Statistica Sinica **21** (2011), 707-730

ASYMPTOTIC PROPERTIES OF SUFFICIENT DIMENSION REDUCTION WITH A DIVERGING NUMBER OF PREDICTORS

Yichao Wu and Lexin Li

North Carolina State University

Abstract: We investigate asymptotic properties of a family of sufficient dimension reduction estimators when the number of predictors p diverges to infinity with the sample size. We adopt a general formulation of dimension reduction estimation through least squares regression of a set of transformations of the response. This formulation allows us to establish the consistency of reduction projection estimation. We then introduce the SCAD max penalty, along with a difference convex optimization algorithm, to achieve variable selection. We show that the penalized estimator selects all truly relevant predictors and excludes all irrelevant ones with probability approaching one, meanwhile it maintains consistent reduction basis estimation for relevant predictors. Our work differs from most model-based selection methods in that it does not require a traditional model, and it extends existing sufficient dimension reduction and model-free variable selection approaches from the fixed p scenario to a diverging p.

Key words and phrases: Central subspace, diverging parameters; SCAD, sliced inverse regression.

1. Introduction

As data with a large number of predictors prevail in many scientific fields such as computational biology, dimension reduction is becoming central to high-dimensional regression analysis of these datasets. Among many dimension reduction methodologies, research in sufficient dimension reduction (SDR), pioneered by Li (1991) and formulated by Cook (1998), has gained considerable interest in recent years. It aims to reduce the predictor dimension by a linear projection of the predictor vector while preserving full regression information. For high-dimensional data, it is often further believed that only a subset of predictors suffice to fully characterize response-predictor relation. Toward this end, simultaneous variable selection along with dimension reduction projection can be achieved (Ni, Cook, and Tsai (2005), Ni et al. (2008), Zhou and He (2008), Bondell and Li (2009). In this article we investigate asymptotic properties of a

family of sufficient dimension reduction methods, in terms of both reduction projection estimation and variable selection, while we allow the number of predictors p to diverge as the sample size n approaches infinity.

Specifically, for regression of a univariate response Y given a predictor vector $\boldsymbol{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^{p\times 1}$, SDR seeks a minimum subspace $\mathcal{S}$, with a $p \times d$ basis matrix $\mathbb{B}$, such that $Y \perp\!\!\!\perp \boldsymbol{X}|\mathbb{B}^T\boldsymbol{X}$. Under minor conditions (Cook (1996), Yin, Li, and Cook (2008)), such a subspace uniquely exists and is a parsimonious population parameter that contains all regression information of $Y|\boldsymbol{X}$. It is named the central subspace, and is denoted by $\mathcal{S}_{Y|X}$ (Cook (1998)). Since the seminal sliced inverse regression (SIR) proposed by Li (1991), there have been a variety of methods proposed to estimate $\mathcal{S}_{Y|X}$ including, for instance, sliced average variance estimation (Cook and Weisberg (1991)), directional regression (Li and Wang (2007)), constructive estimation (Xia (2007)), and sliced regression (Wang and Xia (2008)). Among those methods, SIR is perhaps the most commonly used one for estimating $\mathcal{S}_{Y|X}$, and there have been a number of elaborations on the original methodology of SIR, see for instance, Fung et al. (2002), Yin and Cook (2002), and Cook and Ni (2006). The asymptotic properties of SIR were studied in Li (1991), Hsing and Carroll (1992), Zhu and Ng (1995), and Zhu and Fang (1996). In all those cases, however, the predictor dimension p is treated as fixed. Toward variable selection, Ni, Cook, and Tsai (2005) introduced the lasso type of penalty to SIR to select important predictors along with dimension reduction basis estimation. Zhou and He (2008) imposed the lasso penalty along with thresholding for variable filtering. Ni et al. (2008) and Bondell and Li (2009) generalized the penalized estimation idea to a family of inverse regression estimators and obtained asymptotic properties in terms of consistency in variable selection. Again, p is fixed in those studies and extension to a diverging p is by no means trivial. Recently, there has been work on the diverging p case in the context of sufficient dimension reduction: Zhu, Miao, and Peng (2006) studied the asymptotic properties of SIR as p diverges, but their result is for SIR only, and variable selection is not studied at all; Zhu and Zhu (2009a) investigated weighted partial least squares with a diverging p, but again variable selection is not tackled; Zhu and Zhu (2009b) studied variable selection with a diverging number of predictors through inverse regression, but focused on single-index models only. By contrast, we establish asymptotic properties for a family of inverse regression estimators that includes SIR, study simultaneous dimension reduction and variable selection with a particular emphasis on the latter, and encompass more general model forms.

More specifically, we employ a general formulation of a family of SDR estimators that estimate the central subspace through least squares regression of a set of transformations of the response given the original predictors. This formulation can be viewed as a generalization of the original sliced inverse regression,

and includes SIR as a special case in certain situations. Based on this formulation, we investigate the asymptotic properties of our dimension reduction basis estimator while we allow $p = p_n$ to increase with the sample size n. Under reasonable regularity conditions, we find the rate of convergence of the estimator to be $O_p(\sqrt{p/n})$.

In terms of variable selection, we adopt the SCAD type penalty that was first proposed by Fan and Li (2001), then further developed in Fan and Li (2002), Li and Liang (2008), among others, and combine it with our dimension reduction estimator. It is important to note that exclusion of a predictor in our context of reduction basis estimation requires an entire row of the corresponding basis matrix estimator be zero simultaneously. For this purpose, we employ the SCAD max penalty. We also note that the SDR estimators generally impose no assumption on the conditional distribution $Y|\boldsymbol{X}$ and thus require no traditional models. As a consequence, the penalized SDR estimators achieve variable selection in a model-free fashion. This characteristic distinguishes our result of variable selection with a diverging p from the existing literature, e.g., Fan and Peng (2004), where a parametric model and most often a homoscedastic linear model is assumed. We employ the pseudo-likelihood approach in our proof since no parametric model is imposed. Under suitable conditions, we show that our estimator achieves consistency in variable selection, i.e., the estimator selects all truly relevant predictors with probability approaching one. In addition, the basis estimator of all the relevant predictors is consistent with a $\sqrt{n}$-rate.

The rest of the article is organized as follows. In Section 2, we review a family of SDR estimators and study the convergence as p diverges. In Section 3, we propose the SCAD regularized SDR estimator for variable selection, and investigate its asymptotics with a diverging p in terms of variable selection consistency and basis estimation consistency. We also propose a difference convex algorithm for optimization. We present numerical studies in Section 4, and conclude the paper with a discussion in Section 5. Some technical proofs are given in the Appendix.

2. Dimension Reduction Basis Estimation

2.1. Dimension reduction via response transformation

Throughout, we assume the central subspace $\mathcal{S}_{Y|X}$ exists and its dimension $d = \dim(\mathcal{S}_{Y|X})$ is fixed when $p \to \infty$. This assures that there is a well-defined population parameter as the target of our dimension reduction estimation.

By marginal standardization, if necessary, we assume $E(\boldsymbol{X}) = \mathbf{0}$ and $\mathrm{Var}\,(X_j) = 1$, $j = 1, \ldots, p$. Let $\boldsymbol{\Sigma} = \mathrm{Cov}\,(\boldsymbol{X})$, and define the first moment inverse mean function $\boldsymbol{\phi}(Y) = \Sigma^{-1}E(\boldsymbol{X}|Y)$. Sliced inverse regression is based upon the key observation that, if the linearity condition is satisfied, which states that $E(\boldsymbol{X}|\mathbb{B}^T\boldsymbol{X})$

is a linear function of $\mathbb{B}^T\boldsymbol{X}$ with $\mathbb{B}$ denoting a basis of $\mathcal{S}_{Y|X}$, then $\boldsymbol{\phi}(Y) \in \mathcal{S}_{Y|X}$. The linearity condition is satisfied when $\boldsymbol{X}$ is multivariate normally distributed. Furthermore, Hall and Li (1993) proved that linear combinations of the predictors are approximately normally distributed when $p \to \infty$ as $n \to \infty$, which assures that the linearity condition is satisfied asymptotically. It is also interesting to note that this condition is imposed only on the marginal distribution of $\boldsymbol{X}$, rather than the conditional distribution $Y|\boldsymbol{X}$. For this reason, SIR is viewed as a model-free estimator of the central subspace.

For any function $f(Y)$ satisfying $E\{f(Y)\} = 0$, following Yin and Cook (2002) one can show that

$$E\{f(Y)\boldsymbol{\phi}(Y)\} = \Sigma^{-1}\,\mathrm{Cov}\,\{\boldsymbol{X}, f(Y)\} \in \mathcal{S}_{Y|X} \tag{2.1}$$

under the linearity condition. Consequently, one can choose a series of transformation functions of the response variable, $f_1(Y), \ldots, f_h(Y)$, where h is a prespecified number, and obtain the least squares estimates of regressing $f_k(Y)$ on $\boldsymbol{X}$, i.e.,

$$\boldsymbol{\beta}_k^0 = \arg\min_{\boldsymbol{\beta}_k} E[\{f_k(Y) - \boldsymbol{X}^T\boldsymbol{\beta}_k\}^2], \quad k = 1, \ldots, h. \tag{2.2}$$

Write $\boldsymbol{B}_0 = (\boldsymbol{\beta}_1^0, \ldots, \boldsymbol{\beta}_h^0) \in \mathbb{R}^{p\times h}$, then $\mathrm{Span}(\boldsymbol{B}_0) \subseteq \mathcal{S}_{Y|X}$ by (2.1). By following the usual protocol in the literature of sufficient dimension reduction, we take one step further by assuming the coverage condition $\mathrm{Span}(\boldsymbol{B}_0) = \mathcal{S}_{Y|X}$ whenever $\mathrm{Span}(\boldsymbol{B}_0) \subseteq \mathcal{S}_{Y|X}$. This condition often holds in practice; see Cook and Ni (2006) for a discussion.

