where X_k and X_l were randomly selected from $\{X_1, X_2, X_3\}$, coefficients $b_{j,0}$, ..., $b_{j,4}$ were drawn from $U[-1, 1]$, and $\varepsilon_j \sim \mathcal{N}(0, 5)$. As shown in Figure 2, ZW and MDR selected 4 to 10 FP and FN predictors on average. IIS-SQDA and hierNet selected a large number of FP terms, as shown in Table 2, due to the correlation between relevant and irrelevant predictors as well as correlations between main and interaction terms.

Example 1.3. *Heteroskedastic errors.* Irrelevant we simulated as follows:

$$X_j = b_{j,1}X_k + b_{j,2}X_l + |X_k|\,\varepsilon_j,\ j = 3, \ldots, 50, \tag{14}$$

where X_k and X_l were randomly selected from $\{X_1, X_2, X_3\}$, coefficients $b_{j,1}$ and $b_{j,3}$ were drawn from $U[-1, 1]$, and $\varepsilon_j \sim \mathcal{N}(0, 1)$. It violates the constant variance assumption of ZW and MDR. Thus, ZW, MDR, IIS-SQDA and hierNet all performed poorly. In contrast,

SODA selected almost no VFP and VFN, and achieved near-Oracle prediction accuracy when $n \geq 200$.

Example 1.4. *High-dimensional and non-Gaussian.* Irrelevant predictors were simulated as follows. For $j \in \{4, \ldots, 100\}$, we drew 60% of the X_j's at random and simulated them from $\mathcal{N}(m_j, 1)$, $m_j \sim U[0, 1]$. The remaining 40% of the X_j's were simulated as non-linearly related to (X_k, X_l) similarly as (13) or (14), where k and l were randomly chosen from $\{1, 2, 3\}$. For $j \in \{101, \ldots, 1000\}$, we first drew all predictors from $\mathcal{N}(m_j, 1)$, and then randomly selected 40% of them and re-simulated each of the selected X_j as (13) or (14), where k and l are indexes uniformly drawn from $\{101, \ldots, 1000\}$. We changed ZW to a forward procedure since the backward procedure is not feasible when $p > n$. Results are shown in Figure 2 and Table 2. MDR results are not shown because it is unstable for highly correlated $\mathbf{X}$ matrices and usually keeps on adding new predictors until the estimation of covariance matrices become singular. Overall, SODA performed much better than ZW and IIS-SQDA, and the classification accuracy of SODA was almost the same as the Oracle model for $n > 100$. Figure 3 shows the running times in log-scale versus n for IIS-SQDA, ZW, and SODA, On average, IIS-SQDA took 800 minutes, ZW took 22 minutes, and SODA took 4 minutes to analyze one simulated dataset with $p = 1000$ and $n = 1000$. In contrast, hierNet did not finish the simulation experiments in 24 hours and is thus not included in the comparison.

Example 1.5. *Interactions only.* We simulated the scenario in which there are only interaction effects. In particular, we removed the main effect term X_1 from the previous classification rule and the new classification rule is to label an observation class 1 if $Q(\mathbf{x}) > 0$, where

$$Q(\mathbf{x}) = 1.777 - 0.6X_1^2 - 0.6X_3^2 - 0.7X_1X_2 - 0.7X_2X_3.$$

Example 1.6. *Anti-hierarchical interactions.* The terminology "anti-hierarchical" from Bien et al. (2013) refers to the scenario that the main effects and interaction effects are from different set of predictors. In this example, the classification function $Q(\mathbf{x})$ is

$$Q(\mathbf{x}) = 1.777 + X_4 - X_5 - 0.6X_1^2 - 0.6X_3^2 - 0.7X_1X_2 - 0.7X_2X_3.$$

For both Examples 1.5 and 1.6, we let $p = 50$ and let irrelevant predictors be simulated in the same way as Example 1.2. The results are shown in Figure 4. Overall, the results are similar to previous examples that SODA had fewer variable selection errors and lower TE rates than other methods.

4.2 Continuous-response index models

We conducted simulations to compare performances of several methods for variable selection in nonlinear models with continuous responses. Besides S-SODA, we considered all the five methods studied in Jiang and Liu (2014): Lasso, DC-SIS, hierNet, COP, and SIRI. DC-SIS was proposed in Li et al. (2012) as a sure independence screening procedure based on distance correlation, which has been shown to be capable of detecting relevant variables when interactions are present. The hierNet method in Bien et al. (2013) is a Lasso-like procedure to detect multiplicative interactions between predictors under hierarchical constraints. For SIRI and S-SODA, we equally partition $\{y_i\}_{i=1}^n$ into $H = 5$ slices. In order to improve SIRI's robustness, we consider a modified version of SIRI, termed as N-SIRI, which pre-processes $\mathbf{X}$ by marginally quantile-normalizing each predictor to the standard normal distribution.

We considered the following five simulation examples:

$$\begin{aligned}
\text{Example 2.1}: \quad & Y = 3X_1 + 1.5X_2 + 2X_3 + 2X_4 + 2X_5 + \sigma\epsilon, \\
\text{Example 2.2}: \quad & Y = X_1 + X_1X_2 + X_1X_3 + \sigma\epsilon, \\
\text{Example 2.3}: \quad & Y = X_1^2X_2/X_3^2 + \sigma\epsilon, \\
\text{Example 2.4}: \quad & Y = X_1/\exp(X_2 + X_3) + \sigma\epsilon, \\
\text{Example 2.5}: \quad & Y = X_1 + X_2 + (1 + X_3)^2\epsilon,
\end{aligned}$$

where $\sigma = 0.2$ and $\varepsilon \sim \mathcal{N}(0, 1)$ independent of $\mathbf{X}$. In each example, we simulated the predictors $\mathbf{X}$ with dimension $p = 1000$. In order to test robustness of the methods, we examined the following three scenarios:

