Journal of Applied Statistics

Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

A link-free sparse group variable selection method for single-index model

Bilin Zeng, Xuerong Meggie Wen & Lixing Zhu

To cite this article: Bilin Zeng, Xuerong Meggie Wen & Lixing Zhu (2016): A link-free sparse group variable selection method for single-index model, Journal of Applied Statistics, DOI: 10.1080/02664763.2016.1254731

To link to this article: http://dx.doi.org/10.1080/02664763.2016.1254731

Published online: 14 Nov 2016.

JOURNAL OF APPLIED STATISTICS, 2016
http://dx.doi.org/10.1080/02664763.2016.1254731

Taylor & Francis
Taylor & Francis Group

A link-free sparse group variable selection method for single-index model

Bilin Zeng[a], Xuerong Meggie Wen[b] and Lixing Zhu[c]

[a]Department of Mathematics, California State University, Bakersfield, CA, USA; [b]Department of Mathematics and Statistics, Missouri University of Science and Technology, Rolla, MO, USA; [c]Department of Mathematics, Hong Kong Baptist University, Hong Kong, People's Republic of China

ABSTRACT

For regression problems with grouped covariates, we adapt the idea of sparse group lasso (SGL) [10] to the framework of the sufficient dimension reduction. Assuming that the regression falls into a single-index structure, we propose a method called the *sparse group sufficient dimension reduction* to conduct group and within-group variable selections simultaneously without assuming a specific link function. Simulation studies show that our method is comparable to the SGL under the regular linear model setting and outperforms SGL with higher true positive rates and substantially lower false positive rates when the regression function is nonlinear. One immediate application of our method is to the gene pathway data analysis where genes naturally fall into groups (pathways). An analysis of a glioblastoma microarray data is included for illustration of our method.

ARTICLE HISTORY
Received 23 November 2015
Accepted 14 October 2016

KEYWORDS
Single-index model; sparse group lasso; gene pathway analysis; sufficient dimension reduction; variable selection

CLASSIFICATION CODES
62G08; 62P10

1. Single-index model and sufficient dimension reduction

For a typical regression problem with a univariate random response Y and a p-dimensional random vector $\mathbf{X}$, sufficient dimension reduction (SDR: [5,6,19]) aims to reduce the dimension of $\mathbf{X}$ without loss of information on the regression and without requiring a prespecified parametric model. The basic idea of SDR is to replace the predictors $\mathbf{X} \in \mathbb{R}^p$ with a lower dimensional projection $P_{\mathcal{S}}\mathbf{X}$ onto a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ without the loss of information on the original regression of $Y \mid \mathbf{X}$, i.e. $Y \perp\!\!\!\perp \mathbf{X} \mid P_{\mathcal{S}}\mathbf{X}$, where $\perp\!\!\!\perp$ indicates independence and $P_{(\cdot)}$ stands for a projection operator with respect to the standard inner product. Such an $\mathcal{S}$ is defined as a dimension reduction subspace, and the smallest one is called the *central subspace* $\mathcal{S}_{Y\mid\mathbf{X}}$ [5], which exists under very mild conditions [5,43]. We assume the existence of $\mathcal{S}_{Y\mid\mathbf{X}}$ throughout this article. The dimension of central subspace $\mathcal{S}_{Y\mid\mathbf{X}}$, denoted by d, is called the structural dimension of the regression.

When $d = 1$, it is called the Single-Index Model (SIM) [40]

$$Y = g(\boldsymbol{\beta}^T\mathbf{X}, \epsilon), \tag{1}$$

where $g(\cdot)$ is an unknown link function, $\boldsymbol{\beta}$ is a p-dimensional vector, and the random error ϵ is independent with $\mathbf{X}$. Model (1) [12] is a very general semiparametric model that

CONTACT Bilin Zeng ✉ bzeng@csub.edu

includes a commonly used multiple linear regression model as a special case. One is usually concerned with estimation of $\boldsymbol{\beta}$ and the link function $g(\cdot)$ [9]. We, however, focus on developing a link-free variable selection method assuming Model (1) under the framework of SDR.

Most of the existing variable selection methods are model-based [38]. Such methods might generate biased results if the underlying modeling assumption is violated, which is typically the case for complex or unknown models. For a variable selection method that does not require the full knowledge of the underlying true model, it is called *model-free* or *link-free* variable selection. As pointed out by Bondell and Li [3], the general framework of SDR is very useful for variable selection since no pre-specified underlying models between Y and $\mathbf{X}$ are required. Model-free variable selection can be achieved through the framework of SDR [21]. Instead, usually a so-called linearity condition [11,42] on the marginal distribution of $\mathbf{X}$ is assumed. This is a mild condition and holds approximately true when p goes to infinity. Ni *et al.* [32], Li and Nachtsheim [23] and Li and Yin [24] proposed model-free variable selections by reformulating SDR as a penalized regression problem. Li [20] proposed a unified approach by combining SDR and shrinkage estimation to produce sparse estimators of the central subspace. Wang *et al.* [40] proposed a distribution-weighted lasso method for the SIM. However, none of those model-free variable selections take the prior group (predictor network) information into account. Such situations do arise in the gene pathway analysis where genes naturally fall into groups (pathways/gene networks; see Section 4 for more discussions).

