Estimating a sparse reduction for general regression in high dimensions

Tao Wang, Hongyu Zhao,
Department of Biostatistics, Yale University
Mengjie Chen
Department of Biostatistics, University of North Carolina
and
Lixing Zhu
Department of Mathematics, Hong Kong Baptist University*

September 30, 2015

Abstract

Although the concept of sufficient dimension reduction has been proposed for a long time, studies in the literature have largely focused on properties of estimators of dimension-reduction subspaces in the classical "small p, and large n" setting. Rather than the subspace, this paper considers directly the set of reduced predictors, which we believe are more relevant for subsequent analyses, and proposes a principled method for estimating a sparse reduction, which is based on a new representation of a well-known method called sliced inverse regression. A fast and efficient algorithm is developed for computing the estimator. The asymptotic behavior of the new method is studied when the number of predictors, p, exceeds the sample size, n, providing a guide for choosing the number of sufficient dimension-reduction predictors. Numerical results, including a simulation study and a cancer-drug sensitivity data analysis, are presented to examine the performance.

Keywords: Inverse modeling; Model-free dimension reduction; Sparsity

*The corresponding author (E-mail: lzhu@hkbu.edu.hk).

1 Introduction

Advances in information technology are creating the opportunity of new and exciting discoveries. At the same time statisticians are nowadays faced with new challenges due to the ubiquitous availability of complex and high-dimensional data in various scientific fields, ranging from genetics and genomics to finance and economics. How to extract useful information from these data is an important focus of contemporary statistical research and practice. This paper is concerned with method and theory for general regression in high dimensions, where the number of predictors, p, is larger than the number of samples, n. Suppose that $Y \in \mathbb{R}$ is a univariate response and $\boldsymbol{X} = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ is a vector of predictors. Our goal is to capture all the regression information about Y that is available from $\boldsymbol{X}$, based on a random sample $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ of $(\boldsymbol{X}, Y)$.

Dimension reduction or feature extraction is a popular way of coping with high dimensionality. Classical methods for dimension reduction in regression include principal component regression and partial least squares. While the loadings in principal component regression depend solely on the covariance matrix of $\boldsymbol{X}$, the weights in partial least squares use the information from both $\boldsymbol{X}$ and Y variables. In the same spirit, sufficient dimension reduction (Li 1991, Cook 1994) is a body of methods which utilize response information in achieving dimension reduction. It aims to reduce the dimension of $\boldsymbol{X}$ without loss of information on the regression of Y on $\boldsymbol{X}$, and without requiring a pre-specified parametric model. In this paper, we concentrate on the framework of sufficient dimension reduction as described below.

Let $\mathbb{S}$ denote a subspace of $\mathbb{R}^p$, and let $\mathbf{P}_{\mathbb{S}}$ denote the orthogonal projection onto $\mathbb{S}$ with respect to the usual inner product. If Y and $\boldsymbol{X}$ are independent conditioned on $\mathbf{P}_{\mathbb{S}}\boldsymbol{X}$, then we say that $\mathbb{S}$ is a dimension-reduction subspace. The intersection of all such subspaces, if itself satisfies the conditional independence, is defined to be the central subspace and is denoted by $\mathbb{S}_{Y|\boldsymbol{X}}$ (Cook 1998, Yin et al. 2008). Let $d_{Y|\boldsymbol{X}} = \dim(\mathbb{S}_{Y|\boldsymbol{X}})$. We call $d_{Y|\boldsymbol{X}}$ the structural dimension of the regression of Y on $\boldsymbol{X}$. Formally, sufficient dimension reduction requires that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \mathbf{P}_{\mathbb{S}_{Y|\boldsymbol{X}}} \boldsymbol{X}, \tag{1}$$

where $\perp\!\!\!\perp$ indicates independence. Being nearly model-free, the above statement provides a very general framework for dimension reduction in regression. Popular methods for estimating $\mathbb{S}_{Y|\boldsymbol{X}}$ include sliced inverse regression (Li 1991), sliced average variance estimation (Cook & Weisberg 1991), directional regression (Li & Wang 2007), discretization-expectation estimation (Zhu et al. 2010), among others. Please see Yin (2010) and Ma & Zhu (2013) for recent reviews.

Sufficient dimension reduction is often used as the first step in statistical analysis (Cook 1998, Li 2000). It permits us to restrict attention to a number of new predictors, expressed as linear combinations of the original ones: $\boldsymbol{\beta}_1^\top \boldsymbol{X}, \ldots, \boldsymbol{\beta}_{d_{Y|\boldsymbol{X}}}^\top \boldsymbol{X}$, where $\{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d_{Y|\boldsymbol{X}}}\}$ is a basis of $\mathbb{S}_{Y|\boldsymbol{X}}$. Subsequent modeling and prediction can be built upon these predictors. In this sense, the inference object more relevant to subsequent data analysis is not the subspace $\mathcal{S}_{Y|\boldsymbol{X}}$ but the reduction

$$\mathcal{R}(\boldsymbol{X}) = \{\boldsymbol{\beta}_1^\top \boldsymbol{X}, \ldots, \boldsymbol{\beta}_{d_{Y|\boldsymbol{X}}}^\top \boldsymbol{X}\}$$

itself. Let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{d_{Y|\boldsymbol{X}}})$. At the sample level, the goal is then to estimate $\mathbf{X}\boldsymbol{\beta}$. After that, we can apply more sophisticated, often nonparametric, methods on the reduced set of predictors to model the relationship with the response variable. Surprisingly, studies of sufficient dimension reduction have largely focused on the problem of estimating the central subspace or its variants. When p is fixed or diverges slowly with n, this makes sense because the difference between these two estimation problems is minor, that is, the behavior of an estimator $\hat{\boldsymbol{\beta}}$ and that of $\mathbf{X}\hat{\boldsymbol{\beta}}$ are similar. The situation is different, however, when p is comparable to or even larger than n, unless we have available a good estimator of $\boldsymbol{\beta}$. Unfortunately, it remains very difficult to obtain even a consistent estimator of $\boldsymbol{\beta}$ when $p > n$, as we indicate below.

To handle high dimensionality, a wise alternative to dimension reduction is variable selection or feature selection. In many statistical applications, it is often the case that a relatively small number of predictors have substantial explanatory power while the rest are redundant. By effectively identifying the subset of relevant predictors, variable selection can improve estimation accuracy and enhance model interpretability. For parametric regression, regularization procedures, which impose constraints represented by a penalty function, have proven very successful. Popular methods include the LASSO (Tibshirani 1996), the SCAD

(Fan & Li 2001), and the elastic net (Zou & Hastie 2005), just to name a few. In the framework of sufficient dimension reduction, variable selection has attracted considerable attention (Cook 2004, Yin & Hilafu 2015). In formal terms, it aims to find a parsimonious index set $\mathcal{A} \subseteq \{1, \ldots, p\}$ such that

$$Y \perp\!\!\!\perp \boldsymbol{X} \,|\, \boldsymbol{X}_{\mathcal{A}}, \tag{2}$$

where $\boldsymbol{X}_{\mathcal{A}} = (X_j, j \in \mathcal{A})$. A number of methods have been proposed under this framework, some of which integrate sufficient dimension-reduction techniques with the regularization paradigm. See, for example, Bondell & Li (2009), Chen et al. (2010), and Wu & Li (2011).

The integration of model-free dimension reduction and variable selection is attractive as the two techniques can be mutually enhanced. However, most methods in the literature rely on n being large relative to p, because they require the inversion of the sample covariance matrix of $\boldsymbol{X}$. To partly circumvent the problem, Li & Li (2004) performed dimension reduction in two stages, first replacing the p predictors with a few principal components and then applying a dimension-reduction method to the regression with the selected components; Zhong et al. (2005) proposed a regularized sliced inverse regression method by replacing the sample covariance matrix by a perturbed version that is well-defined; and Cook et al. (2007) developed a novel method that avoids the computation of inverses and reduces to partial least squares in a special case. However, these methods are expected to be inconsistent unless $p/n \to 0$ (Chen et al. 2010), and they may not work well when the regression is sparse. Using the nonlinear least squares formulation of sliced inverse regression originated by Cook (2004), Li & Yin (2008) developed a double regularized version that achieves dimension reduction and then predictor selection. However, their method requires a heavy computational burden, especially in high dimensions, and its asymptotic properties are unknown. Based on the generalized eigenvalue problem associated with sliced inverse regression, Yu et al. (2013) recently proposed to sequentially estimate directions of the central subspace. This seems to be a promising method, but it unrealistically requires the structural dimension to be known, and no clue is given as to how to actually determine the reduced number of predictors. To our knowledge, there are no methods for consistently estimating of $\mathcal{S}_{Y|\boldsymbol{X}}$ ($d_{Y|\boldsymbol{X}}$ and $\boldsymbol{\beta}$) in the large p and small n setting.

The focus of this paper is thus to estimate the reduction $\mathbf{X}\boldsymbol{\beta}$ directly. That is, we

consider the reduction $\mathcal{R}(\boldsymbol{X})$ rather than the central subspace $\mathbb{S}_{Y|\boldsymbol{X}}$. By introducing a novel viewpoint for addressing sliced inverse regression, we develop a principled method for estimating a sparse reduction. Theoretically we show how the rate of convergence depends on the parameters that control the complexity of the problem for $p > n$ regressions. Interestingly, this implies that the cross-validation procedure is valid for determination of the structural dimension (Jiang & Liu 2014).

The organization of this paper is as follows. In Subsection 2.2, we give a brief review of sliced inverse regression. In Subsection 2.3, we represent sliced inverse regression in a dimension-redundant way. We then propose a method for estimating a sparse reduction in Subsection 2.4. In addition, we develop an efficient algorithm for numerical implementation in Subsection 2.5 and study the asymptotic behavior of the new approach in Subsection 2.6. In Sections 3 and 4, we evaluate the performance of the proposed method through a simulation study and a cancer-drug sensitivity data analysis. The proofs are given in Section 5. All lemmas are available in a supplementary article (Lemmas 1-11, (Wang et al. 2015)).

