4-27-2018 10:30 AM

Analysis Challenges for High Dimensional Data

Bangxin Zhao, *The University of Western Ontario*

Supervisor: He, Wenqing, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Statistics and Actuarial Sciences

Abstract

The practical and theoretical challenges posed by the 'large p, small n' settings are important issues in contemporary statistics. In this thesis, we propose new methodologies that target three different areas of high-dimensional statistics: variable screening, influence measure and post-selection inference.

Variable screening is a general procedure in high dimensional data analysis to ensure the applicability of statistical methods. Typically marginal correlation between the response and each predictor are employed for this role. It is a complicated and computationally burdensome procedure since spurious correlations commonly exist among predictor variables, and important predictor variables may not have large marginal correlations with the response variable. We propose a new estimator for the correlation between the response and high-dimensional predictor variables, and based on the estimator we develop a new screening technique termed *Dynamic Tilted Current Correlation Screening* (DTCCS) for high dimensional variables screening. DTCCS is capable of picking up the relevant predictor variables within a finite number of steps. The DTCCS method includes the widely used *sure independence screening* (SIS) method and the *high-dimensional ordinary least squares projection* (HOLP) approach as special cases. The DTCCS technique has sure screening and consistency properties which are demonstrated theoretically and numerically and illustrated through a real-life example.

Two methods of high-dimensional influence measure have also been explored. They are from the perspective of the extreme value distribution (EVD) and the robustness of design respectively. For the first method, EVD-type statistics have been shown to be powerful in measuring high-dimensional influence theoretically and numerically. From the second method, we propose *Hellinger distance for high-dimensional influence measure* (HD-HIM). The inner product of two transformed influence functions is used to measure the Hellinger distance of two discrete distribution functions from the whole and deleted dataset. This construction gives power to flag the observations

that have unusual effect on high-dimensional models. The HD-HIM method has been illustrated theoretically and numerically.

Lastly, we propose a post-selection inference method termed *Cosine PoSI* that is numerically feasible in a high-dimensional framework. *Cosine PoSI* focus on the geometric aspect of *Least Angle Regression* (LARS). LARS efficiently provides a solution path along which the entered predictors always have the same absolute correlation with the current residual. At each step of the LARS algorithm, the proposed *Cosine PoSI* method employs an angle from the correlation between the entering variable and current residual and considers this angle as a random variable from the cosine distribution. The post-selection inference is then conducted based on the order statistics of this cosine distribution. Given the collection of the possible angles, hypothesis tests are performed on the limiting distribution of the maximum angle. To confirm the effectiveness of the proposed method, we conduct simulation studies and a real-life data analysis to illustrate the usefulness of this post-selection method.

KEYWORDS: High-dimensional statistics, variable screening, deterministic design matrix, influence measure, post-selection inference.

Acknowledgements

This dissertation could not have been possible without the help and the support of several individuals. First and foremost, I would like to express my deepest gratitude to my beloved supervisor, Dr. Wenqing He, for his immense support and extremely valuable guidance throughout my Ph.D. study. Without his consistent and illuminating instruction, I would not be where I am today. I not only learned new statistical knowledge from him, but also learned how to conduct research with creativeness and better vision. His enthusiasm to help his students will always be a good example and guide me in my future.

I would also like to express my heartfelt gratitude to the examiners of my thesis defense committee, Dr. John Koval, Dr. Jiandong Ren, Dr. Douglas Woolford and Dr. Mu Zhu, for their time, insightful questions, valuable comments, suggestions and discussions on my research work. I want to thank Dr. Reg Kulperger for sharing his vast knowledge and giving suggestions during my thesis and proposal writing.

I owe my honest gratitude to Dr. David R. Bellhouse, Dr. Ian McLeod, Dr. Serge Provost, Dr. Hao Yu and Dr. Ricardas Zitikis. I also want to thank all the faculty members and staff in our department for their invaluable help. Their kindness and help during these wonderful years at the University of Western Ontario is appreciated.

Many special thanks go in particular to Dr. Grace Y. Yi for sharing her valuable experience in scientific writing and paper preparation.

I want to thank Dr. Wenqing He again, and Dr. Rohana J. Karunamuni, Dr. Qingguo Tang for being great co-authors.

I am deeply grateful for all the fun time I had with my friends during my Ph.D. study. I thank Xin Liu, Jiang Wu, Wenjun Jiang for being great officemates and for all the interesting discussions and sports time. I appreciate all the cultural discussions and game play with Wisdom Stallone Avusuglo and Hossein Zareamoghaddam. I thank Dexen Xi, Chen Yang, Guandong Zhang, Kexin Luo for the R discussions.

It is the dream of everyone to play an important role on the stage surrounded by loyal audiences. However, our sense of security, hopefulness, confidence, and self-worth are built by members of our family who can hold our hands tightly even in the

midst of despondency, insecurity, and disappointment. I consider myself a lucky fellow to have such supportive family members. Especially, my wife, Zhe (Aggie) Yu, and my parents whose unwavering support and love these years cannot be overemphasized. I will forever be grateful.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation and Purpose of the Dissertation

With rapid development in technologies, a growing number of research fields encounter data with unprecedented size and complexity, such as researches in artificial intelligence, economy, finance, biology, genetics, engineering and astronomy. The importance of data and the vitality of data analysis cannot be downplayed in contemporary science. As computational power increases and the expense of data collection and processing decrease significantly, the dimension of datasets is continuously becoming large. In those dataset, the dimension of predictor variables p can be as large as or much larger than the sample size n, but very often, among thousands of available predictor variables only a small number of them are informative and it is critically important to identify them correctly. High-dimensional data analysis has received a tremendous of attention recently. Seminal theories of *Least Angle Regression* (LARS, Efron et al. 2004) and *Sure Independence Screening* (SIS, Fan and Lv 2008) both proposed to use correlation between predictor variables and response (or current residual) to solve high-dimensional problems. The high-dimensional correlation can be viewed as a counterpart to ordinary least square (OLS) estimator of the parameter and many data-driven methods based on correlation have been studied for years in high dimensional statistics. In this dissertation, we develop new methodologies and techniques by centering on correlation learning for high-dimensional sparse modelling.

The methodologies proposed in this dissertation is aiming to solve but not limited to the following high-dimensional problems:

Example 1.1.1. Hastie et al. (2009): Microarrays gene expression data

Microarrays gene expression data is one of the classical high-dimensional data types. DNA microarrays measure the expression of a gene in a cell by measuring the amount of mRNA present for that gene. A gene expression data set collects together the expression values from a sequence of DNA microarray experiments, with each column representing an experiment. There are therefore several thousand (p) rows representing individual genes and tens (n) of columns representing samples.

Typical questions about microarray data: certain genes show abnormal expression for certain cancer sample; certain genes are more important in a certain disease and et cetera. Traditional statistical methods can not be directly applied to answer those questions.

Example 1.1.2. Biba and Xhafa (2011): High-dimensional text regression

The design matrix of the bag-of-words (BOW) model consists rows of high dimensional vector whose elements are the frequency of words. The BOW model has been widely applied in machine learning topics such as email filtering. For more details see Biba and Xhafa (2011). Statistical diagnostic techniques can be also contribute to these problems. We measure the influence of the high-dimensional observations (the email) and expect to automatically flag the email category and give warning to a suspicious email.

Besides the above two examples, other high-dimensional problems to which the methods developed in this thesis could be applied are image recognition (pixels of the high resolution images are large); spatial correlation of home prices (up to 1 million spatial parameters), retailer real-time pricing (for millions of items), amongst others.

In the rest of this chapter, we provide a review of the literature on the relevant topics covered in this thesis, which include matrices with applications in statistics, development in high-dimensional sparse modelling and estimation, robust statistics and high-dimensional influence measure.

1.2 The High-dimensional Design Matrix

Over the past decade, advancement of new technologies in the fields of the natural and social sciences have improved data collection procedures. This has led to the problem of high-dimensional data analysis which links to the idea of a complicated large design matrix, denoted $\mathbf{X}$. For this $n \times p$ design matrix, the number of predictor variables, p, is either on the same order of, or much greater than, the number of observations, n. For instance, data ascertained from spectra, biomedical imaging, high-frequency finance and DNA micro-arrays can be of high-dimension. The traditional methods that perform well for low-dimensional data run into many severe problems in analyzing such a high-dimensional dataset. The common issues that arise in analyzing a high-dimensional dataset by using traditional methods include: the non-invertibility of the matrix $\mathbf{X}^T\mathbf{X}$, the high correlation among predictors in the model, the non-existence of the inverse covariance matrix (precision matrix), amongst others.

The high-dimensional dataset with $p \approx n$ or $p > n$ can be divided into two cases: *high dimension* and *ultra-high dimension*. If the dimensionality p grows polynomially with the sample size n, i.e., $p = O(n^{\alpha})$ for some $\alpha > 0$, we call it *high dimension*; if the dimensionality p grows non-polynomially with the sample size n, i.e., $p = O(e^{n^{\iota}})$ for some $\iota \in (0, 1)$, we call it *ultra-high dimension* or *non-polynomial (NP) dimensionality* (Fan and Lv 2008; Shao and Deng 2012).