There are various choices for the transformation functions $f_k(Y)$. The original SIR corresponds to choosing the slice indicator function $f_k(Y) = 1$ if Y is in slice k, and 0 otherwise, where the response Y is assumed to take h distinctive values $\{1, \ldots, h\}$. Since $E\{f_k(Y)\boldsymbol{\phi}(Y)\} = P(Y = k)\Sigma^{-1}E(\boldsymbol{X}|Y = k)$, $k = 1, \ldots, h$, we have $\mathrm{Span}(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_h) = \Sigma^{-1}\mathrm{Span}(E(\boldsymbol{X}|Y = 1), \ldots, E(\boldsymbol{X}|Y = h))$, and thus it is equivalent to the traditional SIR estimator. Fung et al. (2002) suggested choosing the normalized B-spline basis functions for $f_k(Y)$; Yin and Cook (2002) suggested the normalized polynomial transformation $f_k(Y) = Y^k$ up to power h; Cook and Ni (2006) recommended choosing $f_k(Y) = Y$ if Y is in slice k, and 0 otherwise. We do not address which choice of the transformation function is the "best"; we focus on the asymptotic properties of this family of estimators in general. Moreover, the number of the transformation functions h is a tuning parameter, and although its value matters, it is generally recognized that methods based on inverse means alone are not overly sensitive to the choice of h as long as $h > d$ (Li (1991, Remark 4.3), Cook and Ni (2006, p.71)). Since h usually takes a pre-specified small value in practice, we treat $h(> d)$ as fixed in our asymptotic investigations.

Throughout, we assume that we have n i.i.d. realizations of the data, $\{(\boldsymbol{X}_i, Y_i), i = 1, \ldots, n\}$, and h pre-specified transformation functions $f_1(\cdot), \ldots, f_h(\cdot)$ whose forms do not depend on the data. We then solve the least squares optimization

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}_k} \sum_{i=1}^{n} \{f_k(Y_i) - \boldsymbol{X}_i^T \boldsymbol{\beta}_k\}^2, \quad k = 1, \ldots, h. \tag{2.3}$$

We construct the $p \times h$ matrix $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_h)$, obtain the first d eigenvectors $(\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d)$ of the matrix $h^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T$, and take $\mathrm{Span}(\hat{\boldsymbol{v}}_1, \ldots, \hat{\boldsymbol{v}}_d)$ as an estimate of the targeted central subspace. The structure dimension $d = \dim(\mathcal{S}_{Y|X})$ of the central subspace can be estimated by an asymptotic test (Li (1991)), a permutation test (Cook and Yin (2001)), or an information criterion (Zhu, Miao, and Peng (2006)), and as such d is treated as known in our investigation of reduction basis estimation.

2.2. Asymptotic properties

We now study the asymptotic properties of our estimator of the central subspace. We begin with a lemma that is a key for our main asymptotic result in Theorem 1.

Lemma 1. *Suppose Conditions* (i), (ii), *and* (iii) *of Appendix A hold. When* $p(\log n)/n \to 0$, *there exists a constant* $a^* > 0$ *such that*

$$P\left(\inf_{\|\boldsymbol{\beta}\|=1} \sum_{i=1}^{n} (\boldsymbol{X}_i^T \boldsymbol{\beta})^2 > a^* n\right) \to 1 \text{ as } n \to \infty.$$

Proof of Lemma 1. For n *i.i.d.* standard normal errors ϵ_i, $i = 1, \ldots, n$, we construct artificial data $\{(\boldsymbol{X}_i, \tilde{Y}_i), i = 1, \ldots, n\}$, where $\tilde{Y}_i = \boldsymbol{X}_i^T \tilde{\boldsymbol{\beta}} + \epsilon_i$ for some fixed $\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p$. The desired result follows by applying Lemma 3.1 of Portnoy (1984) with $\psi(t) = t$.

For any function $f(\cdot)$ satisfying that $E\{f(Y)\} = 0$, let

$$\boldsymbol{\beta}^0 = \boldsymbol{\beta}^0(f) = \operatorname*{argmin}_{\boldsymbol{\beta}} E\left\{f(Y) - \boldsymbol{X}^T \boldsymbol{\beta}\right\}^2, \tag{2.4}$$

$$\hat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}(f)} = \operatorname*{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{f(Y_i) - \boldsymbol{X}_i^T \boldsymbol{\beta}\right\}^2. \tag{2.5}$$

Theorem 1. *Suppose Conditions* (i)–(vii) *of Appendix A hold. If* $p(\log n)/n \to 0$, $\hat{\boldsymbol{\beta}}$ *is a consistent estimator for* $\boldsymbol{\beta}^0$ *with* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{p/n})$.

The proof of this theorem is given in Appendix B. It serves as a basis for our consistency result for estimating the central subspace.

Corollary 1. *Consider a set of response transformation functions,* $\{f_1(\cdot), \ldots, f_h(\cdot)\}$*, each of which satisfies Conditions* (vi) *and* (vii) *of Appendix A. Under Conditions* (i)−(v) *of Appendix A, we have* $\|h^{-1}(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T - \boldsymbol{B}_0\boldsymbol{B}_0^T)\| = O_p(\sqrt{p/n})$*,* $\boldsymbol{B}_0$ *and* $\hat{\boldsymbol{B}}$ *as at* (2.2) *and* (2.3)*.*

Proof of Corollary 1. Note first that we assume h is finite and fixed. Consequently $\|\boldsymbol{B}_0\| = O(1)$. Theorem 1implies that the Frobenius norm of $\hat{\boldsymbol{B}} - \boldsymbol{B}_0$ satisfies $\|\hat{\boldsymbol{B}} - \boldsymbol{B}_0\| = O_p(\sqrt{p/n})$. Therefore,

$$
\begin{aligned}
&\|h^{-1}(\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T - \boldsymbol{B}_0\boldsymbol{B}_0^T)\| \\
&\quad \le h^{-1}(\|(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)^T\| + \|\boldsymbol{B}_0(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)^T\| + \|(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)\boldsymbol{B}_0^T\|) \\
&\quad \le h^{-1}(\|(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)\|\|(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)^T\| + \|\boldsymbol{B}_0\|\|(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)^T\| + \|(\hat{\boldsymbol{B}} - \boldsymbol{B}_0)\|\|\boldsymbol{B}_0^T\|) \\
&\quad \le h^{-1}\left(O_p(\sqrt{\frac{p}{n}})^2 + O(1)O_p(\sqrt{\frac{p}{n}}) + O(1)O_p(\sqrt{\frac{p}{n}})\right) \\
&\quad = O_p(\sqrt{\frac{p}{n}}).
\end{aligned}
$$

Remark 1. Zhu, Miao, and Peng (2006) studied the asymptotics of the original SIR estimator when p diverges. In their study, they fixed the number of sample points in each slice while letting the number of slices $h \to \infty$ as $n \to \infty$. In our study, this notion of fixed number of observations per slice no longer applies for a choice of transformation functions other than the indicator function. Besides, since in practice h is pre-determined, we choose to treat h as fixed in our asymptotic investigations. For these reasons, our consistency rate is not directly comparable to that obtained by Zhu, Miao, and Peng (2006) for the original SIR, while our result goes beyond SIR and applies to the entire family of SDR estimators based on the first inverse moment $\boldsymbol{\phi}(Y)$, as discussed in Section 2.1.

We can further bound estimation error of the first d eigenvectors of $h^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T$ when the first d eigenvalues of $h^{-1}\boldsymbol{B}_0\boldsymbol{B}_0^T$ are distinct. Let $\boldsymbol{A}_0 = h^{-1}\boldsymbol{B}_0\boldsymbol{B}_0^T$, $\hat{\boldsymbol{A}} = h^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T$, and $\boldsymbol{E} = \hat{\boldsymbol{A}} - \boldsymbol{A}_0$. Denote the eigenvalues and eigenvectors of $\boldsymbol{A}_0$ by $\{\lambda_j, \boldsymbol{v}_j\}$, $j = 1, \ldots, p$.