In this paper, we propose a link-free (model-free) variable selection method called the sparse group sufficient dimension reduction (sgSDR), which conducts both group and within-group variable selections simultaneously under the framework of the SIM. We then apply our method to a survival analysis for glioblastoma patients [14] using gene-expression profiles with about 1500 genes and 33 pathways.

The remainder of this article is organized as follows. Section 2 describes our statistical approach. We first review the sparse group lasso (SGL) [10], then show how it can be extended within the context of SDR. The SLEP package [25] is adopted for the implementation of our method. Five-fold cross-validation is used to select the related tuning parameters. Section 3 reports simulation studies comparing the finite-sample performances of our method with the SGL. A real data example on glioblastoma study [14] is discussed in Section 4. Conclusions and a brief discussion on future research directions are given in Section 5.

2. Sparse group sufficient dimension reduction

The lasso-penalized linear regression [38] is applied to high-dimensional regression problems with tens to hundreds of thousands of predictors. It finds a solution with few non-zero entries by minimizing

$$\tfrac{1}{2}\|\mathbf{y}-\mathcal{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\mathbf{y}=(y_1,\ldots,y_n)^{\mathrm{T}}$ is the observed centered response vector, $\mathcal{X}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)^{\mathrm{T}}$ is the centered design matrix with $\mathbf{x}_i=(x_1^i,\ldots,x_p^i)^{\mathrm{T}}$ being the predictor values for the ith observed subject, $\boldsymbol{\beta}\in\mathbb{R}^p$ the vector of regression coefficients, $\|\mathbf{z}\|_2=(\sum_j z_j^2)^{1/2}$ the

Euclidean (l_2) norm, and $\|\mathbf{z}\|_1 = \sum_j |z_j|$ the l_1 norm. The first term in Equation (2) represents the loss function minimized in the ordinary least-squares, and the second term is the lasso penalty function where the multiplier $\lambda > 0$ is the penalty constant. Large value of λ will set some components β_j exactly to 0. The lasso has become a popular model selection and shrinkage estimation method since it is capable of producing sparse models and is computationally feasible. However, due to the fact that the lasso selects at most n variables before it saturates [46], the lasso fails when the number of significant predictors is greater than the sample size. In addition, the lasso has poor performance when predictors are highly correlated. In such case, lasso tends to randomly select only one variable from each correlated group.

In some applications, it is natural to group correlated predictors [44]. This raises the question of how to penalize a group of parameters. Combining the l_1 norm that is used in lasso and l_2 penalty that is used in the ridge regression [13] to the ordinary least-squares, the elastic [46] generates almost equal regression coefficients for a group of highly correlated variables. Therefore, it is useful for performing group selections of correlated components when the group information is unknown in advance. In some real applications, there exists prior group knowledge that has been obtained through research and studies in the area of expertise. In the case that the group information is pre-assigned, the group lasso proposed by Yuan and Lin [44] allows the group sparsity since its l_2 group penalty takes the prior group information into account. Its solution is obtained by minimizing the following penalized least-squares regression:

$$\frac{1}{2}\left\|\mathbf{y} - \sum_{g=1}^{G} \mathcal{X}^{(g)}\boldsymbol{\beta}^{(g)}\right\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g}\|\boldsymbol{\beta}^{(g)}\|_2, \tag{3}$$

where $\mathcal{X}^{(g)}$ is the submatrix of $\mathcal{X}$ with columns corresponding to the predictors in the gth group, and $\boldsymbol{\beta}^{(g)}$ is the coefficient vector of that group with p_g as its length. The rescaling factor p_g makes the penalty level proportional to the group size, which ensures that small groups are not overwhelmed by large groups in group selections. The group lasso penalty has been investigated in multiple studies [2,15,30]. The sparsity of the solution is determined by the tuning parameter λ. When the group size $p_g = 1$, group lasso is reduced to the regular lasso. While the group lasso can identify important groups, it is not capable of selecting important predictors within each group, which will be an issue when p_g is large.

Friedman *et al.* [10] proposed the SGL which could achieve sparsity of both groups and within each group by minimizing the following penalized least-squares regression:

$$\frac{1}{2}\left\|\mathbf{y} - \sum_{g=1}^{G} \mathcal{X}^{(g)}\boldsymbol{\beta}^{(g)}\right\|_2^2 + \lambda_1 \sum_{g=1}^{G} \sqrt{p_g}\|\boldsymbol{\beta}^{(g)}\|_2 + \lambda_2\|\boldsymbol{\beta}\|_1. \tag{4}$$

It is reduced to the group lasso when $\lambda_2 = 0$ and the lasso when $\lambda_1 = 0$. SGL is capable of selecting important groups and important predictors within the selected groups simultaneously. Unlike the elastic net, SGL encourages the sparsity at the group levels as its l_2 penalty is not differentiable at zero. For the sparse selection within-group levels, SGL generates an elastic net type solution [36]. SGL might lead to better predictions since it takes the prior cluster structure into consideration; and also, its merit of performing within-group variable selections can produce more parsimonious models and hence lead to more interpretable results. However, all the above lasso-based methods assume a linear relationship between the response and the predictors. They are not applicable to the scenarios when the linearity modeling assumption is violated. We propose a sgSDR method to overcome this limitation.