2 Methodology

2.1 Notation

For the following development, we introduce some notation. Let $\mathbf{M} \in \mathbb{R}^{I\times J}$ be an $I \times J$ matrix. Given $\mathcal{I} \subseteq \{1,\ldots,I\}$ and $\mathcal{J} \subseteq \{1,\ldots,J\}$, we denote $\mathbf{M}_{\mathcal{I}\mathcal{J}}$ to be the sub-matrix of $\mathbf{M}$ whose rows and columns are indexed by $\mathcal{I}$ and $\mathcal{J}$, respectively. Let $\mathbf{M}_{\mathcal{I}*}$ be the sub-matrix of $\mathbf{M}$ with rows in $\mathcal{I}$ and $\mathbf{M}_{*\mathcal{J}}$ the sub-matrix with columns in $\mathcal{J}$. We denote $\|\mathbf{M}\|_F$ to be the Frobenius norm of $\mathbf{M}$. Let $\sigma_k(\mathbf{M})$ be the k-th largest singular value of $\mathbf{M}$, for $k = 1,\ldots,\min(I,J)$. If $I = J$, we let $\lambda_j(\mathbf{M})$ be the j-th largest eigenvalue of $\mathbf{M}$, for $j = 1,\ldots,J$. For any two matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ of the same size, we denote the inner product of $\mathbf{M}_1$ and $\mathbf{M}_2$ by $\langle \mathbf{M}_1, \mathbf{M}_2\rangle = \mathrm{trace}(\mathbf{M}_1^\top \mathbf{M}_2)$. Let $\boldsymbol{v} \in \mathbb{R}^J$ be a J-dimensional vector. We denote $\boldsymbol{v}_{\mathcal{J}}$ to be the sub-vector of $\boldsymbol{v}$ whose entries are indexed by $\mathcal{J}$. For any two vectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of the same length, we denote the inner product of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ by $\langle \boldsymbol{v}_1, \boldsymbol{v}_2\rangle = \boldsymbol{v}_1^\top \boldsymbol{v}_2$. For $r \geq 1$, let $\|\boldsymbol{v}\|_r$ denote the usual L_r norm of $\boldsymbol{v}$. Write

$\mathbf{M} = (\boldsymbol{m}_1, \ldots, \boldsymbol{m}_I)^\top$, where $\boldsymbol{m}_i \in \mathbb{R}^J$ for $i = 1, \ldots, I$. Define $\|\mathbf{M}\|_{2,1} = \sum_{i=1}^I \|\boldsymbol{m}_i\|_2$. For any positive integer J and any subset $\mathcal{J} \subseteq \{1, \ldots, J\}$, we let $|\mathcal{J}|$ denote the cardinality of $\mathcal{J}$ and $\mathcal{J}^c$ the complement of $\mathcal{J}$. Finally, we denote $\mathbf{I}_J$ to be the $J \times J$ identity matrix and $\mathbf{P}_J$ the centering matrix of size J.

2.2 A short review of sliced inverse regression

Our method is based on sliced inverse regression (Li 1991). Suppose $\boldsymbol{X}$ has mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$. Assume without loss of generality that $\boldsymbol{\mu}$ is zero. Sliced inverse regression uses the fact that the inverse regression curve $E(\boldsymbol{X}|Y = y)$ is contained in $\boldsymbol{\Sigma}\mathbb{S}_{Y|\boldsymbol{X}}$ for all y, if a linearity condition on the distribution of $\boldsymbol{X}$ is satisfied. This condition states that the conditional mean $E(\boldsymbol{X}|\mathbf{B}^\top\boldsymbol{X})$ is linear in $\mathbf{B}^\top\boldsymbol{X}$, where $\mathbf{B} \in \mathbb{R}^{p\times d_{Y|\boldsymbol{X}}}$ is any basis matrix for $\mathbb{S}_{Y|\boldsymbol{X}}$.

To allow relatively easy estimation of $\mathbb{S}_{Y|\boldsymbol{X}}$, the slicing technique, which constructs a finite support by partitioning the range of Y into a few slices, is a standard practice, especially with a continuous response. Let $T(Y)$ denote the sliced version of Y. The rationale is that one has $\mathbb{S}_{T(Y)|\boldsymbol{X}} \subseteq \mathbb{S}_{Y|\boldsymbol{X}}$ with equality when the slicing is fine enough. For simplicity we assume that Y takes values in the set $\{1, \ldots, G\}$, and that $\mathbb{S}_{Y|\boldsymbol{X}}$ is contained in the subspace spanned by $\{\boldsymbol{\Sigma}^{-1}E(\boldsymbol{X}|Y = g)\}_{g=1}^G$.

We note that most methods in the literature that are potentially applicable in situations where p is comparable to or larger than n, are developed upon sliced inverse regression, or more generally upon the inverse modeling in which the predictors in $\boldsymbol{X}$ are regressed on the response Y. One benefit here is the computational feasibility and simplicity. This is very important, because both operational and theoretical issues may occur for other types of dimension-reduction methods in high dimensions as they involve non-parametric estimations.

2.3 A dimension-redundant representation

In order to estimate $\mathbf{X}\boldsymbol{\beta}$ in the $p > n$ setting, we propose to formulate sliced inverse regression in a dimension-redundant way which turns out to be very successful. The basic idea is simple and can be described as follows. First, we know that $\{\boldsymbol{\Sigma}^{-1}E(\boldsymbol{X}|Y = g)\}_{g=1}^G$

and $[\boldsymbol{\Sigma}^{-1}E\{\boldsymbol{X}I(Y=g)\}]_{g=1}^{G}$ span the same subspace. Second, we have available a least squares structure, namely,

$$\mathbf{B}^0 = \boldsymbol{\Sigma}^{-1}[E\{\boldsymbol{X}I(Y=1)\},\ldots,E\{\boldsymbol{X}I(Y=G)\}]$$

is the $p\times G$ matrix of coefficients from the regression of $\{I(Y=g)\}_{g=1}^{G}$ on $\boldsymbol{X}$. Third, given that G is generally larger than $d_{Y|\boldsymbol{X}}$, the matrix $\mathbf{B}_0$ is rank deficient. As a consequence, sliced inverse regression can be formally recast as a reduced-rank regression problem (Izenman 1975), with low structural dimension linking to low rank.

To the best of our knowledge, we are the first to use this naive yet intrinsic finding whose implication for handling $p > n$, as seen below, is largely neglected by the sufficient dimension-reduction community. Note however that, unlike the standard reduced-rank regression in which we have an additive error, we work under the conditional independence setting, (1) and (2), which is model-free or error-free. This is an essential difference, making inevitably the theoretical investigation very challenging.

2.4 Estimating a sparse reduction

Note that $\mathbf{B}^0$ contains all information on dimension reduction, that is, $Y \perp\!\!\!\perp \boldsymbol{X}|\mathbf{B}_0^\top\boldsymbol{X}$. Throughout the paper, we assume that $G > d_{Y|\boldsymbol{X}}$ and that

$$Y \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{X}_{\mathcal{A}_0}$$

for some $\mathcal{A}_0 \subseteq \{1,\ldots,p\}$, with $d_{Y|\boldsymbol{X}} \le |\mathcal{A}_0| \ll p$. Then $\mathbf{B}^0$ is both rank deficient and row sparse. Our procedure for estimating a sparse reduction is based on a rank factorization of an estimate of $\mathbf{B}^0$.

To be specific, let $\mathbf{L} = (L_{ig}) \in \mathbb{R}^{n\times G}$ be the slice indicator matrix, where $L_{ig} = I(y_i = g)$ for $i = 1,\ldots,n$ and $g = 1,\ldots,G$. For each $1 \le d \le G < n$, define

$$\mathfrak{B}_d = \{\mathbf{B} \in \mathbb{R}^{p\times G} : \mathrm{rank}(\mathbf{B}) \le d\}.$$

We now consider the criterion

$$Q_\lambda(\mathbf{B}) = \|\mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}\|_F^2 + 2\lambda\|\mathbf{B}\|_{2,1}, \tag{3}$$

where $\tilde{\mathbf{X}} = \mathbf{P}_n\mathbf{X}$ denotes the centered matrix and $\lambda > 0$ is the regularization parameter. We let $\hat{\mathbf{B}} = \hat{\mathbf{B}}_d$ be a minimizer of $Q_\lambda(\mathbf{B})$ over $\mathfrak{B}_d$.

To see how our method works, note that for each $\mathbf{B} \in \mathfrak{B}_d$, we can write $\mathbf{B} = \mathbf{M}\mathbf{N}^\top$, where $\mathbf{M}$ is of dimension $p \times d$ and $\mathbf{N}$ is a $G \times d$ orthogonal matrix. Since $\|\mathbf{B}\|_{2,1} = \|\mathbf{M}\|_{2,1}$, it is convenient to re-express (3) in the equivalent form

$$\begin{aligned} \underset{\mathbf{M}\in\mathbb{R}^{p\times d},\mathbf{N}\in\mathbb{R}^{G\times d}}{\text{minimize}} \quad & \|\mathbf{L} - \tilde{\mathbf{X}}\mathbf{M}\mathbf{N}^\top\|_F^2 + 2\lambda\|\mathbf{M}\|_{2,1} \\ \text{subject to} \quad & \mathbf{N}^\top\mathbf{N} = \mathbf{I}_d. \end{aligned} \tag{4}$$

It is now clear that our method produces a sparse reduction estimate, $\tilde{\mathbf{X}}\mathbf{M}$, when the reduced dimensionality is d.

2.5 Computation

To solve the minimization problem, we can iteratively optimize with respect to $\mathbf{M}$ and $\mathbf{N}$. When $\mathbf{N}$ is fixed, minimizing (4) with respect to $\mathbf{M}$ reduces to a group LASSO optimization (Yuan & Lin 2006):

$$\underset{\mathbf{M}\in\mathbb{R}^{p\times d}}{\text{minimize}} \quad \|\mathbf{L} - \tilde{\mathbf{X}}\mathbf{M}\mathbf{N}^\top\|_F^2 + 2\lambda\|\mathbf{M}\|_{2,1}. \tag{5}$$

The solution for $\mathbf{M}$ can thus be efficiently calculated using block-wise descent algorithms (Breheny & Huang 2015). When $\mathbf{M}$ is fixed, minimizing (4) with respect to $\mathbf{N}$ becomes an orthogonal Procrustes problem

$$\begin{aligned} \underset{\mathbf{N}\in\mathbb{R}^{G\times d}}{\text{minimize}} \quad & \|\mathbf{L} - \tilde{\mathbf{X}}\mathbf{M}\mathbf{N}^\top\|_F^2 \\ \text{subject to} \quad & \mathbf{N}^\top\mathbf{N} = \mathbf{I}_d. \end{aligned} \tag{6}$$

Let $\mathbf{L}^\top\tilde{\mathbf{X}}\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{L}^\top\tilde{\mathbf{X}}\mathbf{M}$; that is, $\mathbf{U}$ is $G \times d$ with orthonormal columns, $\mathbf{V}$ is $d \times d$ orthogonal, and $\mathbf{D}$ is $d \times d$ diagonal. The solution for $\mathbf{N}$ is $\mathbf{U}\mathbf{V}^\top$.

In all the empirical work in this paper, the values of the parameters, d and λ, are selected by cross validation. Theoretical results, to be discussed below, indicate that cross validation is a good strategy. We note that a similar algorithm can be found in Chen & Huang (2012) and Bunea et al. (2012) who proposed methods for sparse multivariate linear regression model.