Theorem 2. *Suppose* $\lambda_1, \ldots, \lambda_d$ *are unique. Under the conditions of Corollary 1, the estimated eigenvectors* $\hat{\boldsymbol{v}}_j$*s satisfy*

$$
\sqrt{1 - (\boldsymbol{v}_j^T\hat{\boldsymbol{v}}_j)^2} = O_p(\sqrt{\frac{p}{n}}), \quad j = 1, \ldots, d.
$$

Proof of Theorem 2. For $j = 1, \ldots, d$, we rearrange the order of the first d elements and write $\boldsymbol{Q}_j = (\boldsymbol{v}_j, \boldsymbol{v}_{j+1}, \ldots, \boldsymbol{v}_d, \boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{j-1}, \boldsymbol{v}_{d+1}, \boldsymbol{v}_{d+2}, \ldots, \boldsymbol{v}_p) =$

$(\boldsymbol{v}_j, \boldsymbol{Q}_{j2})$. Then $\boldsymbol{Q}_j^T \boldsymbol{A}_0 \boldsymbol{Q}_j = \text{diag}(\lambda_j, \lambda_{j+1}, \ldots, \lambda_d, \lambda_1, \lambda_2, \ldots, \lambda_{j-1}, \lambda_{d+1}, \lambda_{d+2}, \ldots, \lambda_p)$. Next write $\boldsymbol{Q}_j^T \boldsymbol{E} \boldsymbol{Q}_j = \begin{bmatrix} \varepsilon_j & \boldsymbol{\epsilon}_j \\ \boldsymbol{\epsilon}_j^T & \boldsymbol{E}_{22j} \end{bmatrix}$ and $a_j = \min_{j \neq i} |\lambda_i - \lambda_j| > 0$. Note that $\|\boldsymbol{E}\| = O_p(\sqrt{p/n})$ due to Corollary 1. By applying Theorem 8.1.12 of Golub and van Loan (1996), there exists $\boldsymbol{p}_j \in \mathbb{R}^{p-1}$ satisfying $\|\boldsymbol{p}_j\| \leq 4\|\boldsymbol{\epsilon}_j\|/a_j$ such that $\hat{\boldsymbol{v}}_j = (\boldsymbol{v}_j + \boldsymbol{Q}_{j2}\boldsymbol{p}_j)/\sqrt{1 + \boldsymbol{p}_j^T \boldsymbol{p}_j}$ is a unit eigenvector of $\hat{\boldsymbol{A}} = \boldsymbol{A}_0 + \boldsymbol{E}$. Furthermore, $\sqrt{1 - (\boldsymbol{v}_j^T \hat{\boldsymbol{v}}_j)^2} \leq 4\|\boldsymbol{\epsilon}_j\|/a_j$. Since $\|\boldsymbol{\epsilon}_j\| \leq \|\boldsymbol{Q}_j^T \boldsymbol{E} \boldsymbol{Q}_j\| \leq \|\boldsymbol{Q}_j\| \|\boldsymbol{E}\| \|\boldsymbol{Q}_j\| = \|\boldsymbol{E}\| = O_p(\sqrt{p/n})$, the result follows.

3. Variable Selection

3.1. Regularization via the SCAD max penalty

When the number of predictors p is large in a regression analysis, regularization is often employed to add numerical stability, to improve statistical robustness, and to achieve variable selection. In the context of model-based variable selection, there has been an extensive literature on model selection via regularization for the Lasso (Tibshirani (1996)), the SCAD (Fan and Li (2001)), the nonnegative garrote (Breiman (1995)), and the adaptive Lasso (Zou (2006)), among many others. In particular, Fan and Li (2001) first demonstrated that the SCAD penalty possesses the oracle properties in the sense that the regularized estimator correctly selects predictors with nonzero coefficients in the model, excludes those with zero coefficients with probability approaching one, and estimates those nonzero coefficients with the asymptotic distribution they would have if all the zero coefficients were known in advance. Fan and Peng (2004) established these properties of the SCAD for linear models, and Zhu and Zhu (2009b) employed the SCAD for variable selection in single-index models, both with a diverging p. Here we adopt the SCAD max penalty for the purpose of variable selection when p tends to infinity, but we do not impose any parametric or semi-parametric models.

Before pursuing variable selection in the framework of sufficient dimension reduction, we first note that the notions of relevant and irrelevant variables need to be clearly defined, since in SDR estimation no parametric model is imposed. Toward that end, Cook (2004) and Bondell and Li (2009) showed that, as long as the central subspace $\mathcal{S}_{Y|X}$ exists, there exists a unique partition of the predictors $\boldsymbol{X} = (\boldsymbol{X}_+^T, \boldsymbol{X}_-^T)^T$, $\boldsymbol{X}_+ \in \mathbb{R}^{q \times 1}$, and $\boldsymbol{X}_- \in \mathbb{R}^{(p-q) \times 1}$, such that

$$Y \perp\!\!\!\perp \boldsymbol{X}_- | \boldsymbol{X}_+. \tag{3.1}$$

Thus the regression of Y on $\boldsymbol{X}$ only relies on the set of predictors $\boldsymbol{X}_+$, which we call the relevant variables, while $\boldsymbol{X}_-$ is irrelevant. Without loss of generality,

we assume that $\boldsymbol{X}_+$ consists of the first q predictors. Moreover we assume the number of relevant predictors q is fixed as $p \to \infty$. That is, we regard all regression information as concentrated on a fixed number of predictors with the rest of additional variables as nuisance information. We think this condition reasonable, based upon the belief that, in many real applications, increasing the number of predictors after a certain stage does not necessarily induce an increasing amount of useful information. We then have a well-defined population target for the purpose of variable selection in the absence of a traditional model.

Predictor partition as in (3.1) can be directly connected with the basis $\mathbb{B}$ of the central subspace; that is, the last $p-q$ rows of $\mathbb{B}$ must all be zeros (Cook and Ni (2006, Prop. 1)). It also leads to the following lemma in our context of least squares estimation of the central subspace.

Lemma 2. *For $\boldsymbol{\beta}^0$ at (2.4), we have $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_+^{0T}, \mathbf{0}_{(p-q)\times 1}^T)^T$ at (3.1), where $\boldsymbol{\beta}_+^0 = \operatorname{argmin}_{\boldsymbol{\beta}_+ \in \mathbb{R}^q} E\{f(Y) - \boldsymbol{X}_+^T \boldsymbol{\beta}_+\}$ when the linearity condition is satisfied.*

Proof of Lemma 2. Under the linearity condition, we have $\boldsymbol{\beta}^0 \in \mathcal{S}_{Y|X}$ so that $\boldsymbol{\beta}^0$ can be written as a linear combination of the columns of the central subspace basis $\mathbb{B}$. Since the last $p-q$ rows of $\mathbb{B}$ must all be zeros, the result follows.

The class of SDR estimators studied in Section 2 yield linear combinations of all the original predictors and thus perform no variable selection. We introduce a non-concave penalty to achieve selection of relevant predictors. For a set of transformation functions, $f_1, \ldots, f_h$, define the negative pseudo loglikelihood function

$$L_k(\boldsymbol{\beta}_k) = n^{-1} \sum_{i=1}^{n} \left\{ f_k(Y_i) - \boldsymbol{X}_i^T \boldsymbol{\beta}_k \right\}^2 .$$

Applying the max-type penalty, we propose to minimize

$$Q(\boldsymbol{B}) = \left\{ \sum_{k=1}^{h} L_k(\boldsymbol{\beta}_k) + \sum_{j=1}^{p} p_\lambda\Big(\max_{1 \le k \le h} |\beta_{jk}| \Big) \right\} \tag{3.2}$$

over $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_h)$, where β_{jk} is the j-th element of $\boldsymbol{\beta}_k \in \mathbb{R}^{p\times 1}$, $j = 1, \ldots, p, k = 1, \ldots, h$. Here $p_\lambda(\theta)$ is a general penalty function indexed by a regularization parameter λ. For now we simply assume p_λ is symmetric, singular at the origin, and non-decreasing and concave on $[0, \infty)$. Later in this section, we introduce a specific non-concave form, the SCAD penalty function, for $p_\lambda(\theta)$.

Two observations are noteworthy here. First, the minimization in (3.2) is over the entire $p \times h$ matrix $\boldsymbol{B}$, since the penalty is imposed on the maximum over

each row of $\boldsymbol{B}$. This is different from the dimension reduction basis estimation without regularization as discussed in Section 2.1, where the minimization is carried over each column $\boldsymbol{\beta}_k$ of $\boldsymbol{B}$ individually. Second, variable selection achieved through (3.2) requires no dimension reduction basis estimation as a preprocessing step, and thus requires no knowledge of the structural dimension d either. For this reason, the penalty term in (3.2) has ph parameters rather than pd parameters. Selection is done essentially in one step instead of two steps, which to some degree mitigates the dependency of variable selection on the accuracy of reduction basis estimation, and can be particularly useful if model-free variable selection is the sole purpose of the study.

With a slight abuse of notation, we denote the minimizer of (3.2) as $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_h)$, and denote the minimizer of the corresponding population version $\sum_{k=1}^{h} E\{f_k(Y) - \boldsymbol{\beta}_k^T \boldsymbol{X}\}^2$ as $\boldsymbol{B}_0 = (\boldsymbol{\beta}_1^0, \ldots, \boldsymbol{\beta}_h^0)$. We use $\hat{\boldsymbol{B}}_+ = (\hat{\boldsymbol{\beta}}_{1+}, \ldots, \hat{\boldsymbol{\beta}}_{h+})$ to denote the submatrix of $\hat{\boldsymbol{B}}$ that consists of its first q rows, and similarly denote the first q rows of $\boldsymbol{B}^0$ as $\boldsymbol{B}_+^0 = (\boldsymbol{\beta}_{1+}^0, \ldots, \boldsymbol{\beta}_{h+}^0)$. We next aim to show that $\hat{\boldsymbol{\beta}}_{k+} \to \boldsymbol{\beta}_{k+}^0$ as $n \to \infty$, and that the j-th element $\hat{\beta}_{jk}$ of $\hat{\boldsymbol{\beta}}_k$ satisfies $P(\hat{\beta}_{jk} = 0) \to 1$ for $j > q$, $k = 1, \ldots, h$.

3.2. Asymptotic properties

Let $\lambda = \lambda_n$. For a general non-concave penalty function $p_{\lambda_n}(\cdot)$, let $a_n = \max_{1 \le j \le p} p'_{\lambda_n}(\max_{1 \le k \le h} |\beta_{jk}^0|)$ and $b_n = \max_{1 \le j \le p} p''_{\lambda_n}(\max_{1 \le k \le h} |\beta_{jk}^0|)$, where β_{jk}^0 is the j-th element of β_k^0, and $p'_{\lambda_n}(\cdot)$ and $p''_{\lambda_n}(\cdot)$ denote the first and second order derivative, respectively.