We begin with the most important and commonly used regression model, the classical linear regression model. Linear regression model investigates the relationship between a continuous dependent variable (normally referred to as the response variable), and at least one explanatory variable (also known as predictor or covariate). In classical statistical model setting, the number of observations is typically denoted as n, while the number of predictors in a model (referred to as the dimension of the model) is denoted as p. For subject i in a sample of n individuals, let y_i be the response variable and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ be the p dimensional predictors. We write $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^T$ for the response vector of a sample with n subjects,

and $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$ be the $n \times p$ design matrix including the p dimensional predictors for n subjects. In this thesis, the subset columns or rows of the design matrix are frequently used. $\tilde{\mathbf{X}}_{-j}$ denotes the submatrix of deleting the jth predictor variable, $\mathbf{X}_j$, $j = 1, \ldots, p$. $\mathbf{X}_{(-i)}$ denotes the submatrix of deleting the ith observation, $\mathbf{x}_i$, $i = 1, \ldots, n$. To ease the notations, we use $\mathbf{X}_j$, $j = 1, \ldots, p$, for the jth predictor variable and the its realization in the design matrix. The relationship between the response $\mathbf{y}$ and the predictor variables $(\mathbf{X}_1, \ldots, \mathbf{X}_p)^T$ is given by

$$\mathbf{y} = \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \cdots + \beta_p \mathbf{X}_p + \epsilon, \tag{1.1}$$

where ϵ is the random error. Alternatively, this classical model can be written with realization in sample size n,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of the coefficients and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise term. Usually, we assume $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$. Alternatively, $\mathbf{X}$ can be considered as a row of column vectors: $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p)$, where $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})^T$ for $j = 1, \ldots, p$. Let $row(\mathbf{X})$ be the linear $p-$dimensional space which is spanned by the row vectors of $\mathbf{X}$ and $col(\mathbf{X})$ be the linear $n-$dimensional space which is spanned by the column vectors of $\mathbf{X}$.

Now, let $\mathbf{x}$ be a p dimensional random vector with multivariate distribution with mean $\mu_{p\times 1}$ and covariance $\Sigma_{p\times p}$ defined as follows:

$$E(\mathbf{x}) = \mu,\ cov(\mathbf{x}) = \Sigma.$$

In traditional statistics, if μ and the covariance matrix Σ are unknown, one can estimate μ and Σ from the sample. These estimates are known as the sample mean and sample covariance respectively.

Let $\bar{\mathbf{x}}$ and S denote the sample mean and sample covariance matrix, respectively. Define $\mathbf{1}_n = (1, \ldots, 1)^T$, an $n \times 1$ vector of ones, so that we have

$$\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}^T 1_n \tag{1.3}$$

and

$$S = \frac{1}{n}(\mathbf{X} - \bar{\mathbf{X}})^T(\mathbf{X} - \bar{\mathbf{X}}), \tag{1.4}$$

where $\bar{\mathbf{X}}$ is a $n \times p$ matrix with each row comprised of $\bar{\mathbf{x}}^T$. It must be noted here that $\bar{\mathbf{x}}$ and S are unbiased and consistent estimators for μ and Σ respectively. The sample covariance matrix S is a good estimator of the population variance if $n \gg p$, but it performs poorly when p is close to or larger than n (Cai et al. 2016). In the high-dimensional context, the estimation of the precision matrix ($\Omega = \Sigma^{-1}$, the inverse of the covariance matrix) is also a difficult and computational complex question. Cai et al. (2011) proposed *constrained l_1-minimization for inverse matrix estimation* (CLIME) to directly calculate Ω by an optimization problem

$$\min \|\Omega\|_1 \text{ subject to } |S\Omega - \mathbf{I}_p|_\infty \leq \lambda_n, \tag{1.5}$$

where $\|\cdot\|_1$ is the elementwise L_1 norm ($\|\Omega\|_1 = \sum_{i,j} |\Omega_{i,j}|$), $\|\cdot\|_\infty$ is the matrix elementwise infinity norm ($\|\Omega\|_\infty = \max_{1 \leq i,j \leq p} |\Omega_{i,j}|$), and $\lambda_n = \frac{c\log(p)}{n}$ for some sufficiently large constant c. This method has been built in the R package **clime**, but it is still a time-consuming computing process to obtain the estimated precision matrix in high-dimensional statistics.

In the random design setting for linear regression, each pair $(\mathbf{x}_i^T, \mathbf{y}_i)$ is the observation sampled from the population, where random vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ and random variable $\mathbf{y}_i \in \mathbb{R}^1$. If the design matrix $\mathbf{X}$ in Eq. (1.2) consists of random vectors, we call $\mathbf{X}$ a random design matrix. The random $\mathbf{x}_i$'s are usually assumed to be independent identically distributed (i.i.d.) and independent of ϵ_i's, and $\hat{\boldsymbol{\beta}} = [\widehat{cov(\mathbf{x}_i)}]^{-1}\widehat{cov(\mathbf{x}_i, \mathbf{y}_i)}$.

The fixed design setting is the opposite of the random design setting, and the design matrix in this setting is called deterministic design matrix. Let $\tilde{\mathbf{X}}_{-j}$ be the submatrix of $\mathbf{X}$ which excludes the column $\mathbf{X}_j$, $\mathbf{X}_j^\perp$ be the projection of $\mathbf{X}_j$ to the

orthogonal complement of the column space of $\tilde{\mathbf{X}}_{-j}$. By using a deterministic design matrix, the least square estimator can be expressed as $\hat{\beta}_j = (\mathbf{X}_j^{\perp T} Y)/(\mathbf{X}_j^{\perp T} \mathbf{X}_j)$ for a linear model without intercept, where $\mathbf{X}_j^{\perp T} \mathbf{X}_j \neq 0$ for $n > p$ (Zhang and Zhang 2014). These two settings of the design matrix bring two views of parameter estimation: a probabilistic one and a nonprobabilistic one. The goal of both views is to find coefficients $\hat{\boldsymbol{\beta}}$ such that the expected prediction error on a new observation from the population is small enough. For the past two decades, statisticians extended those two views to high-dimensional data, and developed many contemporary methodologies and techniques for large random or deterministic design matrix, see details in Fan and Lv (2008), Shao and Deng (2012), Lv (2013), Zhang and Zhang (2014) and Wang and Leng (2016).

The column space (also called the range or image) of a design matrix $\mathbf{X}$ is commonly used in parameter estimation in the case of $n > p$. The ordinary least squares (OLS) estimate projects the response Y onto the linear space $col(\mathbf{X})$ which is spanned by columns of $\mathbf{X}$. Due to lack of sufficient degrees of freedom, OLS is no longer feasible for high-dimensional statistics. This motivates the idea of variable screening, i.e., to obtain a subset of features that have significant impact on the response before building a formal statistical model. In contrast with column space of $\mathbf{X}$, row space of $\mathbf{X}$ has been studied recently under high-dimensional setting. Shao and Deng (2012) proposes an approach to project the parameter vector $\boldsymbol{\beta}$ onto the linear space $row(\mathbf{X})$ which is spanned by the rows of $\mathbf{X}$ and show that this projection of $\boldsymbol{\beta}$ can discriminate large and small elements efficiently by choosing a proper thresholding value.

For $p > n$, considering the ridge regression estimator of $\boldsymbol{\beta}$ (Hoerl and Kennard 1970) under model (1.2),

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^T Y, \tag{1.6}$$

where $\lambda > 0$ is an appropriately chosen regularization parameter. Shao and Deng (2012) and Wang and Leng (2016) show that the computation of $\hat{\boldsymbol{\beta}}_{ridge}$ involves only inverting an $n \times n$ matrix since $(\mathbf{X}^T \mathbf{X} + \lambda I_p)^{-1} \mathbf{X}^T = \mathbf{X}^T (\mathbf{X}\mathbf{X}^T + \lambda I_n)^{-1}$ which implies that the ridge regression estimator $\hat{\boldsymbol{\beta}}_{ridge}$ is always in $row(\mathbf{X})$.

1.3 High-dimensional Linear Regression

1.3.1 Penalized Regression

Due to rapid development of technological advances, modern scientific research very often encounters datasets with unprecedented size and complexity, such as datasets in genomics, oncology imagery and finance. In practice, it is common to have huge number of variables for predicting a particular phenomenon or outcome. Suffering from high dimensionality, variable selection, which is vitally important in statistical modelling, encounters a big challenge. Many classical variable selection methods, for instance, backward elimination, forward selection, stepwise selection, all subsets selection, may be very computationally expensive or even infeasible. Missing relevant predictors and/or including irrelevant predictors in a statistical model will decrease model's predictive ability and/or increase the difficulty of model interpretation.

The circumvention of the above problem has led to the idea of the *penalized regression*. We give some basic notation before introducing some popular penalties which have been successfully applied to achieve variable selection. For any $p-$dimensional vector $\mathbf{a}$, $\|\mathbf{a}\|_0 = \sum_{j=1}^{p} \mathbb{I}(a_j \neq 0)$, $\|\mathbf{a}\|_\infty = \max_{1\leq j\leq p} |a_j|$ and $\|\mathbf{a}\|_q = (\sum_{j=1}^{p} |a_j|^q)^{1/q}$ for $q \geq 1$.