2.6 Theoretical properties

Assume that $\boldsymbol{X}$ has an elliptical distribution. Let $\boldsymbol{v} \in \mathbb{R}^p$. Assume further that the distribution of $\boldsymbol{Z} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ is sub-Gaussian in the sense that there exists some $\rho > 0$ such that

$$E\{\exp(t\langle \boldsymbol{v}, \boldsymbol{Z}\rangle)\} \leq \exp(\rho^2 t^2/2) \quad \text{for all } t > 0 \text{ and } \|\boldsymbol{v}\|_2 = 1.$$

We call $\boldsymbol{Z}$ sub-Gaussian with sub-Gaussian constant ρ.

For some integer $1 \leq q \leq p$, let $\mathfrak{A}_q = \{\mathcal{S} \subseteq \{1, \ldots, p\} : |\mathcal{S}| \leq q\}$. Let m be a positive integer and $\mathfrak{M}_{\mathcal{S},m} = \{\mathbf{M} \in \mathbb{R}^{p\times m} : \|\mathbf{M}_{\mathcal{S}^c *}\|_{2,1} \leq 2\|\mathbf{M}_{\mathcal{S}*}\|_{2,1}\}$. We need a version of restricted eigenvalue assumption on $\boldsymbol{\Sigma}$.

Assumption RE$(q, m, \boldsymbol{\Sigma})$.

$$K(q, m, \boldsymbol{\Sigma}) = \min_{\mathcal{S}\in\mathfrak{A}_q} \min_{\mathbf{M}\in\mathfrak{M}_{\mathcal{S},m}:\|\mathbf{M}\|_F\neq 0} \frac{\|\boldsymbol{\Sigma}^{1/2}\mathbf{M}\|_F}{\|\mathbf{M}_{\mathcal{S}*}\|_{2,1}} > 0.$$

This and other assumptions will be discussed at the end of this section.

Write $\mathbf{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)^\top$, where $\boldsymbol{b}_j \in \mathbb{R}^G$ for $j = 1, \ldots, p$. Let $\mathcal{A} = \{j : \|\boldsymbol{b}_j\|_2 \neq 0\}$. Similarly, $\mathcal{A}_0$ denotes the index set of non-zero rows of $\mathbf{B}^0$. We assume that $|\mathcal{A}_0| \leq n$ and $p \geq n$, and note that the $p < n$ case can be dealt with similarly in our framework. Let $\tilde{d} = \min(2d, G)$, $\rho_{\min}(q) = \min_{\mathcal{S}\in\mathfrak{A}_q} \lambda_{|\mathcal{S}|}(\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})$ and $\rho_{\max}(q) = \max_{\mathcal{S}\in\mathfrak{A}_q} \lambda_1(\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}})$. We first present a preparatory result.

Proposition 2.1 *Let $\mathbf{B} \in \mathfrak{B}_d$, $\mathfrak{D}_{\mathbf{B}} = \{\boldsymbol{\Delta} = \mathbf{B}_1 - \mathbf{B} : \mathbf{B}_1 \in \mathfrak{B}_d\}$, and $\mathfrak{D}_{\mathcal{A},\mathbf{B}} = \mathfrak{D}_{\mathbf{B}} \cap \mathfrak{M}_{\mathcal{A},G}$. Let RE($|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma}$) be satisfied. Assume (A1): as $n \to \infty$,*

$$d = o(n^{1/2}), d = o\left\{\frac{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n}{\lambda_1(\boldsymbol{\Sigma})}\right\},$$

and

$$d^4|\mathcal{A}| \log\left(\frac{p}{|\mathcal{A}|}\right) = o\left\{\frac{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n}{\rho_{\max}(|\mathcal{A}|)}\right\}.$$

If $\hat{\boldsymbol{\Delta}}$ is a minimizer of $\tilde{Q}_{\lambda,\mathbf{B}}(\boldsymbol{\Delta}) = Q_\lambda(\mathbf{B} + \boldsymbol{\Delta})$ over $\mathfrak{D}_{\mathcal{A},\mathbf{B}}$, then we have, as $n \to \infty$, $\|\hat{\boldsymbol{\Delta}}\|_{2,1} = O_P(c_n)$, where

$$c_n = \sqrt{\frac{E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2)}{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n}} + \frac{\sqrt{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 n} + \sqrt{\lambda_1(\boldsymbol{\Sigma})\log(n)dGn} + \lambda}{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n}.$$

Let $\tilde{n} = n/\{\log(n)\log(p)\}$ and $\tilde{p} = \min(dn, p)$. We now state our main result, which is expressed in terms of how our method summarizes the data from a dimension-reduction perspective.

THEOREM 2.1 *Let RE($|\mathcal{A}|, \tilde{d}, \mathbf{\Sigma}$) be satisfied. Assume (A1). If*

$$\lambda^2 = C\frac{\rho_{\max}(\tilde{n})}{\rho_{\min}(\tilde{n})}\rho_{\max}(\tilde{p})d^2\log(n)\log(p)n$$

for some $C > 0$ that is large enough, then there exists a minimizer $\hat{\mathbf{B}}$ of $Q_\lambda(\mathbf{B})$ such that as $n \to \infty$,

$$\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 = O_P\left[E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2) + \frac{1}{\rho_{\min}(\tilde{n})}\lambda_1(\mathbf{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 d \right.$$
$$\left. + \frac{\rho_{\max}(\tilde{n})}{\rho_{\min}(\tilde{n})}\log(n)dG + \frac{\rho_{\max}(\tilde{n})}{\rho_{\min}(\tilde{n})}\log(n)d|\mathcal{A}|\log(p) + \lambda c_n\right]$$

for any $\mathbf{B} \in \mathfrak{B}_d$.

Theorem 2.1 gives a general result for general scalings of $|\mathcal{A}_0|, p, d, G$, and n. For example, the number G of response values or slices can diverge into infinity. The result shows how the rate of convergence depends on two complexity parameters, d for dimension reduction and λ for variable selection.

When the eigenvalues of $\mathbf{\Sigma}$ are bounded both from below and from above, we have the following result.

COROLLARY 2.1 *Suppose $0 < \delta_2 \le \lambda_p(\mathbf{\Sigma}) \le \lambda_1(\mathbf{\Sigma}) \le \delta_1 < \infty$ for some $\delta_1, \delta_2 > 0$. Let RE($|\mathcal{A}|, \tilde{d}, \mathbf{\Sigma}$) be satisfied. Assume that $d = o(n^{1/2})$ and $d^4|\mathcal{A}|\log(p/|\mathcal{A}|) = o\{K^2(|\mathcal{A}|, \tilde{d}, \mathbf{\Sigma})n\}$. If $\lambda^2 = Cd^2\log(n)\log(p)n$ for some $C > 0$ that is large enough, then there exists a minimizer $\hat{\mathbf{B}}$ of $Q_\lambda(\mathbf{B})$ such that as $n \to \infty$,*

$$\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 = O_P\left\{E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2) + \frac{|\mathcal{A}_0|^2 + \log(n)G + \log(n)|\mathcal{A}|\log(p)}{K^2(|\mathcal{A}|, \tilde{d}, \mathbf{\Sigma})}d\right\}$$

for any $\mathbf{B} \in \mathfrak{B}_d$.

If $|\mathcal{A}_0|$ is finite and $p = o\{\exp(n)\}$, then taking $\mathbf{B} = \mathbf{B}^0$ yields

$$\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 = O_P\{\log(n)G + \log(n)\log(p)\}.$$

Thus the rate of convergence achieved is essentially optimal, ignoring the $\log(n)\log(p)$ factor. Corollary 2.1 also suggests that in practice, we can use cross validation to choose the dimension d among a set of candidate values.

Write $\mathbf{B}^0 = \mathbf{M}^0\mathbf{N}^{0\top}$, where $\mathbf{M}^0$ is of dimension $p \times d_{Y|\boldsymbol{X}}$ and $\mathbf{N}^0$ is a $G \times d_{Y|\boldsymbol{X}}$ orthogonal matrix. As for predictor reduction, we have the following corollary.

COROLLARY 2.2 *Suppose that $d_{Y|\boldsymbol{X}}$ is known. Let $d = d_{Y|\boldsymbol{X}}$ and assume the conditions of Corollary 2.1. Then there exists a minimizer $\hat{\mathbf{B}} = \hat{\mathbf{M}}\hat{\mathbf{N}}^\top$ of $Q_\lambda(\mathbf{B})$ such that as $n \to \infty$,*

$$\|\tilde{\mathbf{X}}\hat{\mathbf{M}}\mathbf{R} - \tilde{\mathbf{X}}\mathbf{M}^0\|_F^2 = O_P\left\{E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2) + \frac{|\mathcal{A}_0|^2 + \log(n)G + \log(n)|\mathcal{A}|\log(p)}{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})}d\right\}$$

for some non-singular $d \times d$ matrix $\mathbf{R}$ and for any $\mathbf{B} \in \mathfrak{B}_d$.

Note that $\tilde{\mathbf{X}}\hat{\mathbf{M}}\mathbf{R}$ and $\tilde{\mathbf{X}}\hat{\mathbf{M}}$ can often be treated as the same reduction in subsequent analyses, such as graphical displays and model building.

REMARK 2.1 *The linearity condition, that is, $E(\boldsymbol{X}|\mathbf{B}^\top\boldsymbol{X})$ is linear in $\mathbf{B}^\top\boldsymbol{X}$ for any basis matrix of $\mathbb{S}_{Y|\boldsymbol{X}}$, is well-known and widely regarded as mild in sufficient dimension reduction (Li & Wang 2007). Since $\mathbb{S}_{Y|\boldsymbol{X}}$ is unknown, this condition is often assumed to hold for all possible $\mathbf{B}$, which is tantamount to assuming that $\boldsymbol{X}$ has an elliptically contoured distribution (Eaton 1983). The sub-Gaussian assumption is routinely used in high-dimensional statistical inference (Bühlmann & Van De Geer 2011). In addition to the Gaussian case, the class of sub-Gaussian variables includes, in general, all random variables whose distribution tails decrease no slower than the tails of a Gaussian random variable (Buldygin & Kozachenko 2000). The assumption RE($q, m, \boldsymbol{\Sigma}$) is a natural extension to our setting of the restricted eigenvalue assumption for the LASSO that is among the weakest and hence the most general conditions in the literature (Bickel et al. 2009). We thus do not attempt a detailed discussion of these conditions.*

3 Simulations

In this section we use a simulation study to evaluate the performance of the proposed method. Throughout, we implement the rank-constrained group LASSO algorithm in Subsection 2.4, and use five-fold cross-validation to select the reduced dimension d as well as

the regularization parameter λ. Let $\boldsymbol{0}_p$ denote the p-vector of zeros, and $\boldsymbol{e}_i$ the p-vector whose i-th element is 1 and other elements are all zero, for $i = 1, \ldots, p$.