Lemma 3. *Suppose $\boldsymbol{X}$ satisfies Conditions* (iv) *and* (v), *and that each of the response transformation functions $f_1(\cdot), \ldots, f_h(\cdot)$, satisfies Conditions* (vi) *and* (vii) *in Appendix A. If $p = o(n^{1/4})$, and the penalty function $p_{\lambda_n}(\cdot)$ has $a_n = O(1/\sqrt{n})$ and $b_n = o(1)$, then there exists a local minimizer $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_h)$ of $Q(\boldsymbol{B})$ in* (3.2) *such that $\|\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^0\| = O_p(\sqrt{p}(n^{-1/2} + a_n))$, $k = 1, 2, \ldots, h$.*

Lemma 4. *Suppose that each of the response transformation functions, $f_1(\cdot), \ldots, f_h(\cdot)$, satisfies Conditions* (vi) *and* (vii) *in Appendix A, and that p_{λ_n} satisfies $\liminf_{n \to \infty} \inf_{\theta \to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$. If $\lambda_n \to 0$, $\sqrt{p/n}/\lambda_n \to 0$, and $p = o(n^{1/4})$ as $n \to \infty$, then for any given $q \times h$ submatrix $\boldsymbol{B}_+ = (\boldsymbol{\beta}_{1+}, \ldots, \boldsymbol{\beta}_{h+})$ satisfying $\|\boldsymbol{\beta}_{k+} - \boldsymbol{\beta}_{k+}^0\| = O_p(\sqrt{p/n})$, $k = 1, \ldots, h$, and any $(p-q) \times h$ submatrix $\boldsymbol{B}_- = (\boldsymbol{\beta}_{1-}, \ldots, \boldsymbol{\beta}_{h-})$ satisfying that $\|\boldsymbol{\beta}_{k-}\| \le C\sqrt{p/n}$ for a constant C, $k = 1, \ldots, h$, with probability tending to one,*

$$Q((\boldsymbol{B}_+^T, \boldsymbol{0}_{(p-q)\times h}^T)^T) = \min_{\|\boldsymbol{\beta}_{k-}\| \le C\sqrt{p/n}} Q((\boldsymbol{B}_+^T, \boldsymbol{B}_-^T)^T).$$

The proofs of these two lemmas are given in Appendix B.

Theorem 3. *Under the conditions of Lemmas* 3 *and* 4*, with probability tending to one, the* $\sqrt{p/n}$*-consistent local minimizer of* $Q(\boldsymbol{B})$ *satisfies*

(i) $\hat{\beta}_{jk} = 0$ *for* $j > q$ *and* $1 \le k \le h$*;*

(ii) $\hat{\beta}_{jk}$ *for* $1 \le j \le q$ *and* $1 \le k \le h$ *have the same asymptotic distribution as the minimizers of*

$$\tilde{Q}(\boldsymbol{B}_+) = n^{-1} \sum_{k=1}^{h} \sum_{i=1}^{n} (f_k(Y_i) - \boldsymbol{X}_{i+}^T \boldsymbol{\beta}_{k+})^2 + \sum_{j=1}^{q} p_{\lambda_n}(\max_{1 \le k \le h} |\beta_{jk}|)$$

over $\boldsymbol{B}_+ = (\boldsymbol{\beta}_{1+}, \ldots, \boldsymbol{\beta}_{h+})$*, where* β_{jk} *is the* j*-th element of* $\boldsymbol{\beta}_{k+} \in \mathbb{R}^{q \times 1}$*,* $j = 1, \ldots, q$*,* $k = 1, \ldots, h$*, and* $\boldsymbol{X}_{i+}$ *is the* i*-th observation of* $\boldsymbol{X}_+$*.*

Proof of Theorem 3. By Lemma 3, there exists a $\sqrt{p/n}$-consistent local minimizer $\hat{\boldsymbol{B}}$ of $Q(\boldsymbol{B})$. Part (1) holds by Lemma 4, that is, $\hat{\boldsymbol{B}} = (\hat{\boldsymbol{B}}_+^T, \mathbf{0}_{(p-q)\times h}^T)^T$ with probability tending to one. Consequently, with probability tending to one, we are in effect minimizing $\tilde{Q}(\cdot)$. Then part (2) follows.

Remark 2. The asymptotic distributional result is given in a way similar to that in Knight and Fu (2000). For a non-concave max penalty, in general, the explicit asymptotic normality result, as in Fan and Li (2001) and Fan and Peng (2004), is not available because there may exist a tie $|\beta_{jk}^0| = |\beta_{jk'}^0|$ for some $1 \le j \le q$ and $k \ne k'$. For some specific non-concave max penalty, the asymptotic normality result is possible, as we discuss next.

We introduce a specific form of a non-concave penalty function, the SCAD penalty first proposed by Fan and Li (2001). Define a penalty function $p_{\lambda_n}(\theta)$ through its first derivative

$$p'_{\lambda_n}(\theta) = \lambda_n \left\{ I(\theta \le \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n) \right\}, \theta \ge 0, \tag{3.3}$$

where a is an additional parameter. It is easy to see that this function satisfies the non-concave penalty condition. Note that $p_{\lambda_n}(\theta)$ flattens out for $|\theta| > a\lambda_n$. Consequently, $a_n = 0$ and $b_n = 0$ as long as $\lambda_n < a^{-1} \max_{1 \le j \le q, 1 \le k \le h} |\beta_{jk}^0|$. This feature enables us to refine the result of Theorem 3, and leads to the following corollary.

Corollary 2. *For the SCAD penalty,* $a_n = 0$ *and* $b_n = 0$ *when* $\lambda_n < a^{-1} \max_{1 \le j \le q, 1 \le k \le h} |\beta_{jk}^0|$*. Then under the conditions of Lemmas* 3 *and* 4*, with probability tending to one, the* $\sqrt{p/n}$*-consistent local minimizer of* $Q(\boldsymbol{B})$ *satisfies*

(i) $\hat{\beta}_{jk} = 0$ *for* $j > q$ *and* $1 \le k \le h$;

(ii) $\sqrt{n}(\hat{\boldsymbol{\beta}}_{k+} - \boldsymbol{\beta}^0_{k+}) \to N(0, \boldsymbol{\Sigma}_+^{-1}\boldsymbol{\Sigma}_{k+}\boldsymbol{\Sigma}_+^{-1})$ *for* $k = 1, \ldots, h$, *where* $\boldsymbol{\Sigma}_+ = \mathrm{Cov}(\boldsymbol{X}_+)$ *and* $\boldsymbol{\Sigma}_{k+} = \mathrm{Var}\{f_k(Y)\boldsymbol{X}_+\}$.

Proof of Corollary 2. It is straightforward to verify that the SCAD penalty satisfies all the penalty-related conditions in Theorem 3. Since $\hat{\boldsymbol{\beta}}_{k+}$, $k = 1, \ldots, h$, are consistent, and $\lambda_n < a^{-1} \max_{1 \le j \le q, 1 \le k \le h} |\beta^0_{jk}|$ asymptotically, we are optimizing $\tilde{Q}(\boldsymbol{B}_+)$ in a neighborhood of $(\boldsymbol{\beta}_{1+}, \ldots, \boldsymbol{\beta}_{h+})$ satisfying $\max_{1 \le k \le h} |\beta_{jk}| > a\lambda$ for $j = 1, \ldots, q$. Correspondingly, $p_{\lambda_n}(\max_{1 \le k \le h} |\beta_{jk}|)$ reduces to $p_{\lambda_n}(a\lambda_n) = (a+1)\lambda_n^2/2$, which does not depend on $\boldsymbol{\beta}_{k+}$, $k = 1, \ldots, h$. As such,

$$\underset{\boldsymbol{\beta}_{1+}, \ldots, \boldsymbol{\beta}_{h+}}{\operatorname{argmin}} \frac{1}{n} \sum_{k=1}^{h} \sum_{i=1}^{n} \left\{ f_k(Y_i) - \boldsymbol{X}_{i+}^T \boldsymbol{\beta}_{k+} \right\}^2 + \sum_{j=1}^{q} p_{\lambda_n}(\max_{1 \le k \le h} |\beta_{jk}|)$$

is the same as

$$\underset{\boldsymbol{\beta}_{1+}, \ldots, \boldsymbol{\beta}_{h+}}{\operatorname{argmin}} \sum_{k=1}^{h} \sum_{i=1}^{n} \frac{1}{n} \left\{ f_k(Y_i) - \boldsymbol{X}_{i+}^T \boldsymbol{\beta}_{k+} \right\}^2.$$

The desired result follows.

Remark 3. Corollary 2 is a special case of Theorem 3 since the SCAD penalty function is a special case of the general non-concave penalty function. This refined result is possible because that the SCAD function is flat when its argument is larger than $a\lambda$ in magnitude. Consequently there is no asymptotic bias in using $\hat{\boldsymbol{B}}_+$ to estimate $\boldsymbol{B}_+$. This is in a similar spirit as the result of Theorem 2 of Fan and Li (2001).

Remark 4. We obtain the $\sqrt{n}$-rate for dimension reduction basis estimation after variable selection because the number of truly relevant predictors q is assumed fixed. Consequently, with the SCAD regularized estimator selecting all truly relevant predictors and excluding all irrelevant ones with probability one, the basis estimation based on those relevant predictors achieves a $\sqrt{n}$-rate.