In the regularization framework, consider a sample $\{(\mathbf{x}_i^T, \mathbf{y}_i)^T, i = 1, \ldots, n\}$ of size n from an unknown population, where $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{y}_i \in \mathbb{R}^1$. Taking the square loss function, we can select variables by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda J(\boldsymbol{\beta}), \tag{1.7}$$

where λ is a non-negative tuning parameter, $J(\cdot)$ is a penalty function which is positive valued for $\boldsymbol{\beta} \neq 0$. A popular choice of the penalty function $J(\boldsymbol{\beta})$ is the L_q norm of the parameters to the qth power (Tibshirani 1996, Zou and Hastie 2005),

$$J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_q^q = \sum_{j=1}^{p} |\beta_j|^q,\ q \geq 0. \tag{1.8}$$

Hoerl and Kennard (1970) proposed *ridge regression* by using $q = 2$ in equation

(1.8). Ridge regression is very similar to least squares, except that the coefficients are estimated by minimizing a slightly different quantity,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{1.9}$$

where $\lambda \geq 0$ is a tuning parameter. Eq.(1.9) is equivalent to the Lagrangian problem which minimize $\|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2$ subject to $\|\boldsymbol{\beta}\|_2^2 \leq t$, where t is a non-negative tuning parameter. Ridge regression improves the OLS by shrinking all coefficients towards zero, but it will still include all p predictors in the final model unless $\lambda = \infty$. Regular ridge regression shrinks the variables, but does not select the variables. Shao and Deng (2012) propose the thresholded ridge regression which uses a threshold value to select variables from the ridge solution. For the columnwise normalized $\mathbf{X}$, the estimates solution to the ridge regression is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ridge} &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \frac{1}{1+\lambda}\begin{pmatrix} 1 & \frac{\hat{\rho}_{12}}{1+\lambda} & \cdots & \cdots & \frac{\hat{\rho}_{1p}}{1+\lambda} \\ \frac{\hat{\rho}_{21}}{1+\lambda} & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & \frac{\hat{\rho}_{p-1,p}}{1+\lambda} \\ \frac{\hat{\rho}_{p1}}{1+\lambda} & \cdots & \cdots & \frac{\hat{\rho}_{p,p-1}}{1+\lambda} & 1 \end{pmatrix}^{-1} \mathbf{X}^T\mathbf{Y},
\end{aligned} \tag{1.10}$$

where $\hat{\rho}_{ij} = corr(\mathbf{X}_i, \mathbf{X}_j)$, the sample correlation. The off-diagonal elements of the correlation matrix $\mathbf{X}^T\mathbf{X}$ are shrunk by the factor $\frac{1}{1+\lambda}$, which was termed as *decorrelation* by Zou and Hastie (2005). For the special orthonormal design case: $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ where $\mathbf{X}$ is the $n \times p$ design matrix, we can check that ridge regression solution is $\frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}_{ols}$ where $\hat{\boldsymbol{\beta}}_{ols}$ is the ordinary least squares solution. For the non-orthonormal case, see details in Hoerl and Kennard (1970).

Tibshirani (1996) is the fundamental paper about *Least Absolute Shrinkage and Selection Operator* (LASSO) by using L_1 penalty which uses $q = 1$ in equation (1.8),

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{1.11}$$

where $\lambda \geq 0$ is a tuning parameter.

LASSO shrinks some coefficients and sets others to 0. Hence, LASSO retains the good shrinkage feature of ridge regression and selects variables simultaneously. Comparing Eq. (1.11) to Eq. (1.9), we see that the LASSO and ridge regression have similar formulations. The only difference is that the LASSO uses an L_1 penalty instead of an L_2 penalty. The theoretical properties of LASSO have been well studied in the literature, see detail in Zhao and Yu (2006), Zhang and Huang (2008), Meinshausen and Yu (2009), Bickel et al. (2009), Lockhart et al. (2014) and Lee et al. (2016). LASSO contributed to the rich literature on the path-based regression methods. The solution path based on those methods potentially make the high-dimensional variable screening possible. Regardless of false discoveries, the coefficients selected by the path-based regression algorithms contains the uniquely defined true model with large probability. If false discovery is taken into consideration, Li and Barber (2017) proposed a family of 'accumulation tests' to efficiently control the false discovery rate (FDR) on the high-dimensional solution path.

Through the generalized L_1 penalties, extensions and modified versions of LASSO have been suggested and studied for the past two decades, examples include adaptive LASSO (Zou 2006), random LASSO (Wang et al. 2011) and generalized LASSO (Tibshirani and Taylor 2011). Those generalized L_1 penalties arise in a wide variety of areas such as microarray studies and image denoising. By combining a squared L_2 penalty with the L_1 penalty, the *elastic net* was proposed by Zou and Hastie (2005). The *elastic net* method uses a linear combination of squared L_2 and L_1 penalties on the regression coefficients and aims to achieve the grouping effect that highly correlated feathers will be in or out of the model together. Elastic net can be formulated as the following penalized least squares problem,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2, \tag{1.12}$$

where λ_1, $\lambda_2 \geq 0$ are tuning parameters which must be chosen in advance.

Efron et al. (2004) propose the *least angle regression* (LARS) algorithm with a modification that can efficiently compute the LASSO solution path. The LARS algorithm is highly related to the traditional *Forward Stepwise Regression* (FR) and

Forward-Stagewise Regression (FSR), but it uses a novel solution direction and step size for each iteration. LARS can be considered as both a variable screener and a model selector. The advent of LARS creates an era of correlation learning which plays an important role in high-dimensional statistics for years. The importance of correlation learning and the detail of the LARS algorithm will be introduced in Section 1.3.2.

To achieve an unbiased, sparse and continuous estimator, Fan and Li (2001) designed a *smoothly clipped absolute deviation* (SCAD) penalty function $J_\lambda(\beta)$ with derivative satisfying

$$J'_\lambda(t) = \lambda \left\{ \mathbb{I}(t \leq \lambda) + \frac{(a\lambda - t) \cdot \mathbb{I}(a\lambda > t)}{(a-1)\lambda} \cdot \mathbb{I}(t > \lambda) \right\}, \tag{1.13}$$

for $t = |\beta|$ and some $a > 2$.

1.3.2 Correlation Learning

Forward-type Regression

Marginal correlation between the individual covariates and response (or current residual) plays a critical role in both low dimensional and high-dimensional data analysis. In low dimensional data analysis, the solution path of *Forward Stepwise Regression* (FR) and *Forward-Stagewise Regression* (FSR) are both iteratively calculated by picking the variable which has the largest absolution correlation with current residual. In high-dimensional data analysis, a vast amount of literature on correlation research has been done in recent years, including the LARS algorithm (Efron et al. 2004), the SIS method (Fan and Lv 2008), the *tilting* procedure (Cho and Fryzlewicz 2012), and *High-dimensional Ordinary Least squares Projection* (HOLP, Wang and Leng 2016).

Comparing with the step size at each iteration, FR is an aggressive fitting technique and it reaches the OLS solution (which is the longest step size) at each iteration, while FSR is a conservative fitting technique which uses thousands of tiny moving to obtain the final model. Hastie et al. (2009) describe the FSR as: starting with no variables in the initial model, i.e. denoting mean function $\hat{\mu}_1 = 0$, initial residual

$Z_1 = Y - \hat{\mu}_1$, then the initial marginal correlation is

$$\hat{\mathbf{c}}_1 = \mathbf{c}(\hat{\mu}_1) = X^T(Y - \hat{\mu}_1). \tag{1.14}$$

Then select variable X_{j1} which has the largest absolute correlation with the response (the current residual vector) Y, and the corresponding marginal correlation is $\hat{C}_1 = \|\hat{\mathbf{c}}_1\|_\infty$, $s_{j1} = sign\{X_{j1}^T Y\}$.

The first step is a construction of simple linear regression of Y on X_{j1} and it leaves a residual vector orthogonal to X_{j1}. After the first step, update the mean function to

$$\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_1 \cdot s_{j1} \cdot X_{j1}, \tag{1.15}$$

where $\hat{\gamma}_1$ is a 'small' constant ('small' is compared to the 'big' choice of $\hat{C}_1$ in FR), then select X_{j2} which has the largest absolute correlation between the variables and the current residual vector $Z_2(= Y - \hat{\mu}_2)$. After kth step, add the predictor $X_{j\,k+1}$ which is most correlated with the $(k+1)$th residual vector $Z_{k+1}(= Y - \hat{\mu}_{k+1})$ to the model. Stop the algorithm at the kth step if the rest predictors have negligible correlation with the current residual vector Z_k.