Li *et al.* [22] proposed the groupwise dimension reduction which incorporates the prior grouping information into the estimation of the central mean subspace. Simulation studies and real data analyses showed that the groupwise dimension reduction approach can substantially increase the estimation accuracy and enhance the estimates interpretability. However, their method is only limited to the dimension reduction of the conditional mean ($\mathrm{E}(Y \mid \mathbf{X})$), and furthermore, it is not capable of conducting variable selections for sparse models. The sgSDR method we propose in this article can conduct variable selection in the general dimension reduction context, including but not limited to the conditional mean, while incorporating the prior group information. Moreover, this method is applicable when the sample size n is far less than the number of predictors p (i.e. $n \ll p$ setting).

We focus on the following general SIM:

$$Y = g(\boldsymbol{\beta}^T\mathbf{X}, \epsilon). \tag{5}$$

Without the loss of generality, we assume that $\mathbf{X}$ is centered with $\mathrm{E}(\mathbf{X}) = 0$ and also suppose that $\mathbf{X}$ can be split into G groups, $\mathbf{X}^{\mathrm{T}} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(G)})$, where $\mathbf{X}^{(g)}$ is a p_g-dimensional row vector, for $g = 1, \ldots, G$, and $\sum_{g=1}^{G} p_g = p$. Following Wang *et al.* [40], we consider the following minimization problem:

$$\frac{1}{2}\left\|\mathbf{F}_n(\mathbf{y}) - \sum_{g=1}^{G}\mathcal{X}^{(g)}\boldsymbol{\beta}^{(g)}\right\|_2^2 + \lambda_1\sum_{g=1}^{G}\sqrt{p_g}\|\boldsymbol{\beta}^{(g)}\|_2 + \lambda_2\|\boldsymbol{\beta}\|_1, \tag{6}$$

where $\mathbf{F}_n(\mathbf{y}) = (F_n(y_1), \ldots, F_n(y_n))^{\mathrm{T}}$ and $\mathcal{X}$ are all centered, $\boldsymbol{\beta}^{(g)}$ is the coefficient vector of the gth group with p_g as its group size, and $F_n(y) = \sum_{i=1}^{n} I(Y_i \le y)/n$ is the empirical distribution function. The solution $\boldsymbol{\beta} = (\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(g)})^{\mathrm{T}}$ of Equation (6) forms the central subspace, and here we call it the sgSDR estimator. For a general SIM, the non-zero entries in the central subspace above represent the estimated coefficients of significant predictor variables in the variable selection. Moreover, Proposition 2.1 shows that the sgSDR estimator is consistent with the non-zero coefficients in the true model. Proposition 2.1 states the identifiability of sgSDR. The proof is similar to that of Proposition 2.1 of Wang *et al.* [40] and is hence omitted.

Proposition 2.1: *Under the linearity condition, and assume that $\boldsymbol{\Sigma}_s$, the marginal covariance matrix of all the significant predictors (denoted by $\mathbf{X}_s$ here for easy of exposition) is invertible, then*

$$\boldsymbol{\Sigma}_s^{-1}\mathrm{Cov}\{\mathbf{X}_s, F(Y)\} = c\boldsymbol{\beta}_s,$$

where $\boldsymbol{\beta}_s$ consists all non-zero coefficients of $\boldsymbol{\beta}$ from Equation (5), $c \in \mathbb{R}^1$ is a constant, and $F(Y)$ is the cumulative distribution function of Y.

The linearity condition is widely assumed in the dimension reduction context [5, 11,19,42]. It holds trivially when the predictor distribution is elliptically symmetric. Our condition is much weaker than the existing one as we only assume the linearity condition on $\mathbf{X}_s$ rather than the original p-dimensional vector $\mathbf{X}$.

The above proposition also holds for any transformation of response Y, $h(Y)$, which implies that the empirical cumulative distribution function used in our method can be replaced by any other transformation of Y. In this paper, the empirical cumulative distribution function $F(Y)$ is used for its computational simplicity. Proposition 2.1 shows that the SDR approach enables us to obtain a consistent estimator for $\boldsymbol{\beta}_s$ without requiring specific specifications of the link function $g(\cdot)$ and without employing any nonparametric smoothing methods.

Let $\hat{\boldsymbol{\beta}}_s$ be the sample estimate of $\boldsymbol{\Sigma}_s^{-1}\text{Cov}\{\mathbf{X}_s, F(Y)\}$, the following two theorems state its asymptotic properties. The results are similar to that of Wang *et al.* [40] as they also hold true for design matrices with grouped structures.