Remark 5. Our results differ from those of Fan and Li (2001) and Fan and Peng (2004), in that they require a parametric linear model and all results hinge on the model being correctly specified. By contrast, our approach does not require a traditional model, and our technical proofs are based on the pseudo-likelihood function.

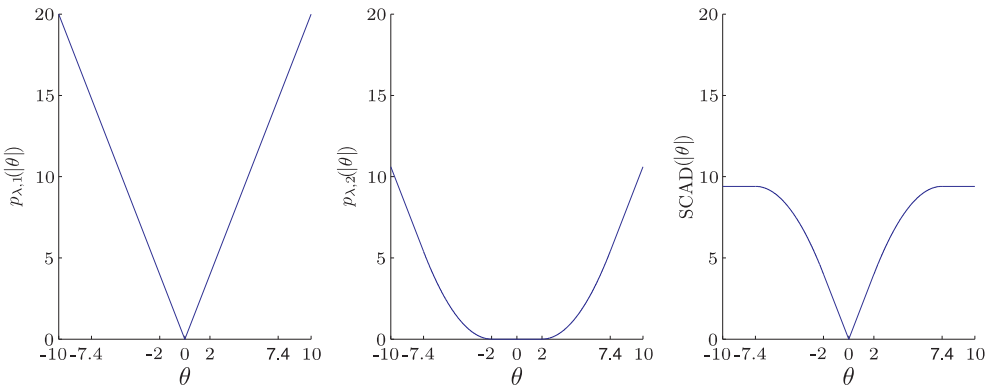

Figure 1. Decomposition of the SCAD penalty as $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$, with parameters $\lambda = 2$ and $a = 3.7$

3.3. Optimization algorithm

We propose an algorithm to minimize $Q(\boldsymbol{B})$ in (3.2). Note that the SCAD type penalty is non-concave, and thus it requires some specially designed optimization algorithm. In the literature, there exist a number of such algorithms, including local quadratic approximation (Fan and Li (2001)), the minorize-maximize algorithm (Hunter and Li (2005)), local linear approximation (Zou and Li (2008)), and the difference convex algorithm (DC, An and Tao (1997) Wu and Liu (2009). For our problem, we employ the DC algorithm, that solves a non-concave optimization problem via a sequence of convex optimizations by decomposing the non-concave objective function as the difference of two convex functions.

For the SCAD penalty, we note that its first derivative as given in (3.3) can be decomposed as $p'_\lambda(\theta) = p'_{\lambda,1}(\theta) + p'_{\lambda,2}(\theta)$, where $p'_{\lambda,1}(\theta) = \lambda$ is a constant and $p'_{\lambda,2}(\theta) = \lambda[1 - (a\lambda - \theta)_+/\{(a-1)\lambda\}]I(\theta > \lambda)$ is a decreasing function on the range $\theta > 0$. Accordingly, the SCAD penalty function can be decomposed as $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$, where both $p_{\lambda,1}(\cdot)$ and $p_{\lambda,2}(\cdot)$ are convex, with $p'_{\lambda,1}(\theta)$ and $p'_{\lambda,2}(\theta)$ as the derivative, respectively. Figure 1 illustrates such a decomposition for a SCAD function with a particular set of parameters, $a = 3.7$ and $\lambda = 2$, where the left panel plots $p_{\lambda,1}(\theta)$, the central panel $p_{\lambda,2}(\theta)$, and the right panel $p_\lambda(\theta) = p_{\lambda,1}(\theta) - p_{\lambda,2}(\theta)$.

We next decompose the objective function in (3.2) as $Q(\boldsymbol{B}) = Q_{vex}(\boldsymbol{B}) + Q_{cav}(\boldsymbol{B})$, where

$$Q_{vex}(\boldsymbol{B}) = \sum_{k=1}^{h} L_k(\boldsymbol{\beta}_k) + \sum_{j=1}^{p} p_{\lambda,1}(\max_{1\le k\le h} |\beta_{jk}|),$$

$$Q_{cav}(\boldsymbol{B}) = -\sum_{j=1}^{p} p_{\lambda,2}(\max_{1\le k\le h} |\beta_{jk}|).$$

We initialize $\boldsymbol{B} = \boldsymbol{B}^{(0)}$ and then update $\boldsymbol{B}$ iteratively. At the $(t+1)$-th step, the DC algorithm uses a linear function $-\sum_{j=1}^{p} p'_{\lambda,2}(\max_{1\le k\le h}|\beta_{jk}^{(t)}|)(\max_{1\le k\le h}|\beta_{jk}| - \max_{1\le k\le h}|\beta_{jk}^{(t)}|)$ to approximate the concave part $Q_{cav}(\boldsymbol{B})$, where $\beta_{jk}^{(t)}$ denotes the (j,k)-th element of the solution $\boldsymbol{B}^{(t)}$ from the t-th step. Then minimizing $Q(\boldsymbol{B})$ amounts to solving

$$\boldsymbol{B}^{(t+1)} = \operatorname*{argmin}_{\boldsymbol{B}}\Big\{Q_{vex}(\boldsymbol{B}) - \sum_{j=1}^{p} p'_{\lambda,2}(\max_{1\le k\le h}|\beta_{jk}^{(t)}|)(\max_{1\le k\le h}|\beta_{jk}| - \max_{1\le k\le h}|\beta_{jk}^{(t)}|)\Big\}. \quad (3.4)$$

Optimization in (3.4) can be further formulated as a quadratic programming problem by letting $\xi_j^{(t)} = \max_{1\le k\le h}|\beta_{jk}^{(t)}|$, then minimizing

$$n^{-1}\sum_{k=1}^{h}\sum_{i=1}^{n}(f_k(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}_k)^2 + \sum_{j=1}^{p}(\lambda - p'_{\lambda,2}(\xi_j^{(t)}))\xi_j$$

over $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_h)$ subject to $\xi_j \ge \beta_{jk}$ and $\xi_j \ge -\beta_{jk}$, $j = 1, \ldots, p, k = 1, \ldots, h$. Existing software is available to solve this quadratic programming problem.

Hunter and Li (2005) studied the convergence property of their minorize-maximize (MM) algorithm for the SCAD penalty. Our DC solution can also be viewed as an instance of their MM algorithm, since we replace the concave part $Q_{cav}(\boldsymbol{B})$ by its affine minorization at each iteration. As the objective function $Q(\boldsymbol{B})$ is nonnegative, by the descent property of the MM algorithm, our DC algorithm is bound to converge to an ϵ-local minimizer in finite steps. Practically, we deem the algorithm convergent if $\sum_{k=1}^{h}\sum_{j=1}^{p}|\beta_{jk}^{(t)} - \beta_{jk}^{(t+1)}|$ is sufficient small, e.g., less than 10^{-4}.

4. Numerical Studies

In this section, we examine the finite sample performance of the proposed method using both simulations and a data example. We employed the BIC type criterion to select the tuning parameter λ for the SCAD penalty, $\sum_{k=1}^{h}\sum_{i=1}^{n}(f_k(y_i) - \boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_k^{(\lambda)})^2 + n^{(\lambda)}\log n$, where $n^{(\lambda)} = \#\{j : \max_{1\le k\le h}|\hat{\beta}_{jk}^{(\lambda)}| > 0\}$ denotes the number of active predictors at λ. The BIC criterion has been commonly used in regularized variable selection, e.g., Wang, Li and Tsai (2007). For transformation functions, we implemented the slice indicator function that gives the usual SIR estimate, and the B-spline basis function suggested in Fung et al. (2002). For the former, we fixed the number of slices at $h = 5$ and, for the latter, we used a linear spline with three inner knots, which also yields $h = 5$.

Table 1. Evaluation of dimension reduction basis estimation for Examples 4.1 and 4.2. Reported are the mean and standard deviation (in parentheses) of the vector correlation coefficients.

	Example 4.1 with $d = 2$		Example 4.2 with $d = 3$	
	Slicing	Spline	Slicing	Spline
	$p = 20, n = 400$		$p = 20, n = 600$	
w/o penalty	0.92 (0.03)	0.88 (0.04)	0.88 (0.03)	0.85 (0.06)
SCAD	0.98 (0.02)	0.92 (0.11)	0.96 (0.03)	0.96 (0.04)
	$p = 40, n = 800$		$p = 40, n = 1,200$	
w/o penalty	0.92 (0.02)	0.87 (0.03)	0.87 (0.02)	0.84 (0.04)
SCAD	0.99 (0.01)	0.99 (0.01)	0.98 (0.01)	0.98 (0.01)

4.1. Simulations

For Examples 4.1 and 4.2, we generated independent X_j from the standard normal. We also considered correlated predictors with $\mathrm{Corr}(X_i, X_j) = 0.5^{|i-j|}$, $1 \le i, j \le p$.

Example 4.1. Here

$$Y = \frac{\boldsymbol{X}^T\boldsymbol{\beta}_1}{0.5 + (1.5 + \boldsymbol{X}^T\boldsymbol{\beta}_2)^2} + 0.2\epsilon,$$

where $\epsilon \sim Normal(0, 1)$ is independent of $\boldsymbol{X}$. In this model the structural dimension $d = 2$. We chose $\boldsymbol{\beta}_1 = (1, 1, 0, \ldots, 0)^T$ and $\boldsymbol{\beta}_2 = (0, 0, 1, 1, 0, \ldots, 0)^T$. We considered $n = 400$, $p = 20$ and $n = 800$, $p = 40$. We employed the vector correlation coefficient (Hotelling (1936)) to evaluate the accuracy of the dimension reduction basis estimation, and it ranges between 0 and 1 with a larger value indicating a better estimate. Results based on 100 data replications are reported in Table 1 (left half), where the mean and standard deviation (in parentheses) of the vector correlations between the true and the estimated central subspace basis are shown. We compared the usual SDR estimator without penalty and the one with the SCAD max penalty. Due to the sparse nature of the central subspace basis, the penalized SDR estimator achieved a better estimation accuracy. To evaluate the performance in terms of variable selection, we employ the true positive rate and the false positive rate, a pair of criteria that are commonly used in biomedical research. Table 2 (left half) reports the average results of the penalized SDR estimator. It is clearly seen that all truly relevant predictors were selected, while the false positive rate was low. Moreover, two choices of transformation functions had similar empirical performance in this example.