Similar to FSR, LARS starts with no variables in the initial model, i.e. the active model set $\mathcal{M}_0 = \{\emptyset\}$. Let $\mathbf{c}(\hat{\mu}_k)$ be the correlation vector of variables and current residual at the kth stage

$$\hat{\mathbf{c}}_k = \mathbf{c}(\hat{\mu}_k) = X^T Z_k = X^T(Y - \hat{\mu}_k),\ k = 1, 2 \ldots, p. \tag{1.16}$$

At the first stage, LARS selects variable X_{j1} which has the biggest correlation with the initial residual $Z_1 = Y$, then LARS solution path takes the direction of $u_1 = X_{j1}$ for a step size $\hat{\gamma}_1$ until some other predictor, say X_{j2}, has the same correlation with the current residual Z_2, i.e. $|\langle X_{j1}, Z_2\rangle| = |\langle X_{j2}, Z_2\rangle|$. Then LARS solution path takes the direction u_2 which bisects X_{j1} and X_{j2} with step size $\hat{\gamma}_2$ until a third variable comes into the model, i.e. $|\langle X_{j1}, Z_3\rangle| = |\langle X_{j2}, Z_3\rangle| = |\langle X_{j3}, Z_3\rangle|$.

At the beginning of the stage k, we have $k-1$ of the variables in the model. We are going to select variable X_{jk} which has the largest absolute correlation with the current residual vector Z_k, and the corresponding marginal correlation is $\hat{C}_k = \|\hat{\mathbf{c}}_k\|_\infty$, $s_{jk} = sign\{X_{jk}^T Z_k\}$. LARS process terminates when $k = \min(n-1, p)$.

Sure Independence Screening

The SIS method of Fan and Lv (2008) ranks the absolute value of the marginal correlations $\boldsymbol{\omega} = |\mathbf{X}^T\mathbf{Y}| = (\omega_1, \ldots, \omega_p)^T$ to choose the variables to be kept in the model. Here, $\boldsymbol{\omega}$ is essentially a vector of marginal correlation between the response and all predictor variables. For any given $\delta \in (0, 1)$, Fan and Lv (2008) sorted the p componentwise magnitudes of the vector $\boldsymbol{\omega}$ in a decreasing order and defined a submodel

$$\mathcal{M}_\delta = \{1 \leq j \leq p : \, |\omega_j| \text{ is among the first } [\delta n] \text{ largest } |\omega_j|'s\},$$

where $[\delta n]$ denotes the integer part of δn. This is a straightforward way to shrink the full model $\mathcal{F}$ to a submodel $\mathcal{M}_\delta$ with size $|\mathcal{M}_\delta| < n$. The SIS method uses each variable independently to evaluate its correlation with the response and filters out the variables which have weak marginal correlations with the response variable. The SIS method is different from the regularized regression as it does not use penalties to shrink the estimator, but measures the importance of each predictor variable by its marginal correlation with the response variable. Due to its independence screening property, the screening can be implemented even when p grows exponentially with the sample size n, i.e., $p = O(e^{n^\iota})$ for some $\iota \in (0, 1)$. This property led to SIS method receiving a large amount of attention in ultra-high dimensional data analysis. Similar to the Forward-type regression, Fan and Lv (2008) also use an iterative SIS (ISIS) to screen variables by ranking the correlation between candidate variables and the current residual for several steps. By using ISIS, important variables that have small marginal correlation but jointly correlated with the response can be saved since it can be evaluated again during the next round by using the updated residual. Wang (2009) used forward regression to find a solution path to reach the minimum residual sum of square (RSS) at each step, and that variable screening method can also identify all relevant predictors consistently.

One of the biggest problems one may encounter in high-dimensional variable screening is the presence of high (most likely spurious) correlations among the predictor variables. Fan and Lv (2008) showed the maximum spurious correlation among

covariates can be large (see Example 1.3.1) due to the increasing dimensionality. Spurious correlation easily brings the fact that an unimportant predictor can be highly correlated with the response variable due to the presence of important predictors associated with that predictor. To circumvent this problem, Cho and Fryzlewicz (2012) discussed the idea of 'tilting' which uses an iterative procedure to reevaluate the importance of predictors. Besides the spurious correlations among the predictors, the multicollinearity arises when the number of predictor variables becomes comparable or much larger than the number of observations. (Belsley et al. 1980).

Example 1.3.1. Spurious Correlation (Fan and Lv 2008)

Let $x_1, x_2, \ldots, x_n$ be n independent observations of a p-dimensional Gaussian random vector $\mathbf{X} = (X_1, \ldots, X_p)^T \sim N_p(0, \mathbf{I}_p)$. Repeatedly simulate the data with $n = 60$ and $p = 1000, 5000$ for 1000 times. Consider the empirical distribution of the maximum absolute sample correlation coefficient between the first variable with the remaining ones defined as

$$\hat{r} = \max_{2 \leq j \leq p} |\hat{Corr}(X_1, X_j)|.$$

From Figure 1.1, we can see even though X_1 and X_j $(2 \leq j \leq q)$ are independently simulated, the maximum correlation between X_1 and other variables can still be very high in high dimensional data. Figure 1.1 shows that the absolute values of maximum correlations even under independent assumption can be at least 0.4 for the case of $p = 5000$ and at least 0.35 for the case of $p = 1000$, which are both non-negligible. Due to presence of spurious correlation, the independence marginal correlation screening may be violated.

Column normalization is very popular in high-dimensional data analysis, such as techniques in Efron et al. (2004), Fan and Lv (2008), Wang (2009), Cho and Fryzlewicz (2012), Wang and Leng (2016) and Fan et al. (2018). After the normalization of the column of $\mathbf{X}$, each columns of $\mathbf{X}$ has a unit norm. We assume error ϵ_i, $i = 1, \ldots, n$ are independent and identically distributed (iid) random noise following a normal

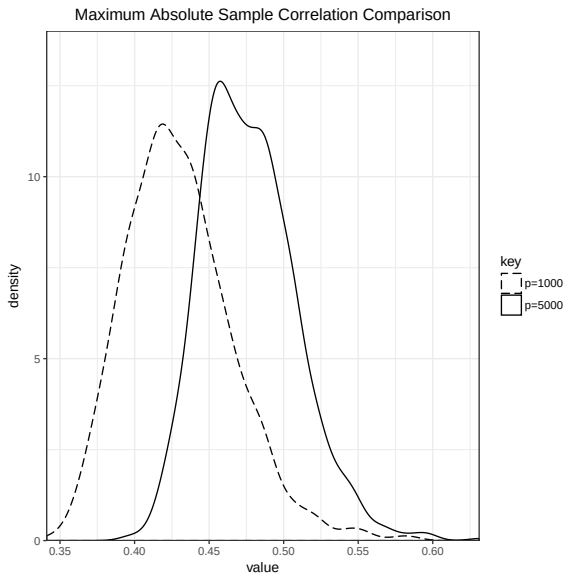

Figure 1.1: Distribution of the maximum absolute sample correlation coefficients between X_1 and $\{X_j\}_{j\neq 1}$ when $n = 60$; $p = 1000$ (dashed curve) and $n = 60$; $p = 5000$ (solid curve).

distribution $N(0, \sigma^2)$ with $\sigma^2 < \infty$. The marginal correlation between each variable X_j and the response Y has the decomposition

$$\mathbf{X}_j^T\mathbf{Y} = \mathbf{X}_j^T(\sum_{k=1}^{p}\beta_k\mathbf{X}_k + \boldsymbol{\epsilon}) = \beta_j + \sum_{k\neq j}\beta_k\mathbf{X}_j^T\mathbf{X}_k + \mathbf{X}_j^T\boldsymbol{\epsilon}. \tag{1.17}$$

The signal-to-noise ratio (SNR) is defined as $SNR = \frac{\beta^T\Sigma\beta}{\sigma^2}$ where Σ is the covariance matrix of the random vector $\mathbf{x}$ (Wang et al. 2011). If the SNR is assumed sufficiently high, for instance, $SNR \geq 10$, then the third term of the above decomposition is negligible compared to the first two terms. The second term of the above decomposition $\sum_{k\neq j}\beta_k X_j^T X_k$ shows that (a) unimportant variables that are highly correlated with the important variables will have a high chance to be selected; (b) an important variable can be marginally uncorrelated but jointly correlated with the response; (c) collinearity can exist among the variables in high-dimensional data. Hence, minimizing the effect of $\sum_{k\neq j}\beta_k X_j^T X_k$ is critically important in high-dimensional screening problem. Recent development in dealing with correlated data can be found in Wang et al. (2011), Cho and Fryzlewicz (2012), Jin and He (2016), for example.