Theorem 2.1: *Assume that the following conditions are satisfied:*

(a) $L_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}_s) \leq L_2$ *for some* $L_1, L_2 > 0$
(b) $\max_{1\leq i\leq p} E(\mathbf{X}_i^4) < L_3 < \infty$ *for some* L_3
(c) $p = o(\sqrt{n})$

where $\lambda_{\min}(\cdot)$ *and* $\lambda_{\max}(\cdot)$ *are the smallest and largest eigenvalues of a symmetric matrix, respectively. Then we have*

$$\|\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s\|_2 = O_p\left(\frac{p}{n^{1/2}}\right).$$

Theorem 2.2: *Assuming condition* (a) *from Theorem 2.1, and also the following conditions* [40]:

- $\max_{1\leq i\leq p} E(\mathbf{X}_i^8) < L_3 < \infty$ *for some* L_3
- $p = o(n^{1/4})$

then as $n \to \infty$, *for any vector* $\boldsymbol{v} \in \mathbb{R}^p$ *such that* $\|\boldsymbol{v}\|_2 \leq 1$ *and* $\boldsymbol{v}^T\boldsymbol{\Lambda}\boldsymbol{v} \to \boldsymbol{G} > 0$ *as* $n \to \infty$, *where* $\boldsymbol{\Lambda} = \text{Cov}\{[F(Y) - \sum_{g=1}^{G}\mathbf{X}^{(g)}\boldsymbol{\beta}^{(g)}]\mathbf{X}\}$, *we have*

$$\sqrt{n}\boldsymbol{v}^T(\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s) \to N(0, \boldsymbol{G}).$$

We minimize Equation (6) to obtain a sparse estimator. The sparsity of the solution is determined by the tuning parameters λ_1 and λ_2. Specifically, the sparsity of group selection is controlled by λ_1 in Equation (6) while the number of variables selected within each group depends on the value of λ_2. The larger value of λ_1 will shrink more group parameters into zero, and it will result in fewer groups being selected. Similarly, smaller value of λ_2 implies less shrinkage on individual parameters which will produce more selected variables within groups and vice versa. To select the two tuning parameters, λ_1 and λ_2, we employ the commonly used five-fold cross-validation in this paper. Simon *et al.* [36] suggested using $\lambda_1 = 19\lambda_2$. However, there is no theoretical justification to support this special λ_1 to λ_2 ratio. In fact, the λ_1 to λ_2 ratio might need to be adjusted when the scenario varies. Therefore, we run all the possible combinations of λ_1 and λ_2 on a two-dimensional

(λ_1, λ_2) grid which is composed by a wide range of λ_1 and λ_2 values. The SLEP package [25] is adopted to implement our method. The selection of the two tuning parameters λ_1 and λ_2 via five-fold cross-validation can be described generically as follows:

Step 1: Randomly partitioning the sample data into five equal sized subsamples. Four subsamples are used as training data, whereas the remaining subsample is used for testing.

Step 2: Choose a grid of values for λ_1 and λ_2.

Step 3: For each (λ_1, λ_2) combination, calculate and store its cross-validation error

$$\mathrm{CVE}(\lambda_1, \lambda_2) = \frac{1}{n}\|F(\mathbf{Y}_{\text{test}}) - \mathcal{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{training}}\|_2^2,$$

where $\hat{\boldsymbol{\beta}}_{\text{training}}$ is the estimated regression coefficient generated by sgSDR from the training set. $\mathbf{Y}_{\text{test}}$ and $\mathcal{X}_{\text{test}}$ denote the test sample response vector and the design matrix, respectively.

Step 4: Repeat Step 1–Step 3 until each of the subsamples are used exactly once as the test set. Denoted by $\mathrm{SCVE}(\lambda_1, \lambda_2)$, the sum of CVE for each specific (λ_1, λ_2) combination is calculated by adding up the $\mathrm{CVE}(\lambda_1, \lambda_2)$ from each loop.

Step 5: The optimal (λ_1, λ_2) solution is obtained by minimizing the $\mathrm{SCVE}(\lambda_1, \lambda_2)$.

3. Simulation studies

In this section, we compare the performance of our method with the SGL. The five-fold cross-validation is applied to both methods. A wide range of λ_1 and λ_2 values from 10^{-4} to 10^4 are used for the selection of tuning parameters for both methods [4]. We considered both linear and nonlinear models with Gaussian and non-Gaussian errors. We use the average true positive rate (TPR = the ratio of the number of correctly declared active variables to the number of truly active variables) and the average false positive rate (FPR = the ratio of the number of falsely declared active variables to the total number of truly inactive variables) as evaluation measurements to summarize variable selection results from 100 simulation runs.

Model I: For a fair comparison, we first consider a regular linear model as Simon *et al.* [36] discussed in their paper. The predictor $\mathbf{X}$ is generated from $N(0, I_p)$, ϵ is standard normal and independent of $\mathbf{X}$, the univariate response Y is constructed as

$$Y = \sum_{g=1}^{G} (\boldsymbol{\beta}^{(g)})^T \mathbf{X}^{(g)} + \sigma\epsilon, \tag{7}$$

where $G = 10$, σ is set to make the signal to noise ratio as 2. And the coefficients for the first l group are $\boldsymbol{\beta}^{(g)} = (1, 2, 3, 4, 5, 0, \ldots, 0)^{\mathrm{T}}$, for $g = 1, \ldots, l$, with l varying from 1 to 3; and all zeros for the rest of G—l groups. Following Simon *et al.* [36], we took $n = 60$, $p = 1500$. Table 1 provides the average true positive and FPRs. As shown in Table 1, the performances of sgSDR and SGL are comparable in the sense that the average TPRs and FPRs are very close to each other.