We consider the following models

$$Y = \frac{X_1}{0.5 + (X_2 + 1.5)^2} + 0.2 \times \epsilon, \tag{7}$$

$$Y = X_1 + X_1 \times X_2 + 0.3 \times \epsilon, \tag{8}$$

$$Y = \text{sign}(X_1 + X_2) \times \log(|X_3 + X_4 + 5|) + 0.2 \times \epsilon, \tag{9}$$

where $\epsilon \sim N(0, 1)$ is independent of $\boldsymbol{X} \sim N(\boldsymbol{0}_p, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = (\varrho^{|i-j|})$ for $i, j = 1, \ldots, p$. Note that $d_{Y|\boldsymbol{X}} = 2$ in all three models. More specifically, $\mathbb{S}_{Y|\boldsymbol{X}} = \text{span}(\boldsymbol{e}_1, \boldsymbol{e}_2)$ and $\mathcal{A}_0 = \{1, 2\}$ in models (7) and (8), and $\mathbb{S}_{Y|\boldsymbol{X}} = \text{span}(\boldsymbol{e}_1 + \boldsymbol{e}_2, \boldsymbol{e}_3 + \boldsymbol{e}_4)$ and $\mathcal{A}_0 = \{1, 2, 3, 4\}$ in model (9). Since Y is continuous, we use its sliced version with $G = 5$. It is well-known that the performance of sliced inverse regression is not very sensitive to the number of slices. We take $n = 100, p \in \{50, 100, 200\}$, and $\varrho \in \{0, 0.5\}$.

For a rank-d estimator $\hat{\mathbf{B}}_d$ of $\mathbf{B}^0$, we write $\hat{\mathbf{B}} = \hat{\mathbf{B}}_d = \hat{\mathbf{M}}\hat{\mathbf{N}}^\top$ for a $p \times d$ matrix $\hat{\mathbf{M}}$ and a $G \times d$ orthogonal matrix $\hat{\mathbf{N}}$. Let $\mathbf{M}^0$ be a basis matrix of $\mathbb{S}_{Y|\boldsymbol{X}}$. Assume for the moment that $d = d_{Y|\boldsymbol{X}}$ and that $d_{Y|\boldsymbol{X}}$ is known. We use three summary statistics to measure the accuracy of estimation: the mean squared error $\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2/(nG)$ for predictor reduction, the mean squared error $\|\hat{\mathbf{B}} - \mathbf{B}^0\|_F^2/(pG)$ for parameter estimation, and the trace correlation coefficient $(\sum_{t=1}^d \phi_t^2/d)^{1/2}$ for estimation of the central subspace, where the ϕ_t^2's are the eigenvalues of $\hat{\mathbf{M}}_o^\top \mathbf{M}_o^0 \mathbf{M}_o^{0\top} \hat{\mathbf{M}}_o$, with $\hat{\mathbf{M}}_o$ and $\mathbf{M}_o^0$ the orthonormalized versions of $\hat{\mathbf{M}}$ and $\mathbf{M}^0$ respectively. Let $\hat{\mathcal{A}}$ denote the index set of non-zero rows of $\hat{\mathbf{B}}$. To assess how well our method selects predictors, we employ the model size $|\hat{\mathcal{A}}|$, the true positive rate $|\hat{\mathcal{A}} \cap \mathcal{A}_0|/|\mathcal{A}_0|$, and the false positive rate $|\hat{\mathcal{A}} \cap \mathcal{A}_0^c|/|\mathcal{A}_0^c|$. For each simulation configuration, we run 200 simulation samples and take the average of the aforementioned criteria. The results are summarized in Table 1. We see that, on average, the new method performs well in terms of both dimension reduction and variable selection. In particular, the values of the true positive rate are very high, especially in models (7) and (8), and the values of the false positive rate are close to zero. Unreported results show that the estimation accuracy improves if we recompute the estimator restricted to the selected variables.

We now study the numerical aspect of determining the structural dimension $d_{Y|\boldsymbol{X}}$. Table 2 reports the percentages of estimated structural dimension out of 200 data replications.

Table 1: Means and standard deviations (in parentheses) of the mean squared error ($\times 100$) for predictor reduction (MSE1), the mean squared error ($\times 100$) for parameter estimation (MSE2) and the trace correlation coefficient for estimation of the central subspace (TCC), and averages of the model size (MS), the true positive rate (TPR) and the false positive rate (FPR) based on 200 data replications

		$p = 50$		$p = 100$		$p = 200$	
		$\varrho = 0$	$\varrho = 0.5$	$\varrho = 0$	$\varrho = 0.5$	$\varrho = 0$	$\varrho = 0.5$
model (7)	MSE1	0.276 (0.267)	0.311 (0.316)	0.306 (0.306)	0.301 (0.324)	0.298 (0.381)	0.386 (0.434)
	MSE2	0.142 (0.139)	0.216 (0.227)	0.080 (0.082)	0.105 (0.098)	0.039 (0.051)	0.064 (0.067)
	TCC	0.985 (0.044)	0.977 (0.059)	0.987 (0.043)	0.982 (0.077)	0.978 (0.086)	0.969 (0.073)
	MS	2.410 (1.241)	2.710 (1.747)	2.390 (1.279)	2.330 (0.845)	2.360 (1.215)	2.870 (2.146)
	TPR	1.000	1.000	1.000	0.995	0.995	1.000
	FPR	0.008	0.014	0.004	0.003	0.001	0.004
model (8)	MSE1	0.411 (0.439)	0.531 (0.517)	0.528 (0.645)	0.715 (0.831)	0.568 (0.699)	0.775 (0.856)
	MSE2	0.215 (0.230)	0.343 (0.341)	0.136 (0.166)	0.223 (0.256)	0.075 (0.091)	0.118 (0.131)
	TCC	0.968 (0.081)	0.937 (0.110)	0.944 (0.126)	0.901 (0.185)	0.930 (0.153)	0.891 (0.188)
	MS	2.710 (2.073)	3.460 (2.931)	3.120 (2.619)	3.750 (2.932)	3.100 (2.301)	4.270 (3.865)
	TPR	1.000	1.000	0.995	0.975	0.985	0.975
	FPR	0.014	0.030	0.012	0.018	0.005	0.011
model (9)	MSE1	0.702 (0.789)	0.475 (0.491)	0.933 (0.998)	0.506 (0.518)	1.155 (1.164)	0.659 (0.738)
	MSE2	0.364 (0.403)	0.263 (0.240)	0.245 (0.258)	0.138 (0.125)	0.152 (0.152)	0.093 (0.104)
	TCC	0.944 (0.097)	0.955 (0.064)	0.923 (0.115)	0.955 (0.056)	0.896 (0.133)	0.934 (0.089)
	MS	4.340 (1.593)	4.230 (1.172)	4.620 (2.029)	4.210 (0.883)	4.570 (2.425)	4.350 (1.519)
	TPR	0.932	0.975	0.900	0.980	0.855	0.947
	FPR	0.013	0.007	0.011	0.003	0.005	0.002

We see that the cross-validation method works reasonably well. This indicates that, for practical purposes, our method is suitable for the case where n is moderate and p is relatively large or $p > n$.

4 Cancer drug sensitivity data

In this section, we apply our method to a drug sensitivity dataset from the Cancer Cell Line Encyclopedia project (Garnett et al. 2012). The purpose of the project is to systematically search genomic features that contain sufficient information for drug response prediction.

This dataset consists of a panel of 639 human cell lines with measurements on a compendium of genomic features including common somatic mutations, copy number aberra-

Table 2: The percentages of estimated structural dimension based on 200 data replications

		$p = 50$		$p = 100$		$p = 200$	
		$\varrho = 0$	$\varrho = 0.5$	$\varrho = 0$	$\varrho = 0.5$	$\varrho = 0$	$\varrho = 0.5$
model (7)	$\hat{d}_{Y\|X} = 1$	0.025	0.115	0.030	0.065	0.035	0.135
	$\hat{d}_{Y\|X} = 2$	0.830	0.735	0.840	0.765	0.785	0.720
	$\hat{d}_{Y\|X} \geq 3$	0.145	0.150	0.130	0.170	0.180	0.145
model (8)	$\hat{d}_{Y\|X} = 1$	0.145	0.250	0.215	0.260	0.230	0.290
	$\hat{d}_{Y\|X} = 2$	0.695	0.620	0.645	0.585	0.590	0.560
	$\hat{d}_{Y\|X} \geq 3$	0.160	0.130	0.140	0.155	0.180	0.150
model (9)	$\hat{d}_{Y\|X} = 1$	0.000	0.000	0.000	0.000	0.000	0.000
	$\hat{d}_{Y\|X} = 2$	0.785	0.845	0.685	0.795	0.605	0.750
	$\hat{d}_{Y\|X} \geq 3$	0.215	0.155	0.315	0.205	0.395	0.250

tions, and gene expressions. The responses of these cell lines to a panel of 130 drugs are also available. The drug sensitivity is assessed by IC_{50}, representing the half maximal concentration to inhibit the cell growth. For simplicity, we focus on drug-cell line combinations without missing values. Since drugs are characterized by their targeted family, processes and molecules, we narrow down our analysis to two specific groups: the targeting RTK (receptor tyrosine kinase) family and the S/T Kinase (serine/threonine protein) family. There are 176 cell lines for these two families, along with 13,847 genomic features. To handle the ultra-high dimensionality of the input data, we pre-select genomic features that are marginally associated with at least one drug in a given drug family using a cut-off on the marginal p-values. Using the threshold 0.1, we obtain 502 features for 9 drugs in the RTK targeted family and 1,198 features for 14 drugs in the S/T Kinase targeted family. We further standardize these features to have mean zero and variance one.

We divide 176 cell lines into a training set (140) and a test set (36), and then consider the 23 drugs separately. The training set is used to select features by our method. However, rather than building a complex model with the selected features as predictors, we fit a simple linear model. We then use the test set to evaluate the performance. Specifically, we calculate the prediction R-squared value, defined as $1 - \mathrm{SS}\boldsymbol{e}_{test}/\mathrm{SS}\boldsymbol{y}_{test}$, where $\mathrm{SS}\boldsymbol{z}$ denotes the sum of squares of $\boldsymbol{z}$, $\boldsymbol{e}_{test} = \boldsymbol{y}_{test} - \mathbf{X}_{test}\hat{\boldsymbol{\beta}}_{train}$ is the prediction error, $\boldsymbol{y}_{test}$ is the vector of observed responses, $\boldsymbol{X}_{test}$ is the matrix of genomic features in the test set, and $\hat{\boldsymbol{\beta}}_{train}$ is

the estimated vector of coefficients. A negative value indicates a poor prediction. To test the significance of a positive value, we use a sampling procedure as follows. For a drug with k selected features, we build 10,000 predictive models with k randomly selected features, and then calculate the null distribution of the prediction R-squares. The p-value is defined as the percentage of simulated prediction R-squared values greater than or equal to the observed one.