Example 4.2. Here $Y = \mathrm{sign}(\boldsymbol{X}^T\boldsymbol{\beta}_1) \log |\boldsymbol{X}^T\boldsymbol{\beta}_2 + 5| + \boldsymbol{X}^T\boldsymbol{\beta}_3 + 0.2\epsilon$, where $\epsilon \sim Normal(0, 1)$ is independent of $\boldsymbol{X}$. In this example, the structural dimension

Table 2. Evaluation of variable selection for Examples 4.1 and 4.2. Reported are the mean and standard deviation (in parentheses) of true positive rate (TPR) and false positive rate (FPR).

	Example 4.1 with $d = 2$		Example 4.2 with $d = 3$	
	Slicing	Spline	Slicing	Spline
	$p = 20, n = 400$		$p = 20, n = 600$	
TPR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
FPR	0.04 (0.05)	0.01 (0.02)	0.02 (0.05)	0.04 (0.04)
	$p = 40, n = 800$		$p = 40, n = 1,200$	
TPR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
FPR	0.00 (0.00)	0.02 (0.02)	0.06 (0.04)	0.00 (0.01)

Table 3. Evaluation of dimension reduction basis estimation and variable selection for Examples 4.1 and 4.2 with correlated predictors.

	Example 4.1 with $d = 2$		Example 4.2 with $d = 3$	
	Slicing	Spline	Slicing	Spline
	$p = 40, n = 800$		$p = 40, n = 1,200$	
w/o penalty	0.77 (0.05)	0.73 (0.06)	0.77 (0.03)	0.74 (0.05)
SCAD	0.95 (0.04)	0.86 (0.13)	0.95 (0.03)	0.96 (0.03)
TPR	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
FPR	0.00 (0.01)	0.01 (0.01)	0.04 (0.04)	0.00 (0.01)

d is 3. We chose $\boldsymbol{\beta}_1 = (1, 1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_2 = (0, 0, 1, 1, 0, \ldots, 0)^T$, and $\boldsymbol{\beta}_3 = (0, 0, 0, 0, 1, 1, 0, \ldots, 0)^T$, where $n = 600$, $p = 20$ and $n = 1{,}200$, $p = 40$. Results of reduction basis estimation are reported in Table 1 (right half), and results of variable selection are reported in Table 2 (right half). Again, the proposed SDR estimator with the SCAD max penalty achieved a good performance in terms of both basis estimation and variable selection.

We next consider the performance with correlated predictors. Table 3 reports the results of reduction basis estimation and variable selection when $p = 40$. It is seen from the table that correlation among the predictors had some bearing on the method, but the overall performance resembled the results for the case without correlation: the penalized SDR estimator improved the estimation accuracy in terms of reduction basis estimation, and achieved a high true positive rate and a low false positive rate.

4.2. A data example

We briefly analyze the motif discovery data of Zhong et al. (2005) to illustrate the proposed method, though our analysis is by no means comprehensive. The goal here is to identify a subset of transcription factor binding motifs that affect the gene expression values. The response variable is the expression value obtained

by DNA microarray experiments, the predictors are the motif-matching scores of $\tilde{p} = 414$ candidate motifs, and the data consist of $n = 5{,}970$ genes as the sample observations. To bring the number of candidate predictors to the order of $\sqrt{n}$, we employed univariate regression for an initial screening, following the spirit of Fan and Lv (2008). We set the cutoff p-value at 0.05, and obtained $p = 118$ motifs for subsequent analysis. Zhong et al. (2005) suggested that the central subspace is two-dimensional and that the predictors affect the response in some nonlinear fashion. We applied our variable selection method to these data. The slicing transformation selected 16 motifs, whereas the spline transformation selected 9 motifs, that form a subset of the 16.

5. Discussion

There are a number of ways to extend this work. First, in our current development, we have treated the number of transformations h as fixed since it usually takes a pre-specified small value, and it helps simplify the technical derivations. For some particular transformation choices, a fixed h may result in an estimate of a proper subspace of the central subspace. As such it is of interest to extend our results to a diverging h. We speculate that the results of Corollary 1 would be modified accordingly, with the convergence rate of $h^{-1}\hat{\boldsymbol{B}}\hat{\boldsymbol{B}}^T$ at $O_p\big(\sqrt{\log(h)p/n}\big)$, while a rigorous conclusion needs more careful study. Second, the SDR estimators discussed in Section 2.1 rely on the first inverse moment $E(\boldsymbol{X}|Y)$. When $E(\boldsymbol{X}|Y) = 0$, the estimated subspace obtained may be a proper subspace of the central subspace. There have been proposals of SDR estimators that take advantage of the second or higher inverse moments, for instance, Cook and Weisberg (1991), Yin and Cook (2003), Li, Zha, and Chiaromonte (2005), and Li and Wang (2007). It is of interest to investigate the asymptotics of those SDR estimators with a diverging p. Finally, in many recent microarray and genetics studies, the number of predictors exceeds the number of observational units. Asymptotic properties of both dimension reduction and variable selection with $p > n$ remain to be explored. Full investigations of those extensions are to be our future research.

Acknowledgement

Wu is supported in part by NSF grant DMS-0905561, NIH grant R01-CA149569, and NCSU Faculty Research and Professional Development Award. Li is supported in part by NSF grant DMS-0706919.

Appendix A: Technical Conditions

(i) There are constants $a^* > 0$ and $C > 0$ such that, for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$,

$$P(\sum_{i\in I}(\boldsymbol{X}_i^T\boldsymbol{\beta})^2 > a^* n) \to 1 \text{ as } n \to \infty,$$

where $I = I(\boldsymbol{\beta}, C) = \{i = 1, \ldots, n : |\boldsymbol{X}_i^T\boldsymbol{\beta}| \le C\}$.

(ii) For any $\epsilon > 0$, there exists a constant $C > 0$ such that, for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| = 1$,

$$P\Big(\sum_{i\notin I}(\boldsymbol{X}_i^T\boldsymbol{\beta})^2 \le \epsilon n\Big) \to 1 \text{ as } n \to \infty.$$

(iii) There is a constant C such that $P(\max_{i=1,\ldots,n}\|\boldsymbol{X}_i\|^2 \le Cn^2) \to 1$ as $n \to \infty$.

(iv) $E(X_j^4) < C$ for some constant $C > 0$, $j = 1, \ldots, p$.

(v) $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$ is positive definite with all its eigenvalues bounded between $\underline{c}$ and $\overline{c}$, $0 < \underline{c} < \overline{c} < \infty$, for all $p = p_n$.

(vi) $E\{f(Y)\} = 0$ and $\mathrm{Var}\{f(Y) - \boldsymbol{X}^T\boldsymbol{\beta}^0\} < \infty$.

(vii) The eigenvalues of the pseudo-Fisher information matrix $I(\boldsymbol{\beta}^0)$ of $\boldsymbol{\beta}^0(f)$ are bounded for all $p = p_n$:

$$0 < \underline{\lambda} < \lambda_{\min}(I(\boldsymbol{\beta}^0)) \le \lambda_{\max}(I(\boldsymbol{\beta}^0)) < \overline{\lambda} < \infty \text{ for all } p = p_n,$$

where, up to a constant,

$$I(\boldsymbol{\beta}^0) = E\left([\boldsymbol{X}\{f(Y) - \boldsymbol{X}^T\boldsymbol{\beta}^0\}][\boldsymbol{X}\{f(Y) - \boldsymbol{X}^T\boldsymbol{\beta}^0\}]^T\right).$$

Remark 6. The regularity Conditions (i), (ii), and (iii) are simplified versions of Conditions X1, X2, and X3 of Portnoy (1984). Portnoy (1984) showed that these conditions hold in probability if $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ are *i.i.d.* according to a distribution satisfying his (4.3). As our Conditions (i), (ii), and (iii) are weaker, the same result applies.

Appendix B: Proofs

Proof of Theorem 1. Let $F(\boldsymbol{\alpha}) = \sum_{i=1}^n \boldsymbol{X}_i\{f(y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0 - \boldsymbol{X}_i^T\boldsymbol{\alpha}\}$ with $\boldsymbol{\alpha} \in \mathbb{R}^{p\times 1}$. Due to the convexity of the squared loss and the fact that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^0 + \hat{\boldsymbol{\alpha}}$, it suffices to show that there is a root $\hat{\boldsymbol{\alpha}}$ of $F(\boldsymbol{\alpha})$ satisfying $\|\hat{\boldsymbol{\alpha}}\|^2 = O_p(p/n)$. According to 6.3.4 of Ortega and Rheinboldt (1970), it in turn suffices to show that $\boldsymbol{\alpha}^T F(\boldsymbol{\alpha}) < 0$ for $\|\boldsymbol{\alpha}\|^2 = Bp/n$ for some $B > 0$. Toward that end, write $\boldsymbol{\alpha}^T F(\boldsymbol{\alpha}) = \sum_{i=1}^n \boldsymbol{X}_i^T\boldsymbol{\alpha}\{f(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0\} - \sum_{i=1}^n(\boldsymbol{X}_i^T\boldsymbol{\alpha})^2 \equiv A_1 - A_2$.