Model II: We now consider a variation of Model I. We take $p = 2000$, $G = 10$, and Y is still generated as in Equation (7). However, the predictors now are mildly correlated and $\boldsymbol{\beta}^{(g)} = (-2, 3, 0, \ldots, 0)^{\mathrm{T}}$, for $g = 1, \ldots, l$, with l varying from 1 to 3; and zeros otherwise.

Table 1. Linear model with Gaussian error.

	$l=1$		$l=2$		$l=3$	
	TPR	FPR	TPR	FPR	TPR	FPR
sgSDR	0.75	0.13	0.64	0.32	0.58	0.35
SGL	0.75	0.10	0.64	0.31	0.56	0.32

Table 2. Linear model with correlated predictors.

		$l=1$		$l=2$		$l=3$	
	Method	TPR	FPR	TPR	FPR	TPR	FPR
Gaussian error	sgSDR	1.00	0.04	1.00	0.05	1.00	0.05
	SGL	1.00	0.04	1.00	0.05	1.00	0.05
$t(5)$error	sgSDR	1.00	0.03	1.00	0.03	1.00	0.04
	SGL	1.00	0.03	1.00	0.03	1.00	0.04

Table 3. Nonlinear model III.

		$l=1$		$l=2$		$l=3$	
	Method	TPR	FPR	TPR	FPR	TPR	FPR
Gaussian error	sgSDR	1.00	0.03	1.00	0.04	1.00	0.04
	SGL	0.90	0.71	1.00	0.92	1.00	0.99
$t(5)$ error	sgSDR	1.00	0.03	1.00	0.04	1.00	0.04
	SGL	0.90	0.75	0.95	0.82	1.00	0.99

Specifically, within each group, $\mathbf{X}^{(g)} = (X_1^{(g)}, \ldots, X_{200}^{(g)})$ are all generated as independent standard normal random variables except $X_3^{(g)}$, which is generated to be correlated with X_1^g and X_2^g by

$$X_3^{(g)} = \tfrac{2}{3}X_1^{(g)} + \tfrac{2}{3}X_2^{(g)} + \tfrac{1}{3}e_g. \tag{8}$$

For the random errors, e_g, $N(0, 0.5^2)$ and t distribution with degrees of freedom of 5 are both considered.

The simulation results with $n=60$ from 100 simulation runs are shown in Table 2. We can see that our method (sgSDR) is comparable to SGL for linear models with correlated predictors.

Model III: We now compare the performances of sgSDR and SGL for nonlinear models. We first consider the following model:

$$Y = \exp\left(\sum_{g=1}^{G} \mathbf{X}^{(g)}\boldsymbol{\beta}^{(g)} + \epsilon\right). \tag{9}$$

The predictors $\mathbf{X}$ and the coefficients $\boldsymbol{\beta}$ are generated the same as those of Model II. As shown in Table 3, our method outperforms SGL with significantly lower FPR and slightly higher TPR. SGL fails when the regression function is nonlinear. The average FPR for SGL is above 75%, which implies that it mistakenly selected over 1500 inactive predictors as significant ones.

Model IV : In this example, the nonlinear model (9) is reconsidered with larger sample size, larger dimension p, and more groups, that is, $n=200$, $p=5000$ and $G=50$. The predictors are generated by $N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = (.5^{|i-j|})$, $i,j = 1, \ldots, p$. We consider

Table 4. Nonlinear model IV.

		$l=1$		$l=2$		$l=3$	
	Method	TPR	FPR	TPR	FPR	TPR	FPR
Gaussian error	sgSDR	0.80	0.03	0.73	0.04	0.68	0.12
	SGL	1.00	1.00	1.00	1.00	1.00	1.00
$t(5)$ error	sgSDR	1.00	0.03	0.99	0.03	0.86	0.04
	SGL	1.00	1.00	1.00	1.00	1.00	1.00

l (5,10,15) significant groups, with $\boldsymbol{\beta}^{(g)} = (3, 1.5, 2, \ldots, 0)^{\mathrm{T}}$, $g = 1, \ldots, l$. According to Table 4, we can see that our method (sgSDR) is more robust under the nonlinear model setting. In such a nonlinear model with complex structure of predictors, even though the same wide range of (λ_1, λ_2) grid ranging from 10^{-4} to 10^4 are provided for the selection of two tuning parameters for both methods, SGL still fails completely with 1.00 FPR which results in choosing all the predictors as significant ones.

4. A real data analysis

Genetic association studies aim to detect the associations between gene expressions and the occurrence or progression of disease phenotypes. Recent developments in microarray techniques make it possible to profile gene expressions on a whole genome scale, simultaneously measuring expressions of thousands or tens of thousands of genes. New challenges arise for the analysis of microarray data due to the large number of genes surveyed and often the relatively small sample sizes. A large amount of existing approaches (to list a few: [1,8,31,35]) has been developed to identify a small subset of genes or linear combinations of genes which are often referred to as super genes, that have influential effects on some certain diseases. Such studies can lead to better understanding of the genetic causation of diseases and better predictive models. However, since the presence of cluster structure of genes (gene pathways) was ignored, these methods are insufficient to dissect the complex genetic structure of many common diseases. Here, the clusters are composed of co-regulated genes with coordinated functions. Gene annotation databases, such as kyoto encyclopedia of genes and genomes (KEGG) [33], Reactome [29], Pathway Interaction Database (http://pid.nci.nih.gov/) and BioCyc [17], group functionally relevant genes into biological pathways. The existing information about gene pathways has been gathered through years of biomedical studies [22]. On the other hand, statistical clustering methods such as hierarchical cluster analysis and K-means cluster analysis [27] are also widely used, which provides a more statistically principled way of partitioning predictors into groups. de Souto *et al.* [7] compare different clustering methods and proximity measures for gene expression data. Since it is commonly believed that genes carry out their functions through intricate pathways of reactions and interactions, intuitively, pathway-based analysis can offer an attractive alternative to improve the power of gene (or single-nucleotide polymorphism)-based methods and may help us to identify relevant subsets of genes in meaningful biological pathways underlying complex diseases.