In total, 12 drugs have positive prediction R-squares, 5 of them have values greater than 0.2. In addition, all these 12 drugs reach a significance level of 0.05, corresponding to an estimate of $FDR = 23 \times 0.05/12 \approx 0.1$. These suggest that the identified genomic features can provide useful predictions of drug sensitivity. We further check whether drugs targeting the same process share any features. The results for drugs targeting the ERK signalling pathway and the AMPK metabolism are summarized in Tables 3 and 4.

The selected features are well supported by the cancer biology literature. It has been reported that EGR1 is regulated by EGF and the ERK signaling pathway in cancer (Gregg & Fraizer 2011), and one possible path is ERK induces P35, a neuron-specific activator of CDK5, through induction of EGR1 (Harada et al. 2001). It has also been suggested that DAF-GPI signaling complex is responsible for early ERK activation (Luo et al. 2002) in CVB3-infected HeLa cells; PHLDA1 is a direct target of ERK regulation (Wang et al. 2012); SHC1 and PGK1 are all genes on the ERK pathway (Zheng et al. 2013); IRS1 (insulin receptor substrate) coordinates skeletal muscle growth and metabolism via the AKT and AMPK pathways (Long et al. 2011). In summary, our results suggest that the expression of genes can be used to predict the sensitivity of drugs targeting related pathways.

5 Proofs

Proof of Proposition 2.1. It is sufficient to show that for any given $\delta > 0$, there exists a large $C_\delta > 0$ such that, for n sufficiently large,

$$\mathbb{P}\left\{\sup_{\mathbf{\Delta}\in\mathfrak{D}_{\mathcal{A},\mathbf{B}}:\|\mathbf{\Delta}\|_{2,1}=C_\delta} Q_\lambda(\mathbf{B}+c_n\mathbf{\Delta}) > Q_\lambda(\mathbf{B})\right\} \geq 1-\delta.$$

Let $\mathbf{\Delta} \in \mathfrak{D}_{\mathcal{A},\mathbf{B}}$. Since $\text{rank}(\mathbf{\Delta}) \leq \tilde{d}$, we have $\mathbf{\Delta} = \mathbf{M}\mathbf{N}^\top$ for a $p \times \tilde{d}$ matrix $\mathbf{M}$ and a $G \times \tilde{d}$

Table 3: Shared features for drugs targeting the ERK signalling pathway

Feature / Drug	AZ628	RDEA119	SB590885	AZD6244
EGR1	0.000	0.053	0.000	0.011
FLJ23556	0.000	0.013	0.039	0.046
GPI	0.000	-0.043	0.000	-0.067
GULP1	0.000	0.021	0.000	0.065
PGK1	0.055	0.000	0.000	-0.014
PHLDA1	0.000	0.017	0.071	0.011
POLS	0.017	0.000	0.000	-0.029
SHC1	0.000	0.033	0.000	0.033
total features	12	43	8	31

Table 4: Shared features for drugs targeting the AMPK metabolism

Feature / Drug	Metformin	AICAR
ACYP1	0.033	0.012
GPI	0.042	0.032
IRS1	-0.023	-0.035
total features	36	10

orthogonal matrix $\mathbf{N}$. Furthermore, $\|\mathbf{\Delta}_{\mathcal{S}*}\|_{2,1} = \|\mathbf{M}_{\mathcal{S}*}\|_{2,1}$ for any $\mathcal{S} \subseteq \{1, \ldots, p\}$. Hence

$$\mathfrak{D}_{\mathcal{A},\mathbf{B}} = \{\mathbf{\Delta} = \mathbf{M}\mathbf{N}^\top : \mathbf{M} \in \mathfrak{M}_{\mathcal{A},\tilde{d}}, \mathbf{N} \in \mathbb{R}^{G\times\tilde{d}}, \mathbf{N}^\top\mathbf{N} = \mathbf{I}_{\tilde{d}}\}. \tag{10}$$

Let $D(\mathbf{\Delta}) = Q_\lambda(\mathbf{B} + c_n\mathbf{\Delta}) - Q_\lambda(\mathbf{B})$. Then

$$\begin{aligned} D(\mathbf{\Delta}) &= c_n^2\|\tilde{\mathbf{X}}\mathbf{\Delta}\|_F^2 - 2c_n\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}, \tilde{\mathbf{X}}\mathbf{\Delta}\rangle + 2\lambda(\|\mathbf{B} + c_n\mathbf{\Delta}\|_{2,1} - \|\mathbf{B}\|_{2,1}) \\ &\geq c_n^2\|\tilde{\mathbf{X}}\mathbf{\Delta}\|_F^2 - 2c_n\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}, \tilde{\mathbf{X}}\mathbf{\Delta}\rangle - 2c_n\lambda\|\mathbf{\Delta}\|_{2,1} \\ &= T_1 + T_2 + T_3. \end{aligned}$$

First, consider T_1. We have $T_1 = c_n^2(\|\mathbf{X}\mathbf{\Delta}\|_F^2 - \|\mathbf{P}_n\mathbf{X}\mathbf{\Delta}\|_F^2)$. By (10) and Lemma 6.5 in Wang et al. (2015), there exists some $C > 0$ such that, with probability tending to 1 as $n \to \infty$,

$$\|\mathbf{X}\mathbf{\Delta}\|_F^2 \geq CK^2(|\mathcal{A}|, \tilde{d}, \mathbf{\Sigma})n\|\mathbf{\Delta}_{\mathcal{A}*}\|_{2,1}^2 \geq \frac{1}{9}CK^2(|\mathcal{A}|, \tilde{d}, \mathbf{\Sigma})n\|\mathbf{\Delta}\|_{2,1}^2.$$

Let $\mathbf{Z} = \mathbf{X}\mathbf{\Sigma}^{-1/2}$. Then $\mathbf{P}_n\mathbf{X}\mathbf{\Delta} = \mathbf{P}_n\mathbf{Z}\mathbf{\Sigma}^{1/2}\mathbf{\Delta}$. Let $\mathbf{\Sigma}^{1/2}\mathbf{\Delta} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ denote the singular value decomposition of $\mathbf{\Sigma}^{1/2}\mathbf{\Delta}$; that is, $\mathbf{U}$ and $\mathbf{V}$ are $p \times \tilde{d}$ and $G \times \tilde{d}$ orthogonal matrices,

and $\mathbf{D}$ a diagonal matrix. It holds that

$$\|\mathbf{P}_n\mathbf{X}\boldsymbol{\Delta}\|_F^2 = \|\mathbf{P}_n\mathbf{Z}\mathbf{U}\mathbf{D}\|_F^2 \le \sigma_1^2(\mathbf{P}_n\mathbf{Z}\mathbf{U})\|\mathbf{D}\|_F^2 \le \sigma_1^2(\mathbf{P}_n\mathbf{Z}\mathbf{U})\lambda_1(\boldsymbol{\Sigma})\|\boldsymbol{\Delta}\|_{2,1}^2.$$

From Lemma 6.6 of Wang et al. (2015) we know that $\sigma_1^2(\mathbf{P}_n\mathbf{Z}\mathbf{U}) = O_P(d)$. Hence

$$\|\mathbf{P}_n\mathbf{X}\boldsymbol{\Delta}\|_F^2 = O_P\{\lambda_1(\boldsymbol{\Sigma})d\}\|\boldsymbol{\Delta}\|_{2,1}^2 = o_P\{K^2(|\mathcal{A}|,\tilde{d},\boldsymbol{\Sigma})n\}\|\boldsymbol{\Delta}\|_{2,1}^2.$$

Consequently, with probability tending to 1 as $n \to \infty$,

$$T_1 \ge \frac{1}{2}c_n^2\|\mathbf{X}\boldsymbol{\Delta}\|_F^2 \ge \frac{1}{18}Cc_n^2K^2(|\mathcal{A}|,\tilde{d},\boldsymbol{\Sigma})n\|\boldsymbol{\Delta}\|_{2,1}^2.$$

Now we consider the second term T_2. It holds that

$$\begin{aligned}
|\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}, \tilde{\mathbf{X}}\boldsymbol{\Delta}\rangle| &= |\langle \mathbf{L} - \mathbf{X}\mathbf{B}, \tilde{\mathbf{X}}\boldsymbol{\Delta}\rangle| \\
&= |\langle \mathbf{X}\mathbf{B}^0 - \mathbf{X}\mathbf{B}, \tilde{\mathbf{X}}\boldsymbol{\Delta}\rangle + \langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \tilde{\mathbf{X}}\boldsymbol{\Delta}\rangle| \\
&\le |\langle \mathbf{X}\mathbf{B}^0 - \mathbf{X}\mathbf{B}, \tilde{\mathbf{X}}\boldsymbol{\Delta}\rangle| + |\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{X}\boldsymbol{\Delta}\rangle| + |\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{P}_n\mathbf{X}\boldsymbol{\Delta}\rangle| \\
&= T_{21} + T_{22} + T_{23}.
\end{aligned}$$

We have

$$\begin{aligned}
T_{21} &\le \|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F\|\tilde{\mathbf{X}}\boldsymbol{\Delta}\|_F \\
&= O_P\left\{\sqrt{E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2)}\right\}\|\tilde{\mathbf{X}}\boldsymbol{\Delta}\|_F = O_P\left\{c_n\sqrt{K^2(|\mathcal{A}|,\tilde{d},\boldsymbol{\Sigma})n}\right\}\|\tilde{\mathbf{X}}\boldsymbol{\Delta}\|_F.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
T_{22} &\le |\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{X}_{*\mathcal{A}_0}\boldsymbol{\Delta}_{\mathcal{A}_0*}\rangle| + |\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{X}_{*\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}\rangle| \\
&\le \|\mathbf{X}_{*\mathcal{A}_0}^\top(\mathbf{L} - \mathbf{X}\mathbf{B}^0)\|_F\|\boldsymbol{\Delta}\|_{2,1} + |\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{X}_{*\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}\rangle| \\
&= O_P\left\{\sqrt{E(\|\mathbf{X}_{*\mathcal{A}_0}^\top(\mathbf{L} - \mathbf{X}\mathbf{B}^0)\|_F^2)}\right\}\|\boldsymbol{\Delta}\|_{2,1} + |\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{X}_{*\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}\rangle|.
\end{aligned}$$