Cho and Fryzlewicz (2012) proposed a new tilting procedure which can efficiently reduce the high correlations (possibly spurious) between the predictor variables in high dimensional data. This method is tilting each column X_j to $X_j^\star$ such that the tilted correlation between $X_j^\star$ and X_k is reduced to 0 or negligible and thus the relationship between the jth covariate and the response can be identified more accurately. For standardized $\mathbf{X}$, denote the sample correlation matrix of X as $\mathbf{C} = \mathbf{X}^T\mathbf{X} = (c_{j,k})_{j,k=1}^p$. For a threshold value $\pi_n \in (0,1)$, define the subset $\mathcal{C}_j$ as $\mathcal{C}_j = \{k \neq j : |X_j^T X_k| = |c_{j,k}| > \pi_n\}$ separately for each variable X_j. Let $\tilde{\mathbf{X}}_j$ denote a submatrix of $\mathbf{X}$ with X_k as its columns, where $k \in \mathcal{C}_j$, and the projection matrix $\Pi_j = \tilde{\mathbf{X}}_j(\tilde{\mathbf{X}}_j^T\tilde{\mathbf{X}}_j)^{-1}\tilde{\mathbf{X}}_j^T$ will project X_j onto the space spanned by X_k's, where $k \in \mathcal{C}_j$. The tilted variable $X_j^\star$ of each X_j is defined as $X_j^\star = (\mathbf{I}_n - \Pi_j)X_j$ which is orthogonal to the space that is spanned by X_k's, where $k \in \mathcal{C}_j$. The adjusted correlation between the tilted variable $X_j^\star$ and Y can still be bounded by 0 and 1 after a proper rescaling.

1.4 Diagnostic Techniques

Many classical statistical methods have been developed and assessed in the context of assuming a multivariate normal distribution for the predictor vector, denoted by $\mathbf{x} \sim N_p(\mu, \Sigma)$. The probability density function for random vector $\mathbf{x}$ from the multivariate normal distribution is defined as,

$$f(\mathbf{x}) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)),$$

where $|\Sigma|$ denotes the determinant of the matrix Σ and $|\Sigma| \neq 0$ for $\Sigma > 0$, where > 0 indicates positive definiteness.

The normality assumption can generally be relaxed when applying many robust methods. An estimator is called robust if it keeps a reasonable efficiency, and reasonably small bias, as well as being asymptotically unbiased when the assumptions are only approximately met for all values of the parameter. Efficiency and robustness are two underlying fundamental ideas behind parameter estimation. However, a tradeoff arises when one attempts to achieve both. Also, there are two types of estimators,

robust and non-robust. An example of a robust estimator is the median, and that of non-robust is the mean. Over the decades, the importance of robust procedure in statistical inference have been stressed by statisticians. The contribution by Hampel (1968, 1973) and Huber (1972, 1973) are very important in the field of robust statistics. Although the methods they proposed are good at dealing with outliers, they easily suffer from a loss of efficiency[1] if there is no contamination in the assumed model distribution (Beran 1977).

Hampel (1968) introduced the influence function/curve to distinguish these two kinds of estimators. He pointed out that in general, the influence curve of an efficient estimator will show unboundedness, while a robust one will always be bounded below and above.

In many areas of statistical inference, minimum distance approaches yield robust estimates. There are several methodologies for measuring distance. Among these methodologies, the *Minimum Hellinger Distance* (MHD), which is introduced by Beran (1977), is one of the popular distance-type methods.

1.4.1 Classical Influential Diagnostic Measure

One can measure the level of influence of an observation on Eq. (1.2) by the use of the residuals ($\epsilon_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$), projection matrix ($H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ with diagonal elements $h_{ii} = \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i$), influence functions and et cetera. We limit our discussion in this section to the influence functions since the proposed methods in Chapter 3 are based on the construction of an influence function (IF).

Hampel (1968) introduced the *influence function* (IF) to measure the influence of the ith observation, and the IF is defined as follows: let $T(\cdot)$ be a real-valued functional defined on some subset of the set of all probability measure on $\mathbb{R}$; let F be a probability measure on $\mathbb{R}$ where T is defined. The parameter estimate for a dataset would be denoted $T(F)$, and let $\boldsymbol{\beta}$ denote the true value of a parameter. $T(F)$ can be called a robust estimator if 'small' changes in F do not produce big fluctuations.

[1] An unbiased estimator T of a parameter $\theta \in \Theta$ is called efficient if it attains $e(T) = \frac{I^{-1}(\theta)}{var(T)} = 1$, where $I(\theta)$ is the Fisher information of the sample.

The *influence function* of the ith observation is

$$\Upsilon_i(x_i, y_i; F; T) = \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon \delta_{x_i, y_i}) - T(F)}{\epsilon}, \tag{1.18}$$

where $\delta_{x_i,y_i} = 1$ at (x_i, y_i) and 0 otherwise. The discrete version of influence function is also called *sensitivity curve* (Tukey 1970), and

$$\begin{aligned} \Upsilon_i(x_i; F; T) &= \frac{T(\frac{n-1}{n} F_{n-1} + \frac{1}{n}\delta_{x_i}) - T(F_{n-1})}{1/n} \\ &= n[T_n(x_1, x_2, \ldots, x_n) - T_{n-1}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)], \end{aligned} \tag{1.19}$$

where δ_{x_i} is a distribution with a point mass at x_i and $T_n(x_1, x_2, \ldots, x_n)$ is a statistic based on a random sample $\{x_1, x_2, \ldots, x_n\}$. The boundedness of the influence function/curve usually determines the robustness of the parameter estimator. Robust estimators usually have bounded influence curve, such as median functional of F. Non-robust estimators usually have unbounded influence curve, such as mean functional of F.

The common approach of influence analysis based on influence functions is deleting (or adding) one observation and see how this deletion (or adding) affects the vector of parameter estimates. Cook (1977) suggested a measure of the squared distance between the least square estimate based on all n observations, $\hat{\boldsymbol{\beta}}$ and the estimate obtained by deleting the *ith* point, say $\hat{\boldsymbol{\beta}}_{(-i)}$. This measure is called Cook's distance or Cook's statistic and it is defined as follows: suppose the parameter of interest is $\hat{\boldsymbol{\beta}} = T(F)$, where F is a joint CDF of the $(p+1)$-vector $(\mathbf{x}^T, y)$ with

$$E_F \left\{ \begin{pmatrix} \mathbf{x} \\ y \end{pmatrix} (\mathbf{x}^T, y) \right\} := \begin{pmatrix} \Sigma(F) & \sigma(F) \\ \sigma^T(F) & \tau(F) \end{pmatrix}.$$

The functional corresponding to the least squares estimator of $\boldsymbol{\beta}$ is $T(F) = \Sigma^{-1}(F)\sigma(F)$. The influence function $\Upsilon_i = T_n(F) - T_{n-1}(F) = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}$ and it is a vector which can be normalized to a meaningful way. For appropriate choice of M and c,

$$D_i(M; c) = \frac{\Upsilon_i^T M \Upsilon_i}{c}. \tag{1.20}$$

.

Substituting $\Upsilon_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)}$, $M = \mathbf{X}^T\mathbf{X}$ and $c = p\hat{\sigma}^2$ in Eq. (1.20) to get the Cook's distance,

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})}{p\,\hat{\sigma}^2}, \tag{1.21}$$

where $\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$, the mean squared residual of the full least squares fit.

Cook's Distance Eq. (1.21) can be easily computed in low-dimensional data since we do not need to re-estimate the model for each removed observation, see the algebra detail in Section 3.1. It is implemented in many statistical software such as R, SAS and SPSS. Besides Cook's distance, Hadi's influence measure, likelihood distance, modified Cook's distance, t star ($t^\star$), and Welsch's distance are also popular diagnostic measures for linear regression model (Cook and Sanford 1980). All these methods share the same underlying principle in determining an influential observation which is deleting one observation and comparing the results obtained from the same model with and without the deleted observation.

Johnson (1985) proposed the Kullback−Leibler divergence as a discrepancy measure for identifying observations which are influential in logistic regression. Pardo (2005) uses a generalization of the divergence type measure using phi–divergences, which is equivalent to the classical Cook's distance and Johnson (1985)'s method with a specific phi function (a convex function with nonnegative support).

1.4.2 Development of High-Dimensional Influence Measure

The information technology industry has became the fastest growing and most profitable sector of the world economy Hastie et al. 2009. Much of this growth can be attributed to the development, management and storage of data for medical, engineering, commercial and scientific purposes. Examples include, but not limited to, medical imaging data, genetic data, financial data and satellite data. Dramatically increasing dimension of data came along with the above development. In that, contemporary statistical analysis encounters instances of accessing large samples of

observations with comparably or even larger number of variables of interest. Traditional methods used in low dimensional data are usually not applicable in high dimensional data.

Linear regression continues to be one of the most important statistical tools in the era of high dimensional data. To handle these high dimensional sparse problems, we have witnessed a technological explosion in the development of new regression methodologies during the last 25 years (for instance, Tibshirani 1996, Efron et al. 2004, Fan and Lv 2008, Shao and Deng 2012, Wang and Leng 2016). In light of this, for an appropriate model to be chosen, a careful study of the individual data points (observations) is needed; as some of these individual data points can have tremendous influence on the model and hence could lead to inaccurate interpretation. Thus, an appropriate method is needed to identify such data points. This has led to the issue of 'influence measure' again in the high dimensional context. High-dimensional influence measure aims at detecting the data points which have influence on the model selection process. This diagnostic step is very crucial since the inclusion of influential data point(s) may lead to a distorted model building and weak prediction accuracy. The methods introduced in the previous section are only targeting low dimensional data and do not work appreciably for the high dimensional data. The ability to compute reliable estimates of parameters and the associated precision matrix are critical barriers of applying traditional methods in high dimensional data. Besides these, other barriers may include the computational cost associated with large number of covariates, statistical inference accuracy and algorithm stability (Fan and Lv 2008).