There are considerable interests in pathway-based analysis (to list a few: [18,26,28,34, 39,41,45]). Pathway-based approaches in microarray data analysis often yield biological insights that are otherwise undetectable by focusing only on genes with the strongest evidence of differential expressions. Most pathway-based methods focus on identifying meaningful biological pathways underlying complex diseases, assuming that if a pathway (cluster) is strongly associated with the phenotype, then all genes within that pathway are associated with the phenotype. However, if only a subset of genes within a pathway contributes to the outcome, then these methods may result in loss of power. Our sgSDR is developed to address this problem, where pathway selection and within pathway gene selection can be achieved simultaneously.

We demonstrate our method by analyzing a microarray gene-expression data with glioblastoma patients by Horvath *et al.* [14]. Glioblastoma is the most common and aggressive malignant brain tumor in humans. Patients with this disease have a median survival time of approximately 15 months from the time of diagnosis despite various treatments such as surgery, radiation and chemotherapy. Consisting of two independent sets of clinical tumor samples of $n = 55$ and $n = 65$, the dataset was obtained by Affymetrix HG-U133A arrays and processed by the Robust Multi-array Average method [16]. As Pan *et al.* [34] pointed out, the two datasets were somewhat different from each other, and they only used dataset one in their analysis. Following Pan *et al.* [34], we also focus on the 50 patients with observed survival times from dataset one and took the log survival time (in days) as the response variable in our analysis and the gene-expression profiles as predictors. Our goal is to simultaneously identify significant pathways and genes within those pathways that are strongly associated with the survival time from glioblastoma.

We merged the gene-expression data with the 33 regulatory pathways recorded in the KEGG database. Among the 1668-node of the 33 pathways, 1507 (Entrez ID) out of 22,283 genes (Probe ID) are identified on the HG-U133A chip. Following Li and Li [18], Pan *et al.* [34], and Zhu and Li [45], we only use these 1507 genes in our following analysis. When there are multiple probe set IDs corresponding to a single Entrez KEGG ID, we took the average expression levels of those probe IDs.

We compared our result with Li and Li [18]. As reported in Table 5, our pathway selection is similar to that of Li and Li [18] except for pathway 6, 13, 17, 18, and 27 (cell cycle, extracellular matrix–receptor interaction, gap junction, complement and coagulation cascades, type I diabetes mellitus). Among those five pathways, the first three pathways were selected by our method but not by Li and Li [18], while the latter two were selected by Li and Li [18] only. As reported in [37], the entire tumor growth profile in brain cancer is a collective behavior of cells regulated by the cell cycle pathway (pathway 6). The study result from Phillips laboratory (UCSF) shows that heparan sulfate proteoglycans in extracellular matrix (pathway 13) can change tumor cell behavior including proliferation, invasion, and recruitment of inflammatory cells. Zhu and Li [45] ranked all the 33 pathways according to their significance. Pathway 17 and 27 which were only selected by Li and Li [18], ranked 30th and 28th, respectively, suggesting that they are not very important pathways.

MAPK signaling pathway (pathway 1), cytokine–cytokine receptor interaction pathway (pathway 3), neuroactive ligand–receptor interaction pathway (pathway 5), and complement and coagulation cascades (pathway 18) were ranked as the top four significant pathways related to the brain cancer by Zhu and Li [45] using a nonlinear dimension reduction method. Our pathway selection is consistent with Zhu and Li [45] since all these four pathways are selected by sgSDR. For the within pathway gene selection, our method selected 85 unique genes. Among them, 10 genes are the same as that of Li and Li [18], i.e. MAP3K7, CX3CL1, SYNJ2, UBE2E1, SMURF2, CLDN6, IRF3, IL21R, PCK1, FOXO1A.

Table 5. Pathway selections for glioblastoma data.