Let $\boldsymbol{L} = \{I(Y=1), \ldots, I(Y=G)\}^\top \in \mathbb{R}^G$ and $\boldsymbol{W}_{\mathcal{A}_0} = \boldsymbol{X}_{\mathcal{A}_0}(\boldsymbol{L}^\top - \boldsymbol{X}_{\mathcal{A}_0}^\top\mathbf{B}_{\mathcal{A}_0*}^0) \in \mathbb{R}^{|\mathcal{A}_0|\times G}$. It is easy to show that

$$E(\|\mathbf{X}_{*\mathcal{A}_0}^\top(\mathbf{L} - \mathbf{X}\mathbf{B}^0)\|_F^2) = nE(\|\boldsymbol{W}_{\mathcal{A}_0}\|_F^2) = O\left\{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2n\right\}.$$

Hence

$$\begin{aligned}
T_{22} &\le O_P\left\{\sqrt{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2n}\right\}\|\boldsymbol{\Delta}\|_{2,1} + |\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{X}_{*\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}\rangle| \\
&= O_P\left\{\sqrt{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2n}\right\}\|\boldsymbol{\Delta}\|_{2,1} + \left\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{Z}_{\mathcal{A}_0^c}\boldsymbol{\Sigma}_{\mathcal{A}_0^c\mathcal{A}_0^c}^{\frac{1}{2}}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}\right\rangle,
\end{aligned}$$

where $\mathbf{Z}_{\mathcal{A}_0^c} = \mathbf{X}_{*\mathcal{A}_0^c}\boldsymbol{\Sigma}^{-1/2}_{\mathcal{A}_0^c\mathcal{A}_0^c}$. Let $\mathbf{L} - \mathbf{X}\mathbf{B}^0 = \mathbf{U}_1\mathbf{D}_1\mathbf{V}_1^\top$ denote the singular value decomposition of $\mathbf{L} - \mathbf{X}\mathbf{B}^0$ and $\boldsymbol{\Sigma}^{1/2}_{\mathcal{A}_0^c\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*} = \mathbf{U}_2\mathbf{D}_2\mathbf{V}_2^\top$ the singular value decomposition of $\boldsymbol{\Sigma}^{1/2}_{\mathcal{A}_0^c\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}$. Then $\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{Z}_{\mathcal{A}_0^c}\boldsymbol{\Sigma}^{1/2}_{\mathcal{A}_0^c\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}\rangle = \langle \mathbf{U}_2^\top\mathbf{Z}^\top_{\mathcal{A}_0^c}\mathbf{U}_1, \mathbf{D}_2\mathbf{V}_2^\top\mathbf{V}_1\mathbf{D}_1\rangle$. Using von Neumann's trace inequality and the fact that $\mathrm{rank}(\mathbf{D}_2\mathbf{V}_2^\top\mathbf{V}_1\mathbf{D}_1) \le 2d$, we obtain

$$\begin{aligned}
|\langle \mathbf{L} - \mathbf{X}\mathbf{B}^0, \mathbf{X}_{*\mathcal{A}_0^c}\boldsymbol{\Delta}_{\mathcal{A}_0^c*}\rangle| &\le \sigma_1(\mathbf{U}_1^\top\mathbf{Z}_{\mathcal{A}_0^c}\mathbf{U}_2)\sqrt{2d}\|\mathbf{D}_2\mathbf{V}_2^\top\mathbf{V}_1\mathbf{D}_1\|_F \\
&\le \sigma_1(\mathbf{U}_1^\top\mathbf{Z}_{\mathcal{A}_0^c}\mathbf{U}_2)\sqrt{2d}\sigma_1(\mathbf{L} - \mathbf{X}\mathbf{B}^0)\sqrt{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0^c\mathcal{A}_0^c})}\|\boldsymbol{\Delta}\|_{2,1}.
\end{aligned}$$

By Lemma 6.10 of Wang et al. (2015), $\sigma_1^2(\mathbf{U}_1^\top\mathbf{Z}_{\mathcal{A}_0^c}\mathbf{U}_2) = O_P\{\log(n)G\}$. Moreover, it is easy to show that

$$\sigma_1^2(\mathbf{L} - \mathbf{X}\mathbf{B}^0) \le 2\sigma_1^2(\mathbf{L}) + 2\sigma_1^2(\mathbf{X}\mathbf{B}^0) \le 2n + 2\sigma_1^2\left(\boldsymbol{\Sigma}^{\frac{1}{2}}_{\mathcal{A}_0\mathcal{A}_0}\mathbf{B}^0_{\mathcal{A}_0*}\right)\sigma_1^2(\mathbf{Z}_{\mathcal{A}_0}) \le 2n + 2\sigma_1^2(\mathbf{Z}_{\mathcal{A}_0}),$$

where $\mathbf{Z}_{\mathcal{A}_0} = \mathbf{X}_{*\mathcal{A}_0}\boldsymbol{\Sigma}^{-1/2}_{\mathcal{A}_0\mathcal{A}_0}$. From Lemma 6.8 in Wang et al. (2015) and $|\mathcal{A}_0| \le n$ we know that

$$\sigma_1^2(\mathbf{L} - \mathbf{X}\mathbf{B}^0) = O_P(n).$$

Consequently,

$$\begin{aligned}
T_{22} &\le O_P\left\{\sqrt{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 n}\right\}\|\boldsymbol{\Delta}\|_{2,1} + O_P\left\{\sqrt{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0^c\mathcal{A}_0^c})\log(n)dGn}\right\}\|\boldsymbol{\Delta}\|_{2,1} \\
&= O_P\{c_n K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n\}\|\boldsymbol{\Delta}\|_{2,1}.
\end{aligned}$$

Note that

$$\begin{aligned}
T_{23} &\le \|\mathbf{L} - \mathbf{X}\mathbf{B}^0\|_F\|\mathbf{P}_n\mathbf{X}\boldsymbol{\Delta}\|_F \\
&= O_P\left\{\sqrt{\lambda_1(\boldsymbol{\Sigma})dGn}\right\}\|\boldsymbol{\Delta}\|_{2,1} = O_P\{c_n K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n\}\|\boldsymbol{\Delta}\|_{2,1}.
\end{aligned}$$

Combining, we have

$$T_2 = O_P\left\{c_n^2\sqrt{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n}\right\}\|\tilde{\mathbf{X}}\boldsymbol{\Delta}\|_F + O_P\{c_n^2 K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n\}\|\boldsymbol{\Delta}\|_{2,1}.$$

Finally, note that $|T_3| = 2c_n\lambda\|\boldsymbol{\Delta}\|_{2,1} = O_P\{c_n^2 K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n\}\|\boldsymbol{\Delta}\|_{2,1}$. Therefore, by choosing a sufficiently large C_δ, T_1 dominates T_2 and T_3 uniformly in $\|\boldsymbol{\Delta}\|_{2,1} = C_\delta$ with high probability. The proof is complete.

PROOF OF THEOREM 2.1. Write $\hat{\mathbf{B}} = (\hat{\boldsymbol{b}}_1, \ldots, \hat{\boldsymbol{b}}_p)^\top$, where $\hat{\boldsymbol{b}}_j \in \mathbb{R}^G$ for $j = 1, \ldots, p$. Let $\hat{\mathcal{A}} = \{j : \|\hat{\boldsymbol{b}}_j\|_2 \neq 0\}$. We assume $|\hat{\mathcal{A}}| \leq dn$. Otherwise, by Lemma 1 in Liu & Zhang (2009), we can always construct a solution to satisfy this condition. By definition,

$$\|\mathbf{L} - \tilde{\mathbf{X}}\hat{\mathbf{B}}\|_F^2 + 2\lambda\|\hat{\mathbf{B}}\|_{2,1} \leq \|\mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}\|_F^2 + 2\lambda\|\mathbf{B}\|_{2,1},$$

which is equivalent to

$$\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\lambda\|\hat{\mathbf{B}}\|_{2,1} \leq \|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle + 2\lambda\|\mathbf{B}\|_{2,1}.$$

Then

$$\begin{aligned} & \|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\lambda\|\hat{\mathbf{B}}_{\mathcal{A}^c *}\|_{2,1} \\ \leq\ & \|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle + 2\lambda(\|\mathbf{B}\|_{2,1} - \|\hat{\mathbf{B}}_{\mathcal{A}*}\|_{2,1}). \end{aligned}$$

Since $\|\hat{\mathbf{B}}_{\mathcal{A}^c *}\|_{2,1} = \|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}^c *}\|_{2,1}$ and $\|\mathbf{B}\|_{2,1} = \|\mathbf{B}_{\mathcal{A}*}\|_{2,1}$, we obtain

$$\begin{aligned} & \|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\lambda\|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}^c *}\|_{2,1} \\ \leq\ & \|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle + 2\lambda(\|\mathbf{B}_{\mathcal{A}*}\|_{2,1} - \|\hat{\mathbf{B}}_{\mathcal{A}*}\|_{2,1}) \\ \leq\ & \|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle + 2\lambda\|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}*}\|_{2,1}. \end{aligned}$$

On one hand, if

$$\|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle \leq 2\lambda\|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}*}\|_{2,1},$$

then

$$|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}^c *}\|_{2,1} \leq 2\|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}*}\|_{2,1}.$$

By Proposition 2.1, we have

$$\|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}*}\|_{2,1} = O_P(C_n),$$

where

$$C_n = \sqrt{\frac{E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2)}{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n}} + \frac{\sqrt{\lambda_1(\boldsymbol{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 n} + \sqrt{\lambda_1(\boldsymbol{\Sigma})\log(n)dGn} + \lambda}{K^2(|\mathcal{A}|, \tilde{d}, \boldsymbol{\Sigma})n}.$$

Hence

$$\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 \leq\ 4\lambda\|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}*}\|_{2,1} = O_P(\lambda C_n). \tag{11}$$

On the other hand, if

$$\|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle \geq 2\lambda\|(\hat{\mathbf{B}} - \mathbf{B})_{\mathcal{A}*}\|_{2,1},$$

then

$$\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 \leq 2\|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 4\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle.$$

Consider the second term on the right-hand side. Let $\mathbf{E} = \mathbf{L} - \mathbf{X}\mathbf{B}^0$. Then

$$4\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle = 4\langle \mathbf{E}, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle = 4\langle \mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E}, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle,$$

where $\breve{\mathcal{A}} = \hat{\mathcal{A}} \cup \mathcal{A}$ and $\mathbf{P}_{\mathcal{S}}$ is the projection matrix on the space spanned by $\tilde{\mathbf{X}}_{*\mathcal{S}}$. Using von Neumann's trace inequality and the fact that $\mathrm{rank}(\hat{\mathbf{B}} - \mathbf{B}) \leq 2d$, we get

$$4\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle \;\leq\; 4\sigma_1(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})\sqrt{2d}\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\|_F.$$

Consequently, for any $c > 0$,

$$\begin{aligned} 4\langle \mathbf{L} - \tilde{\mathbf{X}}\mathbf{B}^0, \tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}\rangle &\leq 4\sigma_1(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})\sqrt{2d}(\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F + \|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F) \\ &\leq 16c\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})d + \frac{1}{c}(\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + \|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2). \end{aligned}$$

Setting $c = 2$, then

$$\frac{1}{2}\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 \;\leq\; \frac{5}{2}\|\tilde{\mathbf{X}}\mathbf{B} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 32\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})d. \tag{12}$$

The result is trivially true when $|\breve{\mathcal{A}}| = 0$. Hence we assume that $|\breve{\mathcal{A}}| \geq 1$. Consider the following two cases: (i) $|\breve{\mathcal{A}}| \leq \tilde{n}$ and (ii) $|\breve{\mathcal{A}}| > \tilde{n}$.