In the classical linear regression model setup (1.2), an ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is obtained by minimizing the objective function $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$, and the solution requires the calculation of $(\mathbf{X}^T\mathbf{X})^{-1}$, which is infeasible when $p > n$. Recall Eq. (1.21), we notice that $(\mathbf{X}^T\mathbf{X})^{-1}$ and $(\mathbf{X}_{(-i)}^T\mathbf{X}_{(-i)})^{-1}$ should be calculated to get $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(-i)}$, it may not be directly computable if $p > n$. Note also that Cook's distance is approximately close to a F-distributed statistic. In the high-dimensional context, it does not make sense to have $F(p, n-p)$ with negative degree of freedom for the denominator. OLS is known to be unstable (or not possible to obtain) for $p > n$.

A direct consequence is that Cook's distance is also unstable. Due to these reasons, new influence measures for high-dimensional data need to be developed.

Zhao et al. (2013) proposed a diagnosis measure for high-dimensional data which captures the influence on the marginal correlation. First, they defined the marginal correlation as $\rho_j = E[\frac{(X_j-\mu_{xj})(Y-\mu_{\mathbf{y}})}{\sigma_{xj}\,\sigma_y}]$, where $\mu_{xj} = E(X_j)$, $\mu_{\mathbf{y}} = E(\mathbf{y})$, $\sigma_{xj}^2 = var(X_j)$ and $\sigma_{\mathbf{y}}^2 = var(\mathbf{y})$. The sample estimate of ρ_j is $\hat{\rho}_j = \frac{\sum_{i=1}^n (X_{ij}-\hat{\mu}_{xj})(Y_i-\hat{\mu}_y)}{(n-1)\hat{\sigma}_{xj}\,\hat{\sigma}_y}$, for $j = 1,\ldots,p$. Then, they used the leave-one-out technique[2] to compute the marginal correlation with the kth observation removed as

$$\hat{\rho}_j^{(k)} = \frac{\sum_{i=1,i\neq k}^n (X_{ij} - \hat{\mu}_{xj}^{(k)})(Y_i - \hat{\mu}_{\mathbf{y}}^{(k)})}{(n-2)\hat{\sigma}_{xj}^{(k)}\hat{\sigma}_y^{(k)}}, \quad j = 1,\ldots,p, k = 1,\ldots,n, \tag{1.22}$$

where $\hat{\mu}_{xj}^{(k)}$, $\hat{\mu}_{\mathbf{y}}^{(k)}$, $\hat{\sigma}_{xj}^{(k)}$, $\hat{\sigma}_{\mathbf{y}}^{(k)}$ are the corresponding sample estimates with the kth observation removed. They propose a statistic termed *high-dimensional influence measure* (HIM) which is based on the estimator of the marginal correlation:

$$D_{him}^{(k)} = \frac{1}{p}\sum_{j=1}^{p}(\hat{\rho}_j - \hat{\rho}_j^{(k)})^2. \tag{1.23}$$

For establishing the theoretical properties of HIM, the following conditions are required:

(C.1) For any fixed $j = 1,\ldots,p$, ρ_j is a constant and does not change as p increases.

(C.2) For the covariance matrix $\Sigma = cov(\mathbf{X})$, with the eigendecomposition $\Sigma = \mathbf{Q}\Lambda\mathbf{Q}^T$, the squared L_2 norm of the diagonal elements of Λ is assumed as $\sum_{j=1}^p \lambda_j^2 = O(p^r)$ for some $0 \leq r < 2$.

(C.3) The predictor X_j, $j = 1,\ldots,p$, follows a multivariate normal distribution and the random noise ϵ_i follows a normal distribution.

For finding the asymptotic distribution, they assume $\mu_{xj} = \mu_y = 0$, $\sigma_{xj} = \sigma_y = 1$ for $1 \leq j \leq p$ and let $K_{p,ts} = \sum_j X_{tj}X_{sj}/p$, then $D_{him}^{(k)}$ can be decomposed as $B_1 + B_2 + B_3 - 2B_4$, where $B_1 = \frac{1}{(n(n-1))^2}\sum_{t=1}^n Y_t^2 K_{p,tt}$, $B_2 = \frac{n-2}{pn(n-1)^2}Y_k^2\|X_k\|^2 = \frac{n-2}{n(n-1)^2}Y_k^2 K_{p,kk}$, $B_3 = \frac{1}{(n(n-1))^2}\sum_{t\neq s} Y_tY_sK_{p,ts}$ and $B_4 = \frac{1}{n(n-1)^2}\sum_{t=1,t\neq k}^n Y_kY_tK_{p,tk}$. Cook's distance detects influential points by finding high leverage h_{ii} and high residual

[2]leave-one-out technique consists of deleting one observation at each step when finding the estimate for the $\rho^{(k)}$.

r_i simultaneously, while $\|X_k\|^2$ and Y_k in the HIM act the similar roles. In Zhao et al. (2013)'s Theorem 1, suppose conditions (C.1)-(C.3) hold, when there is no influential point and $\min\{n, p\} \to \infty$, the asymptotic distribution for $n^2 D_{him}^{(k)}$ is a chi-square distribution with degree of freedom equal to 1. The p-value, $P(\chi^2(1) > n^2 D_{him}^{(k)})$, can be used to determine the rejection region of this hypothesis test $H_0 : ith$ observation is not an influential one.

Zhao et al. (2013) used the numerical studies to demonstrate that HIM is useful in models with contamination in both response and predictors. Also, possible extension to the generalized linear models (GLM) can be expressed as

$$D_{him}^{(k)} = \frac{1}{p} \sum_{j=1}^{p} \|\hat{\beta}_j - \hat{\beta}_j^{(k)}\|_2^2. \tag{1.24}$$

HIM is a good method to detect the high dimensional influential observation, but depends only by using the robust estimates of median and least absolute deviation (LAD) from the sample. Also, the estimate of marginal correlation is not bounded by 1 since the standardization is not used for each leave-one-out step. As shown in Example 1.3.1, high dimensionality of the data brings high correlations among the variables, which results in marginal correlation being unreliable. For those reasons, new methods are still needed in the high dimensional influence measure.

1.5 Contribution of this Thesis

In high-dimensional sparse modelling, seminal theories of *least angle regression* (Efron et al. 2004, LARS) and *sure independence screening* (Fan and Lv 2008, SIS) both used correlation between predictor variables and response (or current residual) to deal with selection and estimation problems. The correlation can be viewed as a high-dimensional counterpart to the ordinary least square (OLS) estimator of the parameter vector and many data-driven methods based on correlation have been studied for years in high dimensional statistics. In this thesis, we contribute to the high-dimensional correlation learning theory from three important problems: variable screening for random and deterministic design matrices; influence measure and post-

selection inference. The novel contributions of this dissertation include:

- We propose a new estimator for the correlation between the response and high-dimensional predictor variables, and based on the estimator we develop a new screening technique termed *dynamic tilted current correlation screening* (DTCCS) for high dimensional variables screening. DTCCS is also extended to the deterministic design matrix.

- We propose two new influence measure and diagnostic procedures from two different viewpoints: the extreme value distribution and the robustness of design.

- We propose a new post-selection inference method which is based on a cosine distribution to deal with high-dimensional inference problem.

The rest of the dissertation is organized as follows. In Chapter 2, we study the problem of high-dimensional variable screening which is among the most widely studied applications of sparse modelling and estimation. In the ultra-high dimensional setting, the SIS method was introduced to significantly reduce the dimensionality to a moderate scale which is below the sample size and preserve the true model with probability tending to 1. The performance of SIS must depend on the marginal correlation which is unreliable due to the dimensionality. In reality, the 'importance' of the variables cannot be easily ranked by their marginal correlation and there exists high (possible spurious) correlation among predictor variables. To overcome them, we propose a new estimator for high-dimensional correlation and a novel screening technique which termed *dynamic tilted current correlation screening* (DTCCS). The new method reduce high correlation among predictor variables in a data-driven way. We show that DTCCS is able to discover all relevant predictors within a finite number of steps when the dimension of true model meets the sparse assumption. DTCCS's sure screening property, consistency property and computational complexity are illustrated theoretically and numerically. To confirm the effectiveness of the proposed methods, we conduct simulation studies and a real-life data analysis to illustrate the usefulness of DTCCS. We apply the DTCCS method in the random design matrix and discuss the potential extension to the deterministic design matrices.