Group	Pathway name	sgSDR	Li and Li
1	MAPK signaling pathway	✓	✓
2	Calcium signaling pathway	✓	✓
3	Cytokine–cytokine receptor interaction	✓	✓
4	Phospatidylinositol signaling system	✓	✓
5	Neuroactiveligand–receptor interaction	✓	✓
6	Cell cycle	✓	
7	Ubiquitin mediated proteolysis	✓	✓
8	Apoptosis	✓	✓
9	Wnt signaling pathway	✓	✓
10	Transforming growth factor-beta signaling pathway	✓	✓
11	Axon guidance	✓	✓
12	Focal adhesion	✓	✓
13	Extracellularmatrix–receptor interaction	✓	
14	Cell adhesion molecules	✓	✓
15	Adherens junction	✓	✓
16	Tight junction	✓	✓
17	Gap junction		✓
18	Complement and coagulation cascades	✓	
19	Toll-like receptor signaling pathway	✓	✓
20	Jak-STAT signaling pathway	✓	✓
21	Natural killer cell mediated cytotoxicity	✓	✓
22	Circadian rhythm		
23	Regulation of actin cytotoxicity	✓	✓
24	Insulin signaling pathway	✓	✓
25	Adipocytokine signaling pathway	✓	✓
26	Type II diabetes mellitus	✓	✓
27	Type I diabetes mellitus		✓
28	Alzheimer's disease		
29	Prion diseases		
30	Cocaine addiction		
31	Unknown		
32	Unknown		
33	Unknown		

And FOXO1A was also identified by Pan *et al.* [34] as one of the significant transcription factors associated with glioblastoma.

5. Conclusions and discussion

We propose a method called sgSDR within the framework of SDR that could conduct group and within-group variable selection simultaneously. Our method is comparable to the SGL [10, 36] for the linear models and outperforms it when the regression function is nonlinear. The glioblastoma data are used to illustrate the applications of our method to the gene pathway analysis. Both simulation studies and the real data survival analysis show promising results for sgSDR. Specially, this method has a practical meaning on the application of genetic association research. The consistency of our group and variable selections remains an important yet challenging and open question that deserves further investigation.

Acknowledgments

We would like to thank Dr Caiyan Li, Dr Hongjie Zhu and Dr Zhi Wei for providing help with the glioblastoma data, and Dr Jieping Ye's lab for help with SLEP.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and J. Levine, *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, PNAS 96 (1999), pp. 6745–6750.

[2] S. Bakin, *Adaptive regression and model selection in data mining problems*, Ph.D. thesis, Australian National University, 1999.

[3] H.D. Bondell and L. Li, *Shrinkage inverse regression estimation for model-free variable selection*, J. R. Stat. Soc. Ser. B 71 (2009), pp. 287–299.

[4] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly, *Sparse group lasso: Consistency and climate applications*, Proceedings of the 2012 SIAM International Conference on Data Mining (2012), pp. 47–58.

[5] R.D. Cook, *Regression Graphics*, Wiley, New York, NY, 1998.

[6] R.D. Cook and S. Weisberg, *Discussion of 'Sliced inverse regression for dimension reduction' by Li*, J. Amer. Stat. Assoc. 86 (1991), pp. 328–332.

[7] M.C.P. de Souto, I.G. Costa, D.S.A. de Araujo, T.B. Ludermir, and A. Schliep, *Clustering cancer gene expression data: A comparative study*, BMC Bioinform. 9 (2008), pp. 497.

[8] S. Dudoit, J. Fridyland, and T.P. Speed, *Comparison of discrimination methods for the classification of tumors using gene expression data*, J. Amer. Stat. Assoc. 97 (2002), pp. 77–87.

[9] Z. Feng and L.X. Zhu, *An alternating determination–optimization approach for an additive multi-index model*, Comput. Stat. Data Anal. 56 (2012), pp. 1981–1993.

[10] J. Friedman, T. Hastie, and R. Tibshirani, *A note on the group lasso and a sparse group lasso*, Tech. Rep., Statistics Department, Stanford University, 2010.

[11] P. Hall and K.C. Li, *On almost linearity of low dimensional projections from high dimensional data*, Ann. Stat. 21 (1993), pp. 867–889.

[12] W. Hardle, P. Hall, and H. Ichimura, *Optimal smoothing in single-index models*, Ann. Stat. 21 (1993), pp. 157–178.

[13] A. Hoerl and R. Kennard, *Ridge regression*, Encyclopedia Stat. Sci. 8 (1988), pp. 129–136.

[14] S. Horvath, B. Zhang, M. Carlson, K.V. Lu, S. Zhu, R.M. Felciano, M.F. Laurance, W. Zhao, S. Qi, Z. Chen, Y. Lee, A.C. Scheck, L.M. Liau, H. Wu, D.H. Geschwind, P.G. Febbo, H.I. Kornblum, T.F. Cloughesy, S.F. Nelson, and P.S. Mischel, *Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target*, PNAS 103 (2006), pp. 17402–17407.

[15] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, *A group bridge approach for variable selection*, Biometrika 96 (2009), pp. 339–355.

[16] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed, *Exploration, normalization, and summaries of high density oligonucleotide array probe level data*, Biostatistics 4 (2003), pp. 249–264.

[17] P.D. Karp, C.A. Ouzounis, C. Moore-Kochlacs, L. Goldovsky, P. Kaipa, D. Ahrén, S. Tsoka, N. Darzentas, V. Kunin, and N. López-Bigas, *Expansion of the BioCyc collection of pathway/genome databases to 160 genomes*, Nucleic Acids Res. 19 (2005), pp. 6083–6089.

[18] C. Li and H. Li, *Network-constrained regularization and variable selection for analysis of genomic data*, Bioinformatics 24 (2008), pp. 1175–1182.