For A_2, we have $A_2 = \sum_{i=1}^n(\boldsymbol{X}_i^T\boldsymbol{\alpha})^2 \ge \|\boldsymbol{\alpha}\|^2 \inf_{\|\boldsymbol{\beta}\|=1}\sum_{i=1}^n(\boldsymbol{X}_i^T\boldsymbol{\beta})^2 \ge a^* n\|\boldsymbol{\alpha}\|^2$ in probability for some constant $a^* > 0$, due to Lemma 1.

For A_1, we have that $|A_1| \le \|\boldsymbol{\alpha}\| \| \sum_{i=1}^n \boldsymbol{X}_i\{f(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0\}\|$. Then,

$$\begin{aligned}
&E\| \sum_{i=1}^n \boldsymbol{X}_i\{f(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0\}\|^2 \\
&= E\left(\sum_{j=1}^p \left[\sum_{i=1}^n X_{ij}\{f(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0\} \right]^2 \right) \\
&= E\left(\sum_{j=1}^p \sum_{i=1}^n \sum_{i'=1}^n X_{ij}X_{i'j}\{f(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0\}\{f(Y_{i'}) - \boldsymbol{X}_{i'}^T\boldsymbol{\beta}^0\} \right) \\
&= \sum_{j=1}^p \sum_{i=1}^n E\left(X_{ij}^2\{f(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0\}^2 \right) + \sum_{j=1}^p \sum_{1\le i\ne i'\le n} E_{ij}E_{i'j} \\
&\le B'np \quad \text{for some } B' > 0,
\end{aligned}$$

where $E_{ij} = E[X_{ij}\{f(Y_i) - \boldsymbol{X}_i^T\boldsymbol{\beta}^0\}]$. The last inequality is true because $\boldsymbol{\beta}^0 = \operatorname{argmin} E\{f(Y) - \boldsymbol{X}^T\boldsymbol{\beta}\}^2$, which implies that $E(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta}^0 = E\{f(Y)\boldsymbol{X}\}$, and thus for any $1 \le j \le p$, $\sum_{m=1}^p E(X_jX_m)\beta_m^0 = E\{f(Y)X_j\}$, so $E_{ij} = 0$. Then by Chebychev's inequality, for any $\epsilon > 0$, there is a constant B^* such that $P\left\{A_1 \le B^*\sqrt{np}\|\boldsymbol{\alpha}\| \text{ for all } \boldsymbol{\alpha}\right\} \ge 1 - \epsilon$.

Combining the above two results, we have

$$P\left\{A_1 - A_2 \le B^*\sqrt{np}\|\boldsymbol{\alpha}\| - a^*n\|\boldsymbol{\alpha}\|^2 \text{ for all } \boldsymbol{\alpha}\right\} \ge 1 - 2\epsilon.$$

Set $B = (2B^*/a^*)^2$ and $\|\boldsymbol{\alpha}\|^2 = Bp/n$. Then we have

$$\begin{aligned}
&P\left\{\boldsymbol{\alpha}^T F(\boldsymbol{\alpha}) < 0 \text{ for all } \boldsymbol{\alpha} \text{ with } \|\boldsymbol{\alpha}\|^2 = B\frac{p}{n}\right\} \\
&\quad \ge P\left\{A_1 - A_2 \le -\frac{1}{2}Ba^*p \text{ for } \|\boldsymbol{\alpha}\|^2 = B'\frac{p}{n}\right\} \ge 1 - 2\epsilon.
\end{aligned}$$

Our desired result then follows from Ortega and Rheinboldt (1970).

Proof of Lemma 3. Let $\alpha_n = \sqrt{p_n}(n^{-1/2} + a_n)$. We need to show that for any $\epsilon > 0$ there exists a constant $C > 0$ such that

$$P\left\{\inf_{\|\boldsymbol{U}\|=C} Q(\boldsymbol{B} + \alpha_n\boldsymbol{U}) > Q(\boldsymbol{B})\right\} \ge 1 - \epsilon.$$

Note that

$$\begin{aligned} &Q(\boldsymbol{B}^0+\alpha_n\boldsymbol{U})-Q(\boldsymbol{B}^0)\\ &\quad\geq\sum_{k=1}^{h}\left\{L_k(\boldsymbol{\beta}_k^0+\alpha_n\boldsymbol{u}_k)-L_k(\boldsymbol{\beta}_k^0)\right\}\\ &\qquad+\sum_{j=1}^{q}\left(p_\lambda(\max_{1\leq k\leq h}|\beta_{jk}^0+\alpha_n u_{jk}|)-p_\lambda(\max_{1\leq k\leq h}|\beta_{jk}^0|)\right)\\ &\quad\equiv D_1+D_2. \end{aligned}$$

We decompose D_1 and D_2, respectively, as $D_1 = D_{11} + D_{12}$ and $D_2 = D_{21} + D_{22}$, where

$$D_{11}=\alpha_n\sum_{k=1}^{h}\boldsymbol{u}_k^T\frac{\partial}{\partial\boldsymbol{\beta}}L_k(\boldsymbol{\beta}_k^0),$$

$$D_{12}=\frac{1}{2}\alpha_n^2\sum_{k=1}^{h}\boldsymbol{u}_k^T\left\{\frac{\partial^2}{\partial\boldsymbol{\beta}^2}L_k(\boldsymbol{\beta}_k^0)\right\}\boldsymbol{u}_k,$$

$$D_{21}=\sum_{j=1}^{q}p_\lambda'(\max_{1\leq k\leq h}|\beta_{jk}^0|)(\max_{1\leq k\leq h}|\beta_{jk}^0+\alpha_n u_{jk}|-\max_{1\leq k\leq h}|\beta_{jk}^0|),$$

$$D_{22}=\sum_{j=1}^{q}\frac{1}{2}p_\lambda''(\max_{1\leq k\leq h}|\beta_{jk}^0|)(\max_{1\leq k\leq h}|\beta_{jk}^0+\alpha_n u_{jk}|-\max_{1\leq k\leq h}|\beta_{jk}^0|)^2(1+o(1)).$$

For D_{11}, by Condition (vii), the eigenvalues of the pseudo-Fisher Information matrix $I(\boldsymbol{\beta}_k^0)$ are bounded away from both zero and infinity. Therefore we have

$$\|\frac{\partial}{\partial\boldsymbol{\beta}}L_k(\boldsymbol{\beta}_k^0)\|=O_p(\sqrt{\frac{p}{n}}). \tag{B.1}$$

Then

$$\begin{aligned} |D_{11}|&\leq\alpha_n\sum_{k=1}^{h}|\boldsymbol{u}_k^T\frac{\partial}{\partial\boldsymbol{\beta}}L_k(\boldsymbol{\beta}_k^0)|\leq\alpha_n\sum_{k=1}^{h}\|\boldsymbol{u}_k\|\cdot\|\frac{\partial}{\partial\boldsymbol{\beta}}L_k(\boldsymbol{\beta}_k^0)\|\\ &=O_p(\alpha_n\sqrt{\frac{p}{n}})\sum_{k=1}^{h}\|\boldsymbol{u}_k\|=O_p(\alpha_n^2)\sum_{k=1}^{h}\|\boldsymbol{u}_k\|. \end{aligned}$$

For D_{12}, we note that

$$\begin{aligned} D_{12}&=\frac{1}{2}\alpha_n^2\sum_{k=1}^{h}\boldsymbol{u}_k^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)\boldsymbol{u}_k\\ &=\frac{1}{2}\alpha_n^2\sum_{k=1}^{h}\boldsymbol{u}_k^T\left\{\frac{1}{n}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T\right)-\boldsymbol{\Sigma}\right\}\boldsymbol{u}_k+\frac{1}{2}\alpha_n^2\sum_{k=1}^{h}\boldsymbol{u}_k^T\boldsymbol{\Sigma}\boldsymbol{u}_k. \end{aligned}$$

As in the proof of Lemma 8 of Fan and Peng (2004), by Chebyshev's inequality, for any $\epsilon > 0$ we have

$$P\left(\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T - \boldsymbol{\Sigma}\| \geq \frac{\epsilon}{p}\right) \leq \frac{p^2}{n^2\epsilon^2}E\sum_{j,m=1}^{p}(X_{ij}X_{im} - \sigma_{ij})^2 = O(\frac{p^4}{n}) = o(1),$$

where σ_{ij} is the (i,j)-element of $\boldsymbol{\Sigma}$. Thus we can write

$$D_{12} = o_p(1)\frac{1}{2}\alpha_n^2\sum_{k=1}^{h}\|\boldsymbol{u}_k\|^2 + \frac{1}{2}\alpha_n^2\sum_{k=1}^{h}\boldsymbol{u}_k^T\boldsymbol{\Sigma}\boldsymbol{u}_k.$$

We have

$$\begin{aligned}
|D_{21}| &\leq \sum_{j=1}^{q}a_n\sum_{k=1}^{h}\alpha_n|u_{jk}| \leq a_n\alpha_n\sqrt{q}\sum_{k=1}^{h}\|\boldsymbol{u}_k\|,\\
|D_{22}| &\leq \sum_{j=1}^{q}\frac{1}{2}p_\lambda''(\max_{1\leq k\leq h}|\beta_{jk}^0|)\left(\sum_{k=1}^{h}\alpha_n|u_{jk}|\right)^2(1+o(1))\\
&\leq \frac{1}{2}\max_{j=1}^{q}p_\lambda''(\max_{1\leq k\leq h}|\beta_{jk}^0|)\alpha_n^2 h\sum_{k=1}^{h}\|\boldsymbol{u}_k\|^2.
\end{aligned}$$

Combining the above results, D_{12} is asymptotically positive and dominates other terms. Setting $C = \|\boldsymbol{U}\| = (\sum_{k=1}^{h}\|\boldsymbol{u}_k\|^2)^{1/2}$ large enough, the desired result follows.