In Chapter 3, we study the problem of high-dimensional influence measure and diagnostic procedure. Influence diagnosis plays an important role in data analysis. Some observation can have tremendous influence on the model and hence could lead to misleading results in regression problems, for instance, distorted variable selection, inaccurate interpretation. Traditional influence detection methods such as Cook's distance measures individual observation's influence on the least squares regression coefficient estimates. However, it will have problem when applied to high-dimensional data. Estimation accuracy and computational cost are two top concerns in high-dimensional data analysis. Difficulties in detecting the influential observations in high-dimensional data may lead to distorted analysis and a high computational complexity. Zhao et al. (2013) propose *High-dimensional Influence Measure* (HIM) which captures the influence on the marginal correlations. However, marginal correlation strongly relies on the independence assumption among predictors which rarely holds in reality. Also, HIM highly depends on the robust estimator. Inspired by the recent work of Cai et al. (2014) and Karunamuni et al. (2015), we propose two new methods to capture the influence on a function of the correlations. The two methods are from the perspectives of the extreme value distribution and the robustness of design respectively. They are both constructed from the high-dimensional correlations. The asymptotic distributions of these proposed influence diagnostic techniques have been established by letting the dimension of the explanatory variable approach infinity. To confirm the effectiveness of the proposed methods, simulation studies are conducted extensively.

In Chapter 4, we use the geometric arguments to discuss the post-selection inference of LARS. The new procedure is based on truncated cosine distribution. At each step of the LARS algorithm, we get a corresponding angle from the correlation between entering variable and current residual. In the high-dimensional context, the angle will be considered as a random variable from cosine distribution, then we can do post-selection inference based on that. Also, the limiting distribution of the maximum angle can be used to do an efficient and robust significance test for each predictor variable. To confirm the effectiveness of the proposed method, we conduct simulation

studies and a real-life data analysis to illustrate the usefulness of this post-selection method.

In Chapter 5, we draw connections between these different statistical problems under the overall theme of this thesis, the correlation learning. It contains the summary and conclusions on the performance of the methods proposed. We also provide some directions for further studies.

Chapter 2

Dynamic Tilted Current Correlation for High Dimensional Variable Screening

Variable screening is a general procedure in high dimensional data analysis to ensure the applicability of statistical methods. It is a complicated and computationally burdensome procedure since spurious correlations commonly exist among predictor variables, and important predictor variables may not have large marginal correlations with the response variable. In this chapter, we propose a new estimator for the correlation between the response and high-dimensional predictor variables, and based on the estimator we develop a new screening technique termed *dynamic tilted current correlation screening* (DTCCS) for high dimensional variables screening. DTCCS is capable of picking up the relevant predictor variables within a finite number of steps. The DTCCS method takes the popular *sure independence screening* (SIS) method and the *high-dimensional ordinary least squares projection* (HOLP) approach as its special cases. The DTCCS technique has sure screening and consistency properties which are demonstrated theoretically and numerically and illustrated through a real-life example.

2.1 Introduction

As the computational power increases and the cost of data collection decreases, high dimensional or ultra-high dimensional data are available more than ever. Data with tens of thousands of variables are frequently seen in modern scientific research, such as oncology image data, financial data, satellite data and genomics data. In such datasets, the dimension p of variables is much larger than the sample size n, but only a small number of variables are believed to be significantly relevant to the response of interest. It is imperative to perform a screening stage for relevant variables before a formal statistical model building procedure in order to extract truly useful underlying information from the data. For this purpose, Fan and Lv (2008) proposed the *sure independent screening* (SIS) method for selecting important variables in ultrahigh-dimensional linear models. The SIS method uses a correlation learning method to rank the importance of predictors according to their marginal correlation with the response variable and includes those having strong marginal correlations with the response variable into the model. Variable screening has received increasing attention in the literature and many new techniques have been investigated in recent years. For example: Wang (2009) showed that the *forward regression variable screening* (FRVS) method can also identify all relevant predictors consistently. Fan and Song (2010) extended the SIS approach to generalized linear models (GLM) by ranking the maximum marginal likelihood estimates (MMLE). Fan et al. (2011) extended the correlation learning to marginal nonparametric learning which can be used in sparse ultra-high dimensional additive models. Zhu et al. (2011) introduced a screening approach under a unified model framework which covers parametric and semiparametric models. Merging the idea of the SIS method and the robust estimator of correlation, Li et al. (2011) and Li et al. (2012a) proposed *robust rank SIS* (RSIS) and *robust rank correlation screening* (RRCS), respectively, to deal with ultra-high dimensional data. To protect from model misspecification, Li et al. (2012b) developed a robust SIS procedure based on the distance correlation (DC-SIS) under more general settings including linear models. Cho and Fryzlewicz (2012) proposed a *tilting*

procedure for variable screening which can efficiently reduce the spurious correlation among predictors. Wang and Leng (2016) used the Moore-Penrose inverse to form a new correlation-based screening technique, called *high-dimensional ordinary least squares projection* (HOLP).

To reduce high spurious correlation among predictors, we propose a correlation estimator between the predictor and the current residual to form a path of predictors entering the model, and this path is then used for variable screening. This new screening technique is termed *dynamic tilted current correlation screening* (DTCCS). Our proposed method is appealing in several aspects. It can retain the important predictors which may have small marginal correlations with the response, and meanwhile, exclude unimportant predictors which may have large correlation with the response. Like the SIS method, the DTCCS approach makes use of the correlation learning, and thus, preserves the sure screening property. Unlike the SIS method, our DTCCS algorithm employs the 'tilted' current correlation to measure the importance of predictors. The DTCCS method uses a path-based regression algorithms like the forward-type regression, LARS and LASSO. LARS or LASSO adds variables one by one to build a final model, but DTCCS adds variables one group after another. For the LASSO method, the number of non-zero variables in the 'best' final model only depend on a single tuning parameter which means that a sequence of 'knots' of tuning parameters determine different final models. For the DTCCS, the candidate model size is predetermined and a group of monotone value of tuning parameters have been used to form a final model.

The rest of Chapter 2 is organized as the follows. The relevant notation and the framework are introduced in Section 2.2. The methodology is presented in Section 2.3. Numerical studies are reported in Section 2.4. Extensions to the deterministic design matrix are discussed in 2.5. Chapter 2 concludes with discussion and possible extensions in Section 2.6.

2.2 The Model and Notations

Recall the linear regression model from Eq (1.2):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ ($\mathbf{X}$ is a $n \times p$ matrix of p dimensional covariates), $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of the coefficient of the respective covariates and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise in the model. Throughout this chapter, $\mathbf{X}$ is assumed to be a full row rank matrix.

Consistent with the common procedure in high-dimensional data analysis (Efron et al. 2004, Fan and Lv 2008, Wang 2009, Cho and Fryzlewicz 2012, Wang and Leng 2016, Fan et al. 2018), we now standardize the response vector $\mathbf{y}$ using the transformation $\mathbf{y} - E(\mathbf{y})$ and standardize the covariate column vectors $\mathbf{X}_j$ by the transformation $\{\mathbf{X}_j - E(\mathbf{X}_j)\}\{var(\mathbf{X}_j)\}^{-1/2}$. Hence, all covariates are standardized to have an equal finite norm (Fan et al. 2018). Note that we use (abuse) the same notation of the random vector and the corresponding realization in the data for ease the complexity of notations here.

$\mathbf{X}_j$ is referred to as a relevant (or irrelevant) predictor if $\beta_j \neq 0$ (or $\beta_j = 0$), where β_j is the jth component of $\boldsymbol{\beta}$. Define the full model as $\mathcal{F} = \{1, \ldots, p\}$, and the true model as $\mathcal{T} = \{1 \leq j \leq p : \beta_j \neq 0\}$. We have $|F| = p$ and let $|\mathcal{T}| = t_0$. Let $\mathcal{M}_k = \{j_1, \ldots, j_k\}$ be an active set which means that the current model has $\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_k}$ as relevant predictors, where $1 < k < n$. Let $\tilde{\mathbf{X}}_{-j}$ be the submatrix of $\mathbf{X}$ which excludes the column $\mathbf{X}_j$. A projection matrix H_j based on 'ridge regression' projects $\mathbf{X}_j$ to the space spanned by all the column vectors $\mathbf{X}_k$ with $k \neq j$, given by $H_j = \tilde{\mathbf{X}}_{-j}(\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j} + \lambda\mathbf{I}_{p-1})^{-1}\tilde{\mathbf{X}}_{-j}^T$ where λ is a tuning parameter defined in Section 2.3.1. By the spectral decomposition theorem, $\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T = \mathbf{P}_j\mathbf{D}_j^2\mathbf{P}_j^T$, where $\mathbf{D}_j = diag(d_{j1}, \ldots, d_{jn})$ is a diagonal matrix with the diagonal entries $d_{j1} \geq d_{j2} \geq \cdots \geq d_{jn} > 0$ being the eigenvalues of $\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T$, and the column vectors of $\mathbf{P}_j$ are the eigenvectors of $\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T$ corresponding to the eigenvalues and are orthonormal.