[19] K.C. Li, *Sliced inverse regression for dimension reduction*, J. Amer. Stat. Assoc. 86 (1991), pp. 316–327.

[20] L. Li, *Sparse sufficient dimension reduction*, Biometrika 94 (2007), pp. 603–613.

[21] L. Li, R.D. Cook, and C.J. Nachtsheim, *Model free variable selection*, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 67 (2005), pp. 285–299.

[22] L. Li, B. Li, and L.X. Zhu, *Groupwise dimension reduction*, J. Amer. Stat. Assoc. 105 (2010), pp. 1188–1201.

[23] L. Li and C.J. Nachtsheim, *Sparse sliced inverse regression*, Technometrics 48 (2006), pp. 503–510.

[24] L. Li and X. Yin, *Sliced inverse regression with regularizations*, Biometrics 64 (2008), pp. 124–131.

[25] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse learning with efficient projections*, Arizona State Univ. 6 (2009). Available at http://www.public.asu.edu/jye02/Software/SLEP.

[26] S. Ma and M.R. Kosorok, *Identification of differential gene pathways with principal component analysis*, Bioinformatics 25 (2009), pp. 882–889.

[27] J. MacQueen, *Some methods for classification and analysis of multivariate observations*, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, Berkeley, CA, University of California Press, 1967, pp. 281–297.

[28] T. Manoli, N. Gretz, H.-J. Grone, M. Kenzelmann, R. Eils, and B. Brors, *Group testing for pathway analysis improves comparability of different microarray datasets*, Bioinformatics 22 (2006), pp. 2500–2506.

[29] L. Matthews, G. Gopinath, M. Gillesphie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, B. Jassal, A. Kanapin, S. Lewis, S. Mahajan, B. May, E. Schmidt, I. Vastrik, G. Wu, E. Birney, L. Stein, and P. D'Eustachio, *Reactome knowledge bases of biological pathways and processes*, Nucleic Acids Res. 37 (2008), pp. 619–622.

[30] L. Meier, S. van de Geer, and P. Bühlmann, *The group lasso for logistic regression*, J. R. Stat. Soc. Ser. B 70 (2008), pp. 53–71.

[31] NguyenD.V. and D.M. Rocke, *Partial least squares proportional hazard regression for application to DNA microarray survival data*, Bioinformatics 18 (2002), pp. 1625–1632.

[32] L. Ni, R.D. Cook, and TsaiC.-L., *A note on shrinkage sliced inverse regression*, Biometrika 92 (2005), pp. 242–247.

[33] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, *KEGG: Kyoto encyclopedia of genes and genomes*, Nucleic Acids Res. 27 (1999), pp. 29–34.

[34] W. Pan, B. Xie, and X. Shen, *Incorporating predictor network in penalized regression with application to microarray data*, Biometrics 66 (2010), pp. 474–484.

[35] A. Rosenwald, G. Wright, A. Wiestner, W.C. Chan, J.M. Connors, E. Campo, R.D. Gascoyne, T.M. Grogan, H.K. Muller-Hermelink, E.B. Smeland, M. Chiorazzi, J.M. Giltnane, E.M. Hurt, H. Zhao, L. Averett, S. Henrickson, L. Yang, J. Powell, W.H. Wilson, E.S. Jaffe, R. Simon, R.D. Klausner, E. Montserrat, F. Bosch, T.C. Greiner, D.D. Weisenburger, W.G. Sanger, B.J. Dave, J.C. Lynch, J. Vose, J.O. Armitage, R.I. Fisher, T.P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, H. Holte, J. Delabie, and L.M. Staudt, *The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma*, Cancer Cell 3 (2003), pp. 185–197.

[36] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, *The sparse group lasso*, J. Comput. Graph. Stat. 22 (2013), pp. 231–245.

[37] X. Sun, L. Zhang, H. Tan, J. Bao, C. Strouthos, and X. Zhou, *Multi-scale agent-based brain cancer modeling and prediction of TKI treatment response: Incorporating EGFR signaling pathway and angiogenesis*, BMC Bioinform. 13 (2012), pp. 218.

[38] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B 58 (1996), pp. 267–288.

[39] K. Wang, M. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies*, Amer. J. Hum. Genet. 81 (2007), pp. 1278–1283.

[40] T. Wang, P.-R. Xiu, and L.-X. Zhu, *Non-convex penalized estimation in high-dimensional models with single-index structure*, J. Multivariate Anal. 109 (2012), pp. 221–235.

[41] P. Wei and W. Pan, *Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model*, Bioinformatics 24 (2008), pp. 404–411.

[42] X. Wen and R.D. Cook, *Optimal sufficient dimension reduction in regressions with categorical predictors*, J. Stat. Plan. Inference 137 (2007), pp. 1961–1978.

[43] X. Yin, B. Li, and R.D. Cook, *Successive direction extraction for estimating the central subspace in a multiple-index regression*, J. Multivariate Anal. 99 (2008), pp. 1733–1757.

[44] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B 68 (2006), pp. 49–67.

[45] H. Zhu and L. Li, *Biological pathway selection through nonlinear dimension reduction*, Biostatistics 12 (2011), pp. 429–444.

[46] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Stat. Soc. Ser. B (Stat. Methodol.) 67 (2005), pp. 301–320.