- Scenario (a): $\mathbf{X}$ is simulated from multivariate Gaussian with correlation $0.5^{|i-j|}$. In this scenario the linearity and constant variance conditions hold.
- Scenario (b): Each predictor X_j, $j = 1, \ldots, p$, is simulated from the χ_1^2 distribution independently. In this scenario the linearity and constant variance conditions hold, but the distribution of $\mathbf{X}$ is non-normal.
- Scenario (c): $X_1, \ldots, X_{125}$ were simulated from multivariate Gaussian with correlation

$0.5^{|i-j|}$. For $X_{126}, \ldots, X_{1000}$, we simulated according to the following schemes:

$$\begin{aligned}
X_j &= X_{j-125}^2 + \varepsilon_j, \quad j = 126, \ldots, 250, \\
X_j &= \sqrt{|X_{j-250}|} + \varepsilon_j, \quad j = 251, \ldots, 375, \\
X_j &= \sin\left(X_{j-375}\right) + \varepsilon_j, \quad j = 376, \ldots, 500, \\
X_j &= \log\left(|X_{j-500}|\right) + \varepsilon_j, \quad j = 501, \ldots, 625, \\
X_j &= \exp\left(X_{j-625}\right) + \varepsilon_j, \quad j = 626, \ldots, 750, \\
X_j &= \exp\left(|X_{j-750}|\right) + \varepsilon_j, \quad j = 751, \ldots, 875, \\
X_j &= X_{j-875}^2 \varepsilon_j, \quad j = 876, \ldots, 1000.
\end{aligned}$$

For each simulation setting, we generated 100 datasets with sample size $n = 200$, and applied the aforementioned seven methods to each simulated dataset. For each method, the average number of false positives (FPs) and false negatives (FNs) were calculated over the 100 datasets. The results for the five examples are shown in Figure 5.

As expected, all the seven methods worked well for Example 2.1 in scenario (a), with low FPs and FNs, since the underlying structure is indeed linear Gaussian. For scenarios (b) and (c) with non-Gaussian predictors, DC-SCAD, hierNet, and SIRI generated more FPs and/or FNs than other methods. In general. SIRI performed the worst for this example. But with quantile-normalization, N-SIRI performed very competitively. S-SODA worked well for all the three scenarios, almost as good as LASSO.

In Examples 2.2$\sim$2.5 the relationships between Y and $\mathbf{X}$ is non-linear. Thus, as expected LASSO and DC-SCAD tended to miss important predictors, resulting in high number of FNs. The method hierNet could only detect second-order interactions such as X_1X_2 but failed to identify more complicated relationships such as $Y = X_1^2X_2/X_3^2$ and $Y = X_1/\exp\left(X_2 + X_3\right)$. COP only identified the information from the first conditional moment

$\mathbb{E}(\mathbf{X} \mid Y)$, and missed important variables with interaction or other second-order effects.

As expected, SIRI usually worked well for scenario (a). N-SIRI worked well for both scenarios (a) and (b) since the joint distribution of the predictors become multivariate Gaussian after quantile-normalization. For scenario (c), SIRI performed very poorly, while N-SIRI performed respectfully, although it still had more FPs and FNs than S-SODA. In contrast, S-SODA worked well for all three scenarios. These examples demonstrated the efficiency and robustness of S-SODA for variable selection in nonparametric index models.

4.3 Prediction of continuous surface

We consider three examples to test the performance of S-SODA in using formula (10) to predict Y, with $p = 1000$ predictors simulated in the same way as scenario (a) in the previous subsection. In order to visualize the relationship surfaces in a three-dimensional plot, each model only has 2 relevant predictors, i.e., $\mathbf{X}_{\mathcal{P}} = (X_1, X_2)$:

$$\begin{aligned} \text{Example 3.1:} \quad Y &= X_1 + X_2 + \sigma\epsilon, \\ \text{Example 3.2:} \quad Y &= X_1 / \exp(X_2) + \sigma\epsilon, \\ \text{Example 3.3:} \quad Y &= \left(1 + X_1^2 + X_2^2\right)^{-1} + \sigma\epsilon, \end{aligned}$$

where $\sigma = 0.2$ and $\epsilon \sim \mathcal{N}(0, 1)$. For each example we simulated $n = 500$ samples, and applied S-SODA to the simulated data. S-SODA correctly identified $\tilde{\mathcal{P}} = \{1, 2\}$. We further used formula (10) with $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ being the MLEs to predict $\hat{\mathbb{E}}[Y \mid \mathbf{X}_{\tilde{\mathcal{P}}}]$ with $H = 25$ slices for each example. The results are shown in Figure 6. Encouragingly, it is observed that even though we do not know the true functional form of $\mathbb{E}[Y \mid \mathbf{X}_{\mathcal{P}}]$, our prediction $\hat{\mathbb{E}}[Y \mid \mathbf{X}_{\tilde{\mathcal{P}}}]$ well captures the landscape of $\mathbb{E}[Y \mid \mathbf{X}_{\mathcal{P}}]$ in these three examples.

5 Real data analysis

We applied SODA, Lasso-Logistic, MDR and IIS-SQDA to a few real datasets to compare their performances. We did not include the ZW method due to its similarity to MDR. The classification accuracy of the selected models were evaluated by 10-fold cross-validation (CV) after the variable selection. For Lasso-Logistic and SODA, we used $\text{EBIC}_{0.5}$ as model selection criterion. We consider three datasets: (1) a Michigan lung cancer dataset analyzed in Efron (2009) with large $p > 5000$; (2) the Ionosphere dataset, with $p = 32$; and (3) the dataset Pumadyn with $p = 32$ with a continuous response. The Ionosphere dataset was downloaded from UCI Machine Learning Repository[1], and the Pumadyn dataset was downloaded from DELVE (Data for Evaluating Learning in Valid Experiments)[2].

5.1 Michigan lung cancer dataset

This dataset was published in Beer et al. (2002), in which researchers measured mRNA expression levels of $p = 5,217$ genes in tumor tissues of 86 lung cancer patients. Among the 86 patients, 62 are labeled as in "good status", and 26 in "bad status". The goal is to classify new patients into one of two statuses. Results on this dataset are summarized in Table 3. IIS-SQDA did not finish in 48 hours for this dataset, so we omitted its result.