2.3 Proposed Methodology

2.3.1 Inferential Methods

Multi-stage statistical procedures are more and more important in the high dimensional setting (Wasserman and Roeder 2009). The screening procedure is the critical first stage before further moderate-scale learning and inference. Fan and Lv (2008)'s SIS method makes use of the absolute value of the marginal correlations $\boldsymbol{\omega} = |\mathbf{X}^T\mathbf{Y}| = (\omega_1, \ldots, \omega_p)^T$ and selects the variables which have relatively high marginal correlations with the response. However, SIS strongly relies on the assumption that the important variables in the model have large marginal correlations with the response, which is not always true in reality. Efron et al. (2004)'s LARS gives a forward solution path by using the equiangular direction of $\mathbf{X}_j$'s and determines the current step size and next direction simultaneously. To overcome the independence violation and reduce the spurious correlation, we propose a novel and simple screening technique by merging the idea of Forward-type regression and screening procedure. The proposed method takes the popular used SIS method as its special case and its other special cases also connect to the ordinary least square estimator (OLS) and the *high-dimensional ordinary least squares projection* (HOLP), see details in the following sections.

High-dimensional Correlation Estimator

In traditional linear regression, the ordinary least squares (OLS) method projects the response $\mathbf{Y}$ onto the linear space $col(\mathbf{X})$ spanned by the column vectors of $\mathbf{X}$. High-dimensional screening methods, such as *forward regression variable screening* (FRVS) (Wang 2009) and tilting (Cho and Fryzlewicz 2012), also project $\mathbf{Y}$ onto $col(\mathbf{X})$. Different from those projections, Shao and Deng (2012) proposed to project the regression vector $\boldsymbol{\beta}$ onto the linear space $row(\mathbf{X})$ which is spanned by the row vectors of $\mathbf{X}$ and showed that large and small elements of the projection of $\boldsymbol{\beta}$ onto $row(\mathbf{X})$ can be discriminated efficiently with probability tending to 1. One advantage of using $row(\mathbf{X})$ in high-dimensional screening is that the dimension of $row(\mathbf{X})$ is at

most n which is much smaller than p in the high-dimensional context.

For $p > n$, we consider the ridge regression estimator of $\boldsymbol{\beta}$ (Hoerl and Kennard 1970) under model (1.2),

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{Y},$$

where $\lambda > 0$ is an appropriately chosen regularization parameter, and $\mathbf{I}_p$ is a $p \times p$ identity matrix. Shao and Deng (2012) and Wang and Leng (2016) showed that the computation of $\hat{\boldsymbol{\beta}}_{ridge}$ involves only inverting an $n \times n$ matrix since

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_n)^{-1}. \tag{2.1}$$

Equation (2.1) implies that the ridge regression estimator $\hat{\boldsymbol{\beta}}_{ridge}$ is always in the row space of $\mathbf{X}$, $row(\mathbf{X})$.

The *high-dimensional ordinary least squares projection* (HOLP) method by Wang and Leng (2016) calculated an estimator of $\hat{\boldsymbol{\beta}}$:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{holp} &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{Y}, \\ &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\beta + \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\boldsymbol{\epsilon}. \end{aligned} \tag{2.2}$$

The projection matrix of the HOLP method, $\mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}$, is spanned by the row space of $\mathbf{X}$ and diagonally dominant. The HOLP method projects $\boldsymbol{\beta}$ onto $row(\mathbf{X})$ to obtain $\hat{\boldsymbol{\beta}}_{holp}$. When the sparse parameter vector $\boldsymbol{\beta}$ has many zero components, $\hat{\boldsymbol{\beta}}_{holp}$ may not have any zero component but many of them must be negligible due to the screening consistency property of the HOLP procedure. Hence, $\hat{\boldsymbol{\beta}}_{holp}$ can also be viewed as a generalized sparse vector and can separate the relevant and irrelevant predictor variables efficiently.

Graybill (1983) suggested an estimator of $\boldsymbol{\beta}$ from a generalized inverse matrix point of view:

$$\hat{\boldsymbol{\beta}} = \mathbf{X}^-\mathbf{Y} + (\mathbf{I}_p - \mathbf{X}^-\mathbf{X})h, \tag{2.3}$$

where $\mathbf{X}^-$ is a generalized inverse of $\mathbf{X}$ and h is a $p \times 1$ vector. If we take $h = \mathbf{X}^T\mathbf{Y}$ in (2.3), then in the cases where $n > p$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the OLS estimator;

in the case with $n < p$, $\hat{\boldsymbol{\beta}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{Y}$ is the HOLP estimator. Graybill (1983) shows that $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$ if and only if $\mathbf{X}^T\mathbf{X}$ is idempotent.

Motivated by the tilting technique (Cho and Fryzlewicz 2012) and the row space $row(\mathbf{X})$ introduced by Shao and Deng (2012) and Wang and Leng (2016), we propose a new correlation estimator for high-dimensional data which efficiently reduces the 'spurious' correlation among the predictors. We expect this proposed correlation estimator to rank the important elements of $\boldsymbol{\beta}$ correctly and thus to screen the important predictors iteratively.

For the high-dimensional $n \times p$ design matrix $\mathbf{X}$, we consider each column $\mathbf{X}_j$, $j = 1, \ldots, p$, as a 'response' variable and the rest $(p-1)$ columns as the corresponding design matrix. Ridge regression is used to 'tilt' $\mathbf{X}_j$ such that the effect of other variables $\mathbf{X}_k$, $k \neq j$, on $\mathbf{X}_j$ is reduced. The 'strength' of tilting can be adjusted by a tuning parameter λ. The current residual $\mathbf{Z}$ is defined to be the ridge residual vector when regressing $\mathbf{Y}$ against the active variables in $\mathcal{M}$ with tuning parameter λ. Note that $\mathbf{Z}_1 = \mathbf{Y}$ for the initial step. Hence, a new measure for the contribution of each variable to the current residual $\mathbf{Z}$ can be expressed as

$$\hat{\rho}_j(\lambda) = \frac{1}{a_j}\mathbf{X}_j^T(\mathbf{I}_n - H_j)\mathbf{Z}, \tag{2.4}$$

where $H_j = \tilde{\mathbf{X}}_{-j}(\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j} + \lambda\mathbf{I}_{p-1})^{-1}\tilde{\mathbf{X}}_{-j}^T$, and a_j is a scaler which rescale the tilted correlation back to be bounded by 1. We call this estimator the *high-dimensional correlation estimator* (HDCE). Let $s_j = sign\{\hat{\rho}_j\}$ for $j = 1, \ldots, p$. Noting that

$$\begin{aligned}
H_j &= \tilde{\mathbf{X}}_{-j}(\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j} + \lambda\mathbf{I}_{p-1})^{-1}\tilde{\mathbf{X}}_{-j}^T = \frac{1}{\lambda}\tilde{\mathbf{X}}_{-j}\left[\mathbf{I}_{p-1} + \left(\frac{\tilde{\mathbf{X}}_{-j}}{\sqrt{\lambda}}\right)^T\left(\frac{\tilde{\mathbf{X}}_{-j}}{\sqrt{\lambda}}\right)\right]^{-1}\tilde{\mathbf{X}}_{-j}^T \\
&= \frac{1}{\lambda}\tilde{\mathbf{X}}_{-j}\left[\mathbf{I}_{p-1} - \frac{\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}}{\lambda} + \frac{\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}}{\lambda^2} - \ldots\right]\tilde{\mathbf{X}}_{-j}^T \\
&= \frac{\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T}{\lambda} - \frac{\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T}{\lambda^2} + \frac{\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T}{\lambda^3} - \ldots,
\end{aligned}$$

we obtain that

$$\begin{aligned}\mathbf{I}_n - H_j &= \mathbf{I}_n - \frac{\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T}{\lambda} + \frac{\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T}{\lambda^2} - \frac{\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T}{\lambda^3} + \dots \\ &= \left(\mathbf{I}_n + \frac{\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T}{\lambda}\right)^{-1} \\ &= \lambda(\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T + \lambda\mathbf{I}_n)^{-1}. \end{aligned} \tag{2.5}$$

Performing singular value decomposition (SVD) of $\tilde{\mathbf{X}}_{-j}$, $j = 1, \dots, p$,

$$\tilde{\mathbf{X}}_{-j} = \mathbf{P}_j\mathbf{D}_j\mathbf{Q}_j^T, \tag{2.6}$$

where $\mathbf{P}_j$ is an $n \times n$ matrix satisfying $\mathbf{P}_j^T\mathbf{P}_j = \mathbf{I}_n$, $\mathbf{Q}_j$ is a $(p-1) \times n$ matrix satisfying $\mathbf{Q}_j^T\mathbf{Q}_j = \mathbf{I}_n$, $\mathbf{D}_j$ is an $n \times n$ diagonal matrix of full rank with diagonal entries being d_{ji}, $i = 1, \dots, n$. Note that the middle matrix in the traditional SVD is an $n \times p$ rectangle. In the high-dimensional statistics, it is more popular to use an $n \times n$ square matrix for the middle matrix in SVD instead of using a very wide $n \times p$ matrix with one block of diagonal and another block of all 0's, see examples in Fan and Lv (2008), Wang and Leng (2016) and R function 'svd'.

Using the eigendecomposition $\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T = \mathbf{P}_j\mathbf{D}_j^2\mathbf{P}_j^T$ where $\mathbf{D}_j^2$ is an $n \times n$ diagonal matrix with positive elements $d_{j1}^2 \geq d_{j2}^2 \geq, \dots, \geq d_{jn}^2 > 0$. We simplify (2.5) to obtain $\mathbf{I}