For any $\mathcal{S} \subseteq \{1, \ldots, p\}$ with $|\mathcal{S}| \geq 1$, let $\mathbf{Z}_{\mathcal{S}} = \mathbf{X}_{*\mathcal{S}}\boldsymbol{\Sigma}_{\mathcal{S}\mathcal{S}}^{-1/2} \in \mathbb{R}^{n\times|\mathcal{S}|}$. In case (i), we know from Lemma 6.7 and Lemma 6.9 in Wang et al. (2015) that, there exist some $C_1, C_2 > 0$ such that $\sigma_1^2(\mathbf{P}_n\mathbf{Z}_{\breve{\mathcal{A}}}) \leq C_1 n/\log(n)$ and $\sigma_{|\breve{\mathcal{A}}|}^2(\mathbf{Z}_{\breve{\mathcal{A}}}) \geq C_2 n$ almost surely for n sufficiently large. It holds that

$$\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}}^\top\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}} = \mathbf{X}_{*\breve{\mathcal{A}}}^\top\mathbf{X}_{*\breve{\mathcal{A}}} - \mathbf{X}_{*\breve{\mathcal{A}}}^\top\mathbf{P}_n\mathbf{X}_{*\breve{\mathcal{A}}} = \boldsymbol{\Sigma}_{\breve{\mathcal{A}}\breve{\mathcal{A}}}^{\frac{1}{2}}(\mathbf{Z}_{\breve{\mathcal{A}}}^\top\mathbf{Z}_{\breve{\mathcal{A}}} - \mathbf{Z}_{\breve{\mathcal{A}}}^\top\mathbf{P}_n\mathbf{Z}_{\breve{\mathcal{A}}})\boldsymbol{\Sigma}_{\breve{\mathcal{A}}\breve{\mathcal{A}}}^{\frac{1}{2}}.$$

Applying Weyl's inequality, we obtain

$$\lambda_{|\breve{\mathcal{A}}|}(\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}}^\top\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}}) \geq \lambda_{|\breve{\mathcal{A}}|}(\boldsymbol{\Sigma}_{\breve{\mathcal{A}}\breve{\mathcal{A}}})\left\{C_2 n - C_1\frac{n}{\log(n)}\right\} \geq \frac{C_2}{2}\rho_{\min}(\tilde{n})n$$

for n sufficiently large. Hence

$$\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E}) = \lambda_1\left\{\mathbf{E}^\top\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}}(\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}}^\top\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}})^{-1}\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}}^\top\mathbf{E}\right\} \le \frac{2}{C_2\rho_{\min}(\tilde{n})n}\sigma_1^2(\tilde{\mathbf{X}}_{*\breve{\mathcal{A}}}^\top\mathbf{E}).$$

Let $\tilde{\mathbf{E}} = \mathbf{E} - \mathbf{P}_n\mathbf{E}$. Then

$$E\{\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})\} = O\left\{\frac{1}{\rho_{\min}(\tilde{n})n}\right\}E\{\sigma_1^2(\mathbf{X}_{*\breve{\mathcal{A}}}^\top\tilde{\mathbf{E}})\}. \tag{13}$$

Let $\breve{\mathcal{A}}_{\setminus 0} = \breve{\mathcal{A}} \setminus \mathcal{A}_0$. It holds that

$$\sigma_1^2(\mathbf{X}_{*\breve{\mathcal{A}}}^\top\tilde{\mathbf{E}}) \le \sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\tilde{\mathbf{E}}) + \sigma_1^2(\mathbf{X}_{*\breve{\mathcal{A}}_{\setminus 0}}^\top\tilde{\mathbf{E}}).$$

We have

$$\begin{aligned}\sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\tilde{\mathbf{E}}) &\le 2\sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\mathbf{E}) + 2\sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\mathbf{P}_n\mathbf{E}) \\ &\le 2\sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\mathbf{E}) + 2\sigma_1^2(\mathbf{E})\lambda_1(\mathbf{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})\sigma_1^2(\mathbf{P}_n\mathbf{Z}_{\mathcal{A}_0}).\end{aligned}$$

It is easy to show that

$$E\{\sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\mathbf{E})\} \le E(\|\mathbf{X}_{*\mathcal{A}_0}^\top\mathbf{E}\|_F^2) = O\left\{\lambda_1(\mathbf{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 n\right\}.$$

Furthermore, by Lemma 6.8 of Wang et al. (2015) and $|\mathcal{A}_0| \le n$, we have, for some $C_3 > 0$,

$$\sigma_1^2(\mathbf{E}) \le 2\sigma_1^2(\mathbf{L}) + 2\sigma_1^2(\mathbf{X}\mathbf{B}^0) \le C_3 n$$

almost surely for n sufficiently large. Consequently,

$$E\{\sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\tilde{\mathbf{E}})\} = O\left\{\lambda_1(\mathbf{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 n\right\} + O\{\lambda_1(\mathbf{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})n\}E\{\sigma_1^2(\mathbf{P}_n\mathbf{Z}_{\mathcal{A}_0})\}.$$

From Lemma 6.6 in Wang et al. (2015) we know that $E\{\sigma_1^2(\mathbf{P}_n\mathbf{Z}_{\mathcal{A}_0})\} = O(|\mathcal{A}_0|)$. Hence

$$E\{\sigma_1^2(\mathbf{X}_{*\mathcal{A}_0}^\top\tilde{\mathbf{E}})\} = O\left\{\lambda_1(\mathbf{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 n\right\}.$$

Let $\mathbf{E} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ denote the singular value decomposition of $\mathbf{E}$; that is, $\mathbf{U}$ and $\mathbf{V}$ are $n \times G$ and $G \times G$ orthogonal matrices, and $\mathbf{D}$ a diagonal matrix. Then

$$\sigma_1^2(\mathbf{X}_{*\breve{\mathcal{A}}_{\setminus 0}}^\top\tilde{\mathbf{E}}) \le \rho_{\max}(\tilde{n})\sigma_1^2(\mathbf{E})\lambda_1(\mathbf{Z}_{\breve{\mathcal{A}}_{\setminus 0}}^\top\mathbf{U}\mathbf{U}^\top\mathbf{Z}_{\breve{\mathcal{A}}_{\setminus 0}}) = \rho_{\max}(\tilde{n})\sigma_1^2(\mathbf{E})\sigma_1^2(\mathbf{Z}_{\breve{\mathcal{A}}_{\setminus 0}}^\top\mathbf{U}).$$

By Lemma 6.11 of Wang et al. (2015), there exists some $C_4 > 0$ such that, for n sufficiently large,

$$E\{\sigma_1^2(\mathbf{Z}_{\breve{\mathcal{A}}\setminus_0}^\top \mathbf{U})\} \le C_4 \log(n)\{E(|\breve{\mathcal{A}}|)\log p + G\}.$$

Consequently, there exists some $C_5 > 0$ such that, for n sufficiently large,

$$E\{\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})\} \le C_5\{C_{n1} + C_{n2}E(|\breve{\mathcal{A}}|)\},$$

where

$$C_{n1} = \frac{1}{\rho_{\min}(\tilde{n})}\lambda_1(\mathbf{\Sigma}_{\mathcal{A}_0\mathcal{A}_0})|\mathcal{A}_0|^2 + \frac{\rho_{\max}(\tilde{n})}{\rho_{\min}(\tilde{n})}\log(n)G$$

and

$$C_{n2} = \frac{\rho_{\max}(\tilde{n})}{\rho_{\min}(\tilde{n})}\log(n)\log(p).$$

In case (ii), $E\{\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})\} \le E\{\sigma_1^2(\mathbf{E})\}$. Combining, we have

$$E\{\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})\} \le C_5\{C_{n1} + C_{n2}E(|\breve{\mathcal{A}}|)\} \tag{14}$$

for n sufficiently large.

Write $\hat{\mathbf{B}} = \hat{\mathbf{M}}\hat{\mathbf{N}}^\top$ for a $p \times d$ matrix $\hat{\mathbf{M}}$ and a $G \times d$ orthogonal matrix $\hat{\mathbf{N}}$. Then

$$\|\mathbf{L} - \tilde{\mathbf{X}}\hat{\mathbf{B}}\|_F^2 = \|(\mathbf{L} - \tilde{\mathbf{X}}\hat{\mathbf{B}})(\hat{\mathbf{N}}, \hat{\mathbf{N}}^\perp)\|_F^2 = \|\mathbf{L}\hat{\mathbf{N}} - \tilde{\mathbf{X}}\hat{\mathbf{M}}\|_F^2 + \|\mathbf{L}\hat{\mathbf{N}}^\perp\|_F^2,$$

where $(\hat{\mathbf{N}}, \hat{\mathbf{N}}^\perp)$ is a $G \times G$ orthogonal matrix. Let $\boldsymbol{e}_j$ denote the p-dimensional vector whose j-th element is 1 and other elements are 0, for $j = 1, \ldots, p$. By the Karush–Kuhn–Tucker optimality condition,

$$\|\boldsymbol{e}_j^\top \tilde{\mathbf{X}}^\top(\tilde{\mathbf{X}}\hat{\mathbf{M}} - \mathbf{L}\hat{\mathbf{N}})\|_F = \lambda, \quad j \in \hat{\mathcal{A}}.$$