Proof of Lemma 4. It suffices to show that with probability tending to one as $n \to \infty$, for any given $\{\boldsymbol{\beta}_{k+}, k = 1,\ldots,h\}$ satisfying $\|\boldsymbol{\beta}_{k+} - \boldsymbol{\beta}_{k+}^0\| = O_p(\sqrt{p/n})$ and any constant C, for $j = q+1,\ldots,p$,

$$\begin{aligned}
&\frac{\partial}{\partial\beta_{jk}^r}Q(\boldsymbol{B}) < 0 \text{ for } 0 < \beta_{jk} < C\sqrt{\frac{p}{n}} \text{ and } \beta_{jk} = \max_{m=1}^{h}|\beta_{jm}|,\\
&\frac{\partial}{\partial\beta_{jk}^l}Q(\boldsymbol{B}) > 0 \text{ for } -C\sqrt{\frac{p}{n}} < \beta_{jk} < 0 \text{ and } \beta_{jk} = -\max_{m=1}^{h}|\beta_{jm}|,
\end{aligned}$$

where $\partial/\partial\beta_{jk}^l$ and $\partial/\partial\beta_{jk}^r$ denote the left and right hand partial derivative, respectively.

By a Taylor expansion,

$$\frac{\partial}{\partial\beta_{jk}}L_k(\boldsymbol{\beta}_k) = \frac{\partial}{\partial\beta_{jk}}L_k(\boldsymbol{\beta}_k^0) + \sum_{l=1}^{p}\frac{\partial^2}{\partial\beta_{jk}\partial\beta_{lk}}L_k(\boldsymbol{\beta}_k^0)(\beta_{lk} - \beta_{lk}^0) \equiv E_1 + E_2.$$

Due to (B.1), we have $E_1 = O_p(\sqrt{p/n})$. Next we decompose E_2 as

$$E_2 = \sum_{l=1}^{p} \left[\frac{\partial^2}{\partial\beta_{jk}\partial\beta_{lk}} L_k(\boldsymbol{\beta}_k^0) - E\left\{ \frac{\partial^2}{\partial\beta_{jk}\partial\beta_{lk}} L_k(\boldsymbol{\beta}_k^0) \right\} \right] (\beta_{lk} - \beta_{lk}^0)$$
$$+ \sum_{l=1}^{p} E\left\{ \frac{\partial^2}{\partial\beta_{jk}\partial\beta_{lk}} L_k(\boldsymbol{\beta}_k^0) \right\} (\beta_{lk} - \beta_{lk}^0) \equiv E_{21} + E_{22}.$$

For E_{21}, we have

$$E_{21} \le \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^0\| \left(\sum_{l=1}^{p} \left[\frac{\partial^2}{\partial\beta_{jk}\partial\beta_{lk}} L_k(\boldsymbol{\beta}_k^0) - E\left\{ \frac{\partial^2}{\partial\beta_{jk}\partial\beta_{lk}} L_k(\boldsymbol{\beta}_k^0) \right\} \right]^2 \right)^{1/2}$$
$$= \|\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^0\| \left(\sum_{l=1}^{p} \left[\frac{1}{n} \sum_{i=1}^{n} X_{ij}X_{il} - E(X_jX_l) \right]^2 \right)^{1/2}$$
$$= O_p(\sqrt{\frac{p}{n}}) O_p(\sqrt{\frac{p}{n}}) = O_p(\frac{p}{n}),$$

where the second to last equality comes from the moment Condition (iv) and the fact that all the eigenvalues of $\boldsymbol{\Sigma}$ are bounded away from both 0 and ∞ by Condition (v).

For E_{22}, by Cauchy-Schwarz inequality and $\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^0\| = O_p(\sqrt{p/n})$,

$$E_{22} \le \left| \sum_{l=1}^{q} E(X_{ij}X_{il})(\beta_{jk} - \beta_{jl}) \right| = O_p(\sqrt{\frac{p}{n}}) \left(\sum_{l=1}^{p} \sigma_{jl} \right)^{1/2} = O_p(\sqrt{\frac{p}{n}}),$$

where σ_{ij} is the (i,j)-element of $\boldsymbol{\Sigma}$, and the last equality comes from the fact that $\sum_{l=1}^{p} \sigma_{jl} = O(1)$ which is ensured by Conditions (iv) and (v).

Combining the above two results, we have $\frac{\partial}{\partial\beta_{jk}} L_k(\boldsymbol{\beta}_k) = O_p(\sqrt{p/n})$.

Finally, note that $\sqrt{p/n}/\lambda_n \to 0$ and $\liminf_{n\to\infty} \inf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0$. When $|\beta_{jk}| = \max_{m=1}^{h} |\beta_{jm}|$, we have

$$\frac{\partial Q(\boldsymbol{B})}{\beta_{jk}^r} = \lambda_n \left\{ \frac{p'_{\lambda_n}(\max_{m=1}^{h} |\beta_{jm}|)}{\lambda_n} + O_p\Big(\frac{\sqrt{p/n}}{\lambda_n}\Big) \right\} \text{ if } \beta_{jk} > 0,$$
$$\frac{\partial Q(\boldsymbol{B})}{\beta_{jk}^l} = \lambda_n \left\{ -\frac{p'_{\lambda_n}(\max_{m=1}^{h} |\beta_{jm}|)}{\lambda_n} + O_p\Big(\frac{\sqrt{p/n}}{\lambda_n}\Big) \right\} \text{ if } \beta_{jk} < 0.$$

In both cases, the first term dominates the second. Thus the result of Lemma 4 follows.

References

An, L. T. H. and Tao, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. *J. Global Optimization* **11**, 253-285.

Bondell, H. D. and Li, L. (2009). Shrinkage inverse regression estimation for model free variable selection. *J. Roy. Statist. Soc. Ser. B* **71**, 287-299.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373-384.

Cook, R. D. (1996). Graphics for regressions with a binary response. *J. Amer. Statist. Assoc.* **91**, 983-992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. Wiley, New York.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist.* **32**, 1062-1092.

Cook, R. D. and Ni, L. (2006). Using intra-slice covariances for improved estimation of the central subspace in regression. *Biometrika* **93**, 65-74.

Cook, R. D. and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction," by K.C. Li *J. Amer. Statist. Assoc.* **86**, 328-332.

Cook, R. D. and Yin, X. (2001). Dimension-reduction and visualization in discriminant analysis. *Austral. N. Z. J. Statist.* **43**, 147-200.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Fan, J. and Li, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30**, 74-99.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. Roy. Statist. Soc. Ser. B* **70**, 849-911.

Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Fung, W. K., He, X., Liu, L. and Shi, P. D. (2002). Dimension reduction based on canonical correlation. *Statist. Sinica* **12**, 1093-1114.

Golub, G. and van Loan, C. (1996). *Matrix Computations*, 3rd ed. The Johns Hopkins University Press.

Hall, P. and Li, K. C. (1993). On almost linearity of low dimensional projection from high dimensional data. *Ann. Statist.* **21**, 867-889.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321-377.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20**, 1040-1061.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33**, 1617-1642.

Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators, *Ann. Statist.* **28**, 1356-1378.

Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc.* **102**, 997-1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Statist.* **36**, 261-286.

Ni, L., Cook, R. D. and Tsai, C. L. (2005). A note on shrinkage sliced inverse regression. *Biometrika* **92**, 242-247.

Ni, L., Wang, H., Tsai, C. L. and Zhou, J. (2008). Model free variable selection via adaptive lasso. *Proceedings of the Joint Statistical Meetings*.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Wiley, New York.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12**, 1298-1309.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wang, H., Li, R. and Tsai, C. L (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *J. Amer. Statist. Assoc.* **103**, 811-821.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statist. Sinica* **19**, 801-817.

Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *Ann. Statist.* **35**, 2654-2690.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional k-th moment in regression. *J. Roy. Statist. Soc. Ser. B* **64**, 159-176.

Yin, X. and Cook, R. D. (2003). Estimating central subspaces via inverse third moments. *Biometrika* **90**, 113-125.

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

Zhong, W., Zeng, P., Ma, P., Liu, J. S. and Zhu, Y. (2005). RSIR: Regularized Sliced Inverse Regression for Motif Discovery. *Bioinformatics* **21**, 4169-4175.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36**, 1649-1668.

Zhu, L. P. and Zhu, L. X. (2009a). On distribution-weighted partial least squares with diverging number of highly correlated predictors. *J. Roy. Statist. Soc. Ser. B* **71**, 525-548.

Zhu, L. P. and Zhu, L. X. (2009b). Nonconcave penalized inverse regression in single-index models with high dimensional predictors. *J. Multivariate Anal.* **100**, 862-875.

Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *Ann. Statist.* **24**, 1053-1068.

Zhu, L. X., Miao, B. and Peng, H. (2006). On sliced inverse regression with high dimensional covariates. *J. Amer. Statist. Assoc.* **101**, 630-643.

Zhu, L. X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5**, 727-736.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1533.

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
E-mail: wu@stat.ncsu.edu

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA.
E-mail: li@stat.ncsu.edu

(Received April 2009; accepted November 2009)