In the solution path of Lasso-Logistic, the lowest $\text{EBIC}_{0.5}$ was achieved at 112.2 with 1 gene, and the corresponding CV error rate was 29%. SODA selected 2 main effects and 2 interaction effects with the $\text{EBIC}_{0.5}$ score at 69.8 and the CV error rate at 11%. Similar to the analysis of the prostate cancer dataset [Singh et al. (2002); see on-line Supplementary Materials], SODA worked much better than Lasso-Logistic for finding the minimum of

[1] `https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength`

[2] `http://www.cs.toronto.edu/~delve/data/pumadyn/desc.html`

$EBIC_{0.5}$ (69.8 vs 112.2). Comparing models reported by Lasso-Logistic and SODA, we see that interaction effects contribute substantially to improve the classification accuracy.

MDR failed to converge on this dataset. MDR selected as many genes as possible until the number of selected genes was the same as the number of samples in the smaller class (26) and achieved the CV error rate of 28%. The observation that MDR failed to converge for this large-p dataset and also for the prostate cancer dataset reported in the Supplementary Materials indicates that QDA variable selection methods under the joint normality assumption work poorly for high-dimensional real datasets.

5.2 Ionosphere dataset

This is a two-class classification problem with 351 samples and 32 predictors. Targets are "Good" and "Bad" radar returns from the ionosphere. "Good" returns are those showing evidence of some type of structure in the ionosphere, while "Bad" returns do not. The original dataset has 34 predictors, but the first 2 predictors have constant values and zero variance for the "Good" class, which will crash MDR and IIS-SODA algorithms. Therefore we eliminated the first 2 predictors for the analysis.

We applied Lasso-Logistic, MDR, IIS-SQDA and SODA to this dataset. Since the number of candidate predictors is not large, we also ran Lasso-Logistic with all main effect terms and $32 \times (32+1)/2 = 528$ interaction terms, which is referred to as Lasso-Logistic-2. Results are summarized in Table 4. In the solution path of Lasso-Logistic, the lowest $EBIC_{0.5}$ was achieved at 302.9 with 6 predictors, and the corresponding CV error rate was 14%. Lasso-Logistic-2 selected 2 main effect terms and 5 interaction terms with $EBIC_{0.5}$ 248.7 and CV error rate 8%. SODA selected 4 main effect and 4 interaction terms with the $EBIC_{0.5}$ score at 204.2 and CV error rate 6%. Again, SODA found smaller $EBIC_{0.5}$ value than both Lasso methods.

MDR method selected all 32 predictors, which is the full QDA model, and achieves cross-validation error rate 28%. IIS-SQDA selected 10 main effect and 96 interaction terms and achieved cross-validation error rate 16%. Comparing the CV error of the full QDA model selected by MDR and the SODA selected model, we see that variable selection substantially reduces the classification error rate for this real dataset.

5.3 Pumadyn dataset

This dataset was synthesized from a realistic simulation of the dynamics of a robotic arm. It has $n = 8192$ samples, $p = 32$ predictors, and a continuous response. The predictor set includes angular positions, velocities and torques of the robotic arm. The goal is to predict the angular acceleration of the robotic arm's links. The samples are split into 4500 in-samples for modeling training, and 3692 out-samples for model evaluation.

We trained the S-SODA model with $H = 20$ for this dataset, and made the predictions using formula (10). We also applied linear regression with Lasso selection with and without interaction terms, denoted as Linear and Linear-2, respectively. The results are summarized in Table 5. In the Lasso path of the linear models, the highest out-sample correlation $r = 0.477$ were achieved when selecting only 1 predictor (named tau4). S-SODA selected two predictors (tau4 and theta5) and achieved the out-sample correlation $r = 0.707$. Our predicted $\hat{\mathbb{E}}\left[Y \mid \mathbf{X}_{(\text{tau4,theta5})}\right]$ surfaces from linear model and S-SODA are shown in Figure 7. The interaction between predictors tau4 and theta5 was captured by S-SODA but not the linear model. From Table 5 we can also see that the interaction between tau4 and theta5 cannot be simply captured by the multiplication term $X_{\text{tau4}} \cdot X_{\text{theta5}}$.

6 Concluding remarks

We study the variable and interaction selection for logistic regression with second-order terms, which covers QDA as a special case. A somewhat surprising observation is that the two-stage algorithm SODA proposed in this article, which is based on the classic stepwise regression idea with a twist for efficiently searching for interaction terms, outperformed all known advanced approaches such as those Lasso-based methods in terms of variable selection accuracy, prediction accuracy, and robustness in a variety of settings when the joint distribution of the predictors do not "behave nicely." Additional empirical studies in Supplementary Materials also showed that when the "incoherence" condition is violated in standard logistic regression models with only linear main effects, SODA still performed robustly and outperformed Lasso-logistic significantly. In contrast to methods of Murphy et al. (2010); Zhang and Wang (2011); and Maugis et al. (2011), the consistency of SODA does not require the joint normality assumption of all candidate predictors. Compared to IIS of Fan et al. (2015), SODA's forward variable addition does not need the normal assumption and does not need to estimate large precision matrices.

It is worth noting that even for logistic regression models with only main effects, we consistently observed that SODA performed better than or similarly to Lasso-logistic in terms of both the $\text{EBIC}_{0.5}$ score and the CV error under various settings, especially when the predictors are highly correlated or the joint distribution of the predictors is long-tailed (see Supplementary Materials). This indicates that EBIC is a good criterion to follow and our stepwise approach is a better optimizer of EBIC than Lasso. Indeed, when one moves away from the L_1 regularization realm but adopts the L_0 regularization framework (such as AIC, BIC, EBIC), Lasso can no longer guarantee to find the optimal solution.

LDA and QDA complement each other in terms of the bias-variance trade-off. Given

finite observations, LDA is simpler and more robust when the response Y can be explained well by the linear effects of $\mathbf{X}$. QDA has the ability to exploit interaction effects, which may contribute dramatically to the classification accuracy, but also has many more parameters to estimate and is more vulnerable to including noise predictors. SODA is designed to be adaptive in the sense that it automatically chooses between LDA and QDA models and takes advantage of both sides. Instead of selecting predictors, SODA selects individual main and interaction terms, which enables SODA to simultaneously utilize interaction terms and avoid including a large number of unnecessary terms.

An interesting and also somewhat surprising twist of SODA is its extension S-SODA for dealing with the variable selection problem for nonparametric regression models with continuous responses. Our simulation results demonstrate that the simple idea of slicing (aka *discretizing*) the response variable can bring a lot to the table, especially when coupled with stepwise variable selection tools such as N-SIRI (Jiang and Liu, 2014) and S-SODA. Compared with existing SIR-based methods, SODA enjoys a much improved robustness. When the underlying true model is linear, S-SODA performed competitively with LASSO and outperformed other linear or near-linear methods, such as hierNet and DC-SCAD, when the joint distribution of the covariates is long-tailed.

A main limitation of SODA is that the stepwise detectable condition might not hold when main effects are weak or nonexistent but some interaction effects are strong. Other than letting all interaction terms subject to selection, which is computationally inhibitive for moderately large p, we may overcome this difficulty by examining conditional distributions $[X_j \mid Y]$, $j = 1, \ldots, p$, in some fashion. In empirical studies we found that SODA worked well for both simulated and real-data examples, suggesting that this limitation may not be a serious issue in most real applications. Indeed, even for QDA models it is quite unusual and nearly pathological to construct mean vectors and covariance matrices that

result in a discriminant function with no main effects but only interaction terms.

The Implementation of SODA and S-SODA procedures is available in the R package `sodavis` on CRAN (`http://cran.us.r-project.org`).

SUPPLEMENTARY MATERIALS

Appendix: Additional empirical studies and detailed proofs of Theorems 1, 2 and 4.

References

Anderson, T. W. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Beer, D. G., S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, et al. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature medicine 8*(8), 816–824.

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

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152. ACM.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Broman, K. W. and T. P. Speed (2002). A model selection approach for the identification

of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(4), 641–656.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association 106*(496), 1566–1577.

Chen, J. and Z. Chen (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika 95*(3), 759–771.

Chen, J. and Z. Chen (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica 22*(2), 555.

Chowdhary, R., J. Zhang, and J. S. Liu (2009). Bayesian inference of protein–protein interactions from biological literature. *Bioinformatics 25*(12), 1536–1542.

Clemmensen, L., T. Hastie, D. Witten, and B. Ersbøll (2011). Sparse discriminant analysis. *Technometrics 53*(4), 406–413.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, 1–26.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association 104*(487), 1015–1028.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, Volume 1. Cambridge University Press.

Fan, J. and Y. Fan (2008). High dimensional classification using features annealed independence rules. *Annals of statistics 36*(6), 2605.

Fan, Y., J. Jin, and Z. Yao (2013). Optimal classification in sparse gaussian graphic model. *The Annals of Statistics 41*(5), 2537–2571.

Fan, Y., Y. Kong, D. Li, and Z. Zheng (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics 43*(3), 1243–1272.

Foygel, R. and M. Drton (2011). Bayesian model choice and information criteria in sparse generalized linear models. *arXiv preprint arXiv:1112.5635*.

Guo, Y., T. Hastie, and R. Tibshirani (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics 8*(1), 86–100.

Han, F., T. Zhao, and H. Liu (2013). Coda: High dimensional copula discriminant analysis. *The Journal of Machine Learning Research 14*(1), 629–671.

Jia, J. and B. Yu (2010). On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica 20*(2), 595–611.

Jiang, B. and J. S. Liu (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics 42*(5), 1751–1786.

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Springer.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association 86*(414), 316–327.

Li, L. (2007). Sparse sufficient dimension reduction. *Biometrika 94*(3), 603–613.

Li, L., R. Dennis Cook, and C. J. Nachtsheim (2005). Model-free variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 285–299.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*(499), 1129–1139.

Lin, Q., Z. Zhao, and J. S. Liu (2015). On consistency and sparsity for sliced inverse regression in high dimensions. *arXiv preprint arXiv:1507.03895*.

Lin, Q., Z. Zhao, and J. S. Liu (2016). Sparse sliced inverse regression for high dimensional data. *arXiv preprint arXiv:1611.06655*.

Mai, Q., H. Zou, and M. Yuan (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42.

Maugis, C., G. Celeux, and M.-L. Martin-Magniette (2011). Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis 102*(10), 1374–1387.

Murphy, T. B., N. Dean, and A. E. Raftery (2010). Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics 4*(1), 396.

Phillips, P. J. (1998). *Support vector machines applied to face recognition*, Volume 285. Citeseer.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010). High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics 38*(3), 1287–1319.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics 6*(2), 461–464.

ACCEPTED MANUSCRIPT

Shao, J., Y. Wang, X. Deng, and S. Wang (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics 39*(2), 1241–1265.

Simon, N. and R. Tibshirani (2012). A permutation approach to testing interactions in many dimensions. *arXiv preprint arXiv:1206.6519*.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*(2), 203–209.

Szretter, M. E. and V. J. Yohai (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference 139*(10), 3570–3578.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences 99*(10), 6567–6572.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications*, 210–268.

Waldmann, P., G. Mészáros, B. Gredler, C. Fuerst, and J. Sölkner (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics 4*.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*(488), 1512–1524.

Wasserman, L. and K. Roeder (2009). High dimensional variable selection. *Annals of statistics 37*(5A), 2178.

Witten, D. M. and R. Tibshirani (2011). Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73*(5), 753–772.

Yu, Y. and Y. Feng (2014). Modified cross-validation for penalized high-dimensional linear regression models. *Journal of Computational and Graphical Statistics 23*(4), 1009–1027.

Zhang, Q. and H. Wang (2011). On bic's selection consistency for discriminant analysis. *Statistica Sinica 21*(2), 731.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research 7*, 2541–2563.

Zhong, W., T. Zhang, Y. Zhu, and J. S. Liu (2012). Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(5), 849–870.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

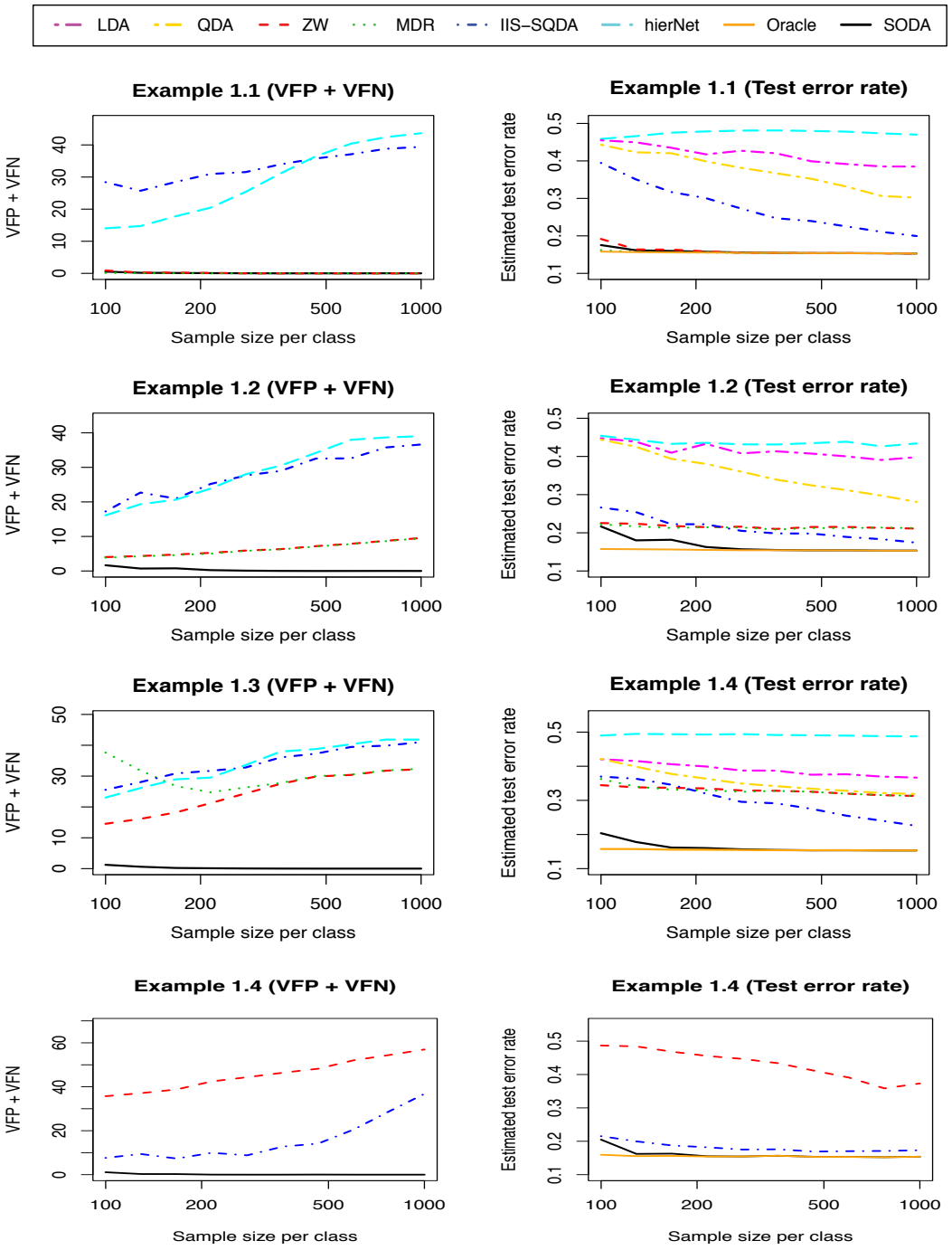

Figure 2: Results for Example 1.1∼1.4. VFP: average number of variable selection false positives. VFN: average number of variable selection false negatives. LDA and QDA used all the variables without any selection, so they do not appear in the left panel and their TEs were high. MDR and hierNet all broke down for Example 1.4. LDA and QDA also did not work due to large p.

Example	n	SODA				IIS-SQDA				hierNet			
		MFN	**MFP**	**IFN**	**IFP**	**MFN**	**MFP**	**IFN**	**IFP**	**MFN**	**MFP**	**IFN**	**IFP**
1.1	100	0.05	0.16	1.01	0.30	0.27	2.39	0.90	48.5	0	12.6	1.58	2.26
	215	0	0.01	0.04	0.02	0.08	2.90	0.25	63.2	0	19.2	1.10	14.0
	1000	0	0	0	0	0	6.39	0	112	0	44.6	0	46.2
1.2	100	0.26	0.58	1.74	0.28	0.26	12.9	0.40	7.42	0	14.9	1.44	7.24
	215	0	0.13	0.27	0.03	0	19.7	0.02	11.1	0	22.9	0.65	15.3
	1000	0	0	0	0	0	28.5	0	24.3	0	39.1	0	46.9
1.3	100	0.12	0.13	1.50	0.70	0.09	5.59	0.13	44.9	0.04	21.4	2.20	19.3
	215	0.02	0.03	0.17	0.07	0	8.96	0	61.1	0	29.5	0.93	25.7
	1000	0	0	0	0	0	14.71	0	99.8	0	42.8	0	46.0
1.4	100	0.20	0.22	1.58	0.30	0.68	1.58	0.42	6.08				
	215	0	0	0.14	0	0.20	1.74	0.10	8.84				
	1000	0	0	0	0	0	3.68	0	40.7				

Table 2: Variable Selection Results for Examples 1.1 $\sim$ 1.4. MFP / MFN: Average number of main effect false positives and negatives. IFP / IFN: Average number of interaction effect false positives and negatives. The number of observations for each class is denoted by n.

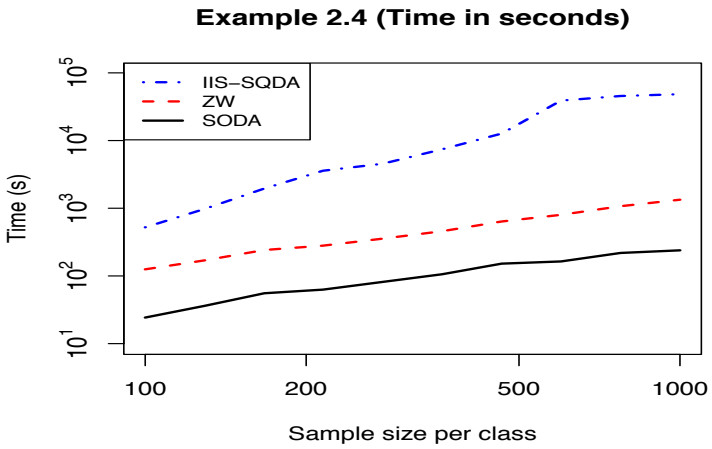

Figure 3: Mean running time in seconds for ZW, IIS-SQDA, and SODA for Example 1.4; and hierNet did not finish the job within 24 hours.

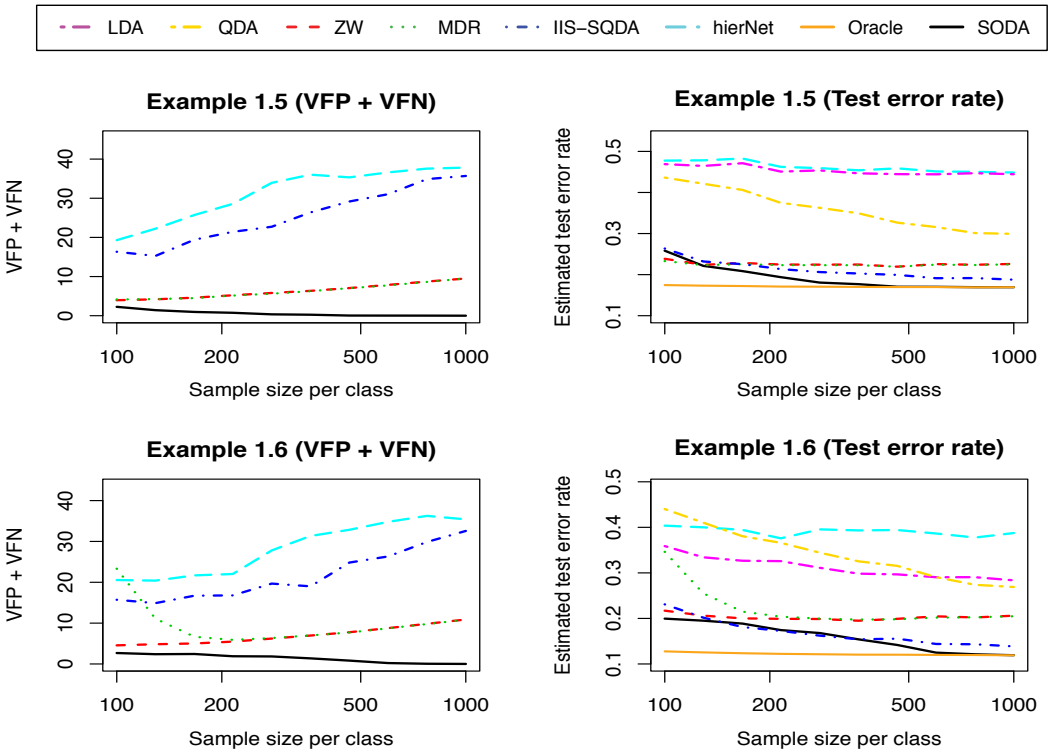

Figure 4: Results for Example 1.5 $\sim$ 1.6. VFP: average number of variable selection false positives. VFN: average number of variable selection false negatives.

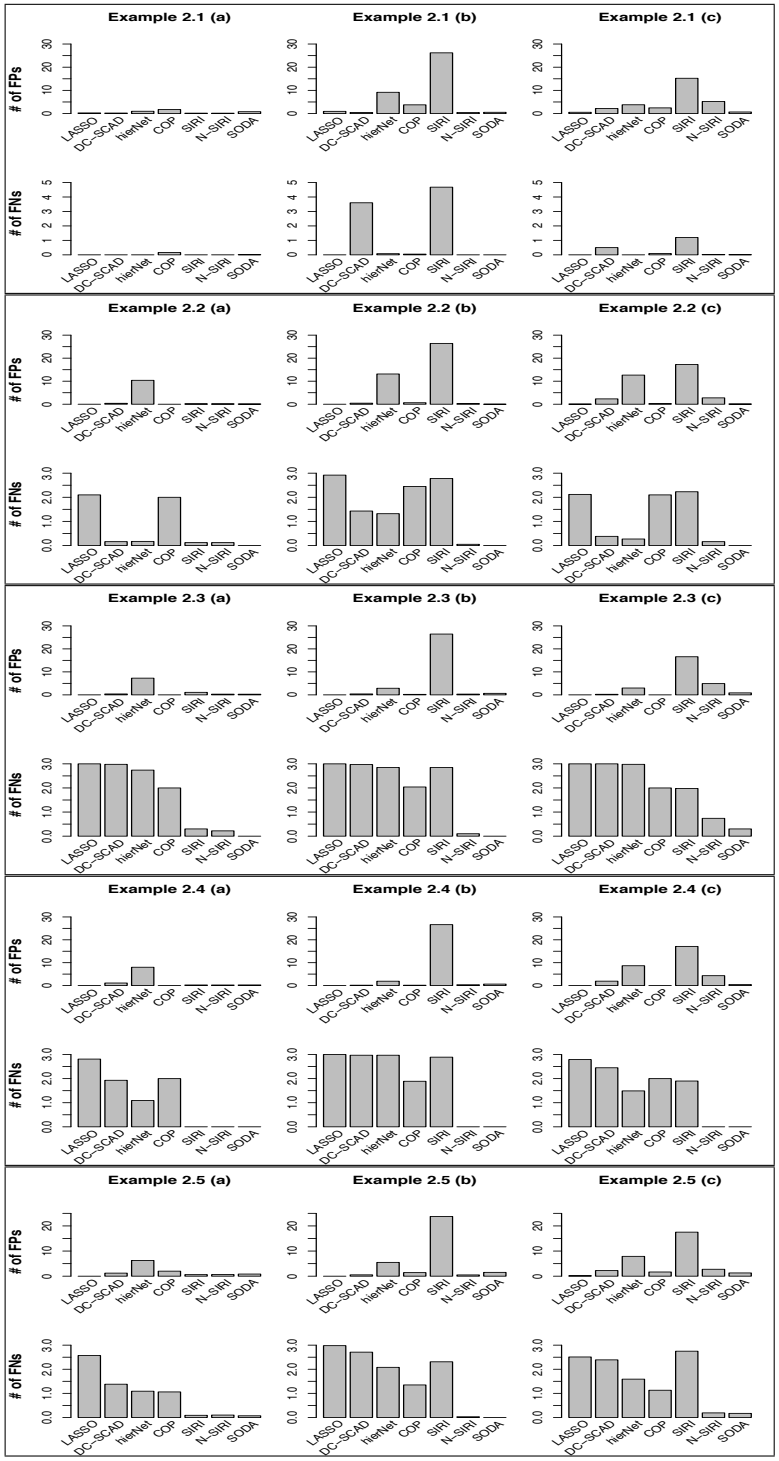

Figure 5: Simulation study results for Examples 2.1∼2.5.

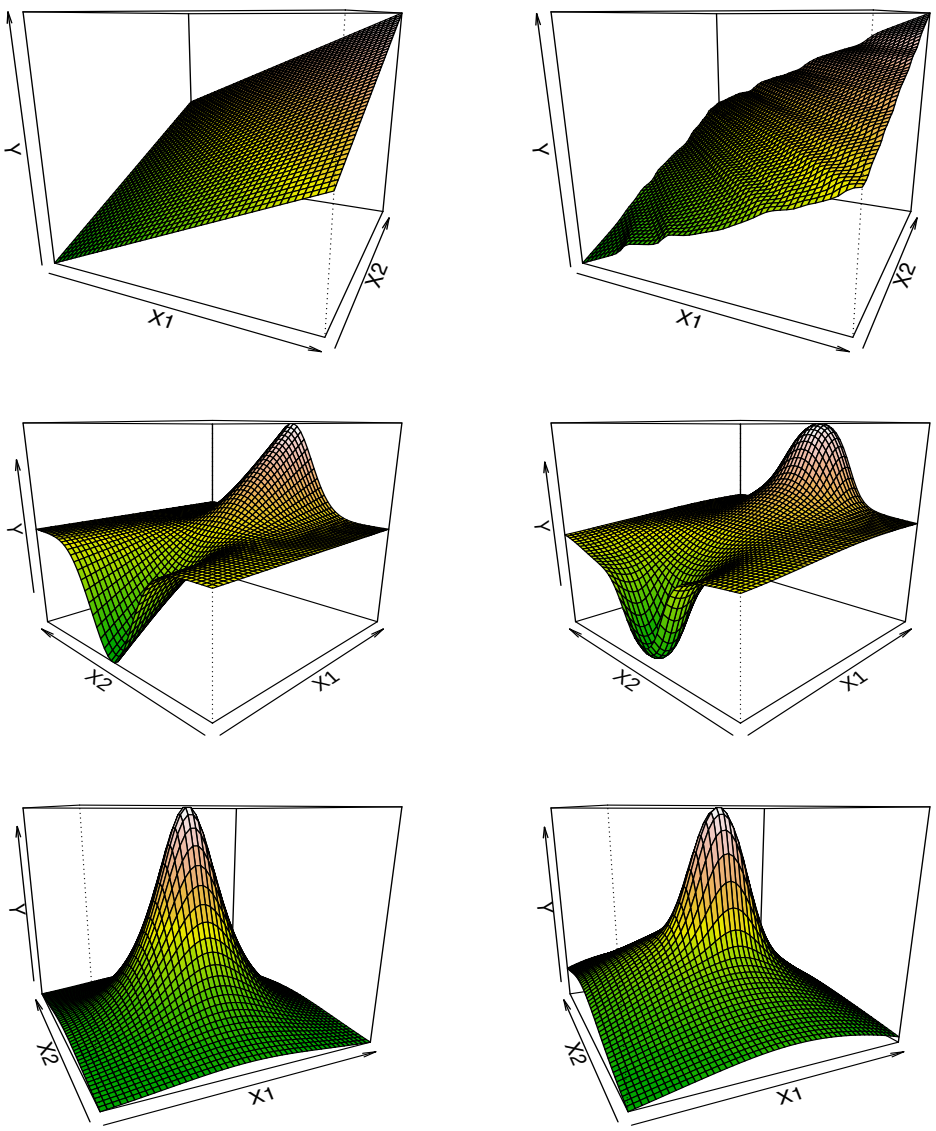

Figure 6: Results for the simulation Examples 3.1-3.3. Left panel: theoretical surface $\mathbb{E}[Y \mid \mathbf{X}]$; Right panel: surface $\hat{\mathbb{E}}[Y \mid \mathbf{X}]$ predicted by S-SODA.

Shrunken centroid		**Empirical Bayes**		**MDR**			**Lasso-Logistic**			**SODA**		
#P	CVE	#P	CVE	ΔBIC	#P	CVE	$EBIC_{0.5}$	#P	CVE	$EBIC_{0.5}$	#M/#I	CVE
0	0.28	5	0.41	-353	1	0.27	**112.2**	**1**	**0.29**	111.2	1 / 0	0.33
5	0.28	20	0.43	-177	2	0.26	113.9	2	0.25	104.1	2 / 0	0.25
11	0.29	40	0.39	-178	3	0.25	122.2	3	0.25	98.2	3 / 0	0.21
21	0.28	60	0.41	-165	4	0.24	121.0	4	0.20	89.6	4 / 0	0.17
55	0.35	80	0.40	-156	5	0.25	131.5	5	0.22	79.2	5 / 0	0.12
109	0.35	100	0.39	-134	8	0.24	144.5	6	0.23	94.4	5 / 1	0.12
260	0.37	120	0.40	-132	11	0.29	150.5	7	0.25	⋮	⋮	⋮
567	0.38	140	0.40	-131	14	0.28	158.4	8	0.22	**69.8**	**2 / 2**	**0.11**
1,173	0.40	160	0.42	-143	17	0.30	171.1	9	0.23			
2,532	0.38	180	0.38	-146	20	0.27	177.6	10	0.23			
5,217	0.38	200	0.40	**-151**	**25**	**0.28**	188.6	11	0.23			

Table 3: Analysis results of the Michigan lung cancer dataset by five methods. For Lasso-Logistic, MDR and SODA, the selected set with the lowest BIC score is highlighted in bold font. ΔBIC: For MDR method, the difference of BIC_G between two adjacent steps. CVE: prediction error rate estimated by 10-fold CV. #P: number of selected predictors. #M / #I: number of selected main effect and interaction terms by SODA.

MDR			**Lasso-Logistic**			**Lasso-Logistic-2**			**IIS-SQDA**		**SODA**		
ΔBIC	#P	CVE	$EBIC_{0.5}$	#P	CVE	$EBIC_{0.5}$	#M / #I	CVE	#M / #I	CVE	$EBIC_{0.5}$	#M / #I	CVE
-326	1	0.20	343.5	2	0.19	279.0	1 / 2	0.13	**10 / 96**	**0.16**	371.2	1 / 0	0.21
-221	3	0.26	329.9	4	0.16	253.8	2 / 3	0.10			343.5	2 / 0	0.19
-338	5	0.25	**302.9**	**6**	**0.14**	252.7	2 / 4	0.09			319.6	3 / 0	0.19
-298	7	0.24	313.5	8	0.16	**248.7**	**2 / 5**	**0.08**			298.8	4 / 0	0.15
-242	9	0.25	312.5	10	0.15	254.4	2 / 6	0.08			296.1	5 / 0	0.14
-200	11	0.24	321.7	12	0.15	258.6	2 / 7	0.08			232.2	5 / 1	0.08
-278	15	0.29	345.3	15	0.15	267.3	2 / 8	0.08			224.1	5 / 3	0.07
-361	20	0.28	363.8	18	0.15	286.3	2 / 10	0.08			⋮	⋮	⋮
-434	25	0.30	383.1	22	0.15	290.3	2 / 11	0.08			**204.2**	**4 / 4**	**0.06**
-130	**32**	**0.28**	445.4	30	0.16	312.0	2 / 13	0.08					

Table 4: The summary of results on the Ionosphere dataset by the five methods. ΔBIC: For MDR method, the difference of BIC_G between two adjacent steps. CVE: prediction error rate estimated by 10-fold cross-validation. #P: number of selected predictors. #M / #I: number of selected main effect and interaction terms by SODA.

Linear		**Linear-2**		**S-SODA**	
# Predictors	Out-r	# M / #I	Out-r	# Predictors	Out-r
1	0.477	1 / 0	0.477	1	0.469
2	0.477	1 / 1	0.476	2	0.707
3	0.476	1 / 2	0.474		
4	0.476	1 / 3	0.473		
5	0.476	1 / 4	0.473		
10	0.474	1 / 10	0.469		
20	0.472	1 / 20	0.464		
30	0.472	1 / 30	0.459		

Table 5: Analysis results of the Pumadyn dataset by the three methods. #P: the number of selected predictors. #M / #I: the number of selected main effect and interaction terms by Lasso on linear model with interaction terms. Out-r: the out-sample correlation r.

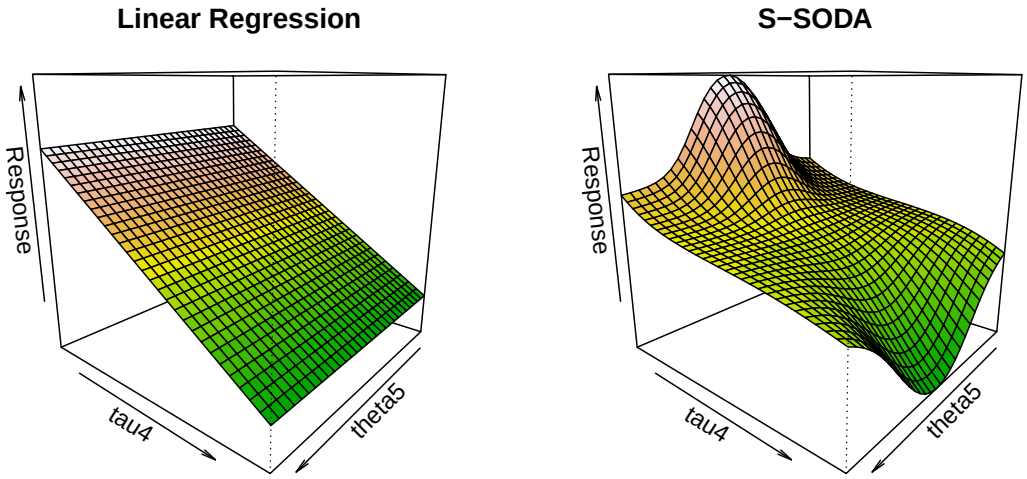

Figure 7: The predicted $\hat{\mathbb{E}}\left[Y \mid \mathbf{X}_{(\text{tau4,theta5})}\right]$ surface from linear model (left) and S-SODA (right).