It follows that

$$\begin{aligned}
|\hat{\mathcal{A}}|\lambda^2 &= \sum_{j\in\hat{\mathcal{A}}} \|\boldsymbol{e}_j^\top \tilde{\mathbf{X}}^\top(\tilde{\mathbf{X}}\hat{\mathbf{M}} - \mathbf{L}\hat{\mathbf{N}})\|_F^2 \\
&\le 2\sum_{j\in\hat{\mathcal{A}}} \|\boldsymbol{e}_j^\top \tilde{\mathbf{X}}^\top(\tilde{\mathbf{X}}\hat{\mathbf{M}} - \tilde{\mathbf{X}}\mathbf{B}^0\hat{\mathbf{N}})\|_F^2 + 2\sum_{j\in\hat{\mathcal{A}}} \|\boldsymbol{e}_j^\top \tilde{\mathbf{X}}^\top \mathbf{E}\hat{\mathbf{N}}\|_F^2 \\
&\le 2\lambda_1(\mathbf{X}_{*\hat{\mathcal{A}}}^\top \mathbf{X}_{*\hat{\mathcal{A}}})\|\tilde{\mathbf{X}}\hat{\mathbf{M}} - \tilde{\mathbf{X}}\mathbf{B}^0\hat{\mathbf{N}}\|_F^2 + 2\lambda_1(\mathbf{X}_{*\hat{\mathcal{A}}}^\top \mathbf{X}_{*\hat{\mathcal{A}}})\|\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E}\hat{\mathbf{N}}\|_F^2 \\
&\le 2\lambda_1(\mathbf{X}_{*\hat{\mathcal{A}}}^\top \mathbf{X}_{*\hat{\mathcal{A}}})\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + 2\lambda_1(\mathbf{X}_{*\hat{\mathcal{A}}}^\top \mathbf{X}_{*\hat{\mathcal{A}}})\sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})d \\
&\le 2\lambda_1(\mathbf{\Sigma}_{\hat{\mathcal{A}}\hat{\mathcal{A}}})\sigma_1^2(\mathbf{Z}_{\hat{\mathcal{A}}})\{\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + \sigma_1^2(\mathbf{P}_{\breve{\mathcal{A}}}\mathbf{E})d\}.
\end{aligned}$$

Hence

$$\begin{aligned} E(|\hat{\mathcal{A}}|\lambda^2) &\leq 2E[\lambda_1(\mathbf{\Sigma}_{\hat{\mathcal{A}}\hat{\mathcal{A}}})\sigma_1^2(\mathbf{Z}_{\hat{\mathcal{A}}})\{\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})d\}] \\ &= 2E(E[\lambda_1(\mathbf{\Sigma}_{\hat{\mathcal{A}}\hat{\mathcal{A}}})\sigma_1^2(\mathbf{Z}_{\hat{\mathcal{A}}})\{\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})d\}|\hat{\mathcal{A}}]). \end{aligned}$$

By Lemma 6.8 in Wang et al. (2015), there exists some $C_6 > 0$ such that

$$E(|\hat{\mathcal{A}}|\lambda^2) \leq C_6 E(E[\lambda_1(\mathbf{\Sigma}_{\hat{\mathcal{A}}\hat{\mathcal{A}}})(n + |\hat{\mathcal{A}}|)\{\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})d\}|\hat{\mathcal{A}}]).$$

This, together with $|\hat{\mathcal{A}}| \leq dn$, implies that

$$\begin{aligned} E(|\hat{\mathcal{A}}|\lambda^2) &\leq C_6\rho_{\max}(\tilde{p})(n + dn)E[E\{\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})d|\hat{\mathcal{A}}\}] \\ &\leq 2C_6\rho_{\max}(\tilde{p})dnE\{\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})d\}. \end{aligned}$$

Let $C_{n3} = 2C_6\rho_{\max}(\tilde{p})d^2n$. Then

$$E(|\hat{\mathcal{A}}|) \leq \frac{C_{n3}}{\lambda^2}E\left\{\frac{1}{d}\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}_0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})\right\}.$$

Together with (14), this yields that, for n sufficiently large,

$$E\{\sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})\} \leq C_5\left[C_{n1} + C_{n2}|\mathcal{A}| + C_{n2}\frac{C_{n3}}{\lambda^2}E\left\{\frac{1}{d}\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}_0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})\right\}\right].$$

Setting $\lambda^2 = 129C_5C_{n2}C_{n3}$, then

$$E\{\sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})\} \leq C_5(C_{n1} + C_{n2}|\mathcal{A}|) + \frac{1}{129}E\left\{\frac{1}{d}\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}_0\|_F^2 + \sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})\right\}.$$

A simple calculation shows that

$$E\{\sigma_1^2(\mathbf{P}_{\check{\mathcal{A}}}\mathbf{E})\} \leq \frac{129C_5}{128}(C_{n1} + C_{n2}|\mathcal{A}|) + \frac{1}{128d}E(\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}_0\|_F^2).$$

Consequently, from (12) we have, for n sufficiently large,

$$\begin{aligned} &\frac{1}{2}E(\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2) \\ \leq\ &\frac{5}{2}E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2) + \frac{129C_5}{4}(C_{n1} + C_{n2}|\mathcal{A}|)d + \frac{1}{4}E(\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}_0\|_F^2), \end{aligned}$$

and hence

$$E(\|\tilde{\mathbf{X}}\hat{\mathbf{B}} - \tilde{\mathbf{X}}\mathbf{B}^0\|_F^2) \leq 10E(\|\mathbf{X}\mathbf{B} - \mathbf{X}\mathbf{B}^0\|_F^2) + 129C_5\{C_{n1} + C_{n2}|\mathcal{A}|\}d.$$

Combining this with (11), the proof is complete.

References

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009), 'Simultaneous analysis of Lasso and Dantzig selector', *The Annals of Statistics* **37**(4), 1705–1732.

Bondell, H. D. & Li, L. (2009), 'Shrinkage inverse regression estimation for model-free variable selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(1), 287–299.

Breheny, P. & Huang, J. (2015), 'Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors', *Statistics and Computing* **25**(2), 173–187.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media.

Buldygin, V. V. & Kozachenko, Y. V. (2000), *Metric Characterization of Random Variables and Random Processes*, American Mathematical Society.

Bunea, F., She, Y. & Wegkamp, M. H. (2012), 'Joint variable and rank selection for parsimonious estimation of high-dimensional matrices', *The Annals of Statistics* **40**(5), 2359–2388.

Chen, L. & Huang, J. Z. (2012), 'Sparse reduced-rank regression for simultaneous dimension reduction and variable selection', *Journal of the American Statistical Association* **107**(500), 1533–1545.

Chen, X., Zou, C. & Cook, R. D. (2010), 'Coordinate-independent sparse sufficient dimension reduction and variable selection', *The Annals of Statistics* **38**(6), 3696–3723.

Cook, R. D. (1994), Using dimension-reduction subspaces to identify important inputs in models of physical systems, *in* 'Proceedings of the section on Physical and Engineering Sciences', American Statistical Association Alexandria, VA, pp. 18–25.

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', *The Annals of Statistics* **32**(3), 1062–1092.

Cook, R. D., Li, B. & Chiaromonte, F. (2007), 'Dimension reduction in regression without matrix inversion', *Biometrika* **94**(3), 569–584.

Cook, R. D. & Weisberg, S. (1991), 'Comment', *Journal of the American Statistical Association* **86**(414), 328–332.

Eaton, M. L. (1983), *Multivariate Statistics: A Vector Space Approach*, Wiley, New York.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American statistical Association* **96**(456), 1348–1360.

Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J. et al. (2012), 'Systematic identification of genomic markers of drug sensitivity in cancer cells', *Nature* **483**(7391), 570–575.

Gregg, J. & Fraizer, G. (2011), 'Transcriptional regulation of EGR1 by EGF and the ERK signaling pathway in prostate cancer cells', *Genes & Cancer* **2**(9), 900–909.

Harada, T., Morooka, T., Ogawa, S. & Nishida, E. (2001), 'Erk induces p35, a neuron-specific activator of Cdk5, through induction of Egr1', *Nature Cell Biology* **3**(5), 453–459.

Izenman, A. J. (1975), 'Reduced-rank regression for the multivariate linear model', *Journal of Multivariate Analysis* **5**(2), 248–264.

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

Li, B. & Wang, S. (2007), 'On directional regression for dimension reduction', *Journal of the American Statistical Association* **102**(479), 997–1008.

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

Li, K.-C. (2000), 'High dimensional data analysis via the SIR/PHD approach'.

Li, L. & Li, H. (2004), 'Dimension reduction methods for microarrays with application to censored survival data', *Bioinformatics* **20**(18), 3406–3412.

Li, L. & Yin, X. (2008), 'Sliced inverse regression with regularizations', *Biometrics* **64**(1), 124–131.

Liu, H. & Zhang, J. (2009), Estimation consistency of the group lasso and its applications, *in* 'International Conference on Artificial Intelligence and Statistics', pp. 376–383.

Long, Y. C., Cheng, Z., Copps, K. D. & White, M. F. (2011), 'Insulin receptor substrates Irs1 and Irs2 coordinate skeletal muscle growth and metabolism via the Akt and AMPK pathways', *Molecular and Cellular Biology* **31**(3), 430–441.

Luo, H., Yanagawa, B., Zhang, J., Luo, Z., Zhang, M., Esfandiarei, M., Carthy, C., Wilson, J. E., Yang, D. & McManus, B. M. (2002), 'Coxsackievirus B3 replication is reduced by inhibition of the extracellular signal-regulated kinase (ERK) signaling pathway', *Journal of Virology* **76**(7), 3365–3373.

Ma, Y. & Zhu, L. (2013), 'A review on dimension reduction', *International Statistical Review* **81**(1), 134–150.

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

Wang, T., Zhao, H., Chen, M. & Zhu, L. (2015), 'Supplement to "Model-free dimension reduction and variable selection in high-dimensional regression"'.

Wang, X., Li, G., Hibshoosh, H. & Halmos, B. (2012), 'Phlda1/2 contribute to tumor suppression in breast and lung cancer as downstream targets of oncogenic HER2 signaling', *Cancer Research* **72**(8 Supplement), 20–20.

Wu, Y. & Li, L. (2011), 'Asymptotic properties of sufficient dimension reduction with a diverging number of predictors', *Statistica Sinica* **2011**(21), 707–730.

Yin, X. (2010), *Sufficient dimension reduction in regression*, Book chapter in "The Analysis of High-dimensional Data" (Eds. X. Shen and T. Cai.). World Scientific, New Jersey.

Yin, X. & Hilafu, H. (2015), 'Sequential sufficient dimension reduction for large p, small n problems', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(4), 879–892.

Yin, X., Li, B. & Cook, R. D. (2008), 'Successive direction extraction for estimating the central subspace in a multiple-index regression', *Journal of Multivariate Analysis* **99**(8), 1733–1757.

Yu, Z., Zhu, L., Peng, H. & Zhu, L. (2013), 'Dimension reduction and predictor selection in semiparametric models', *Biometrika* **100**(3), 641–654.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

Zheng, Y., Zhang, C., Croucher, D. R., Soliman, M. A., St-Denis, N., Pasculescu, A., Taylor, L., Tate, S. A., Hardy, W. R., Colwill, K. et al. (2013), 'Temporal regulation of EGF signalling networks by the scaffold protein Shc1', *Nature* **499**(7457), 166–171.

Zhong, W., Zeng, P., Ma, P., Liu, J. S. & Zhu, Y. (2005), 'Rsir: regularized sliced inverse regression for motif discovery', *Bioinformatics* **21**(22), 4169–4175.

Zhu, L., Wang, T., Zhu, L. & Ferré, L. (2010), 'Sufficient dimension reduction through discretization-expectation estimation', *Biometrika* **97**(2), 295–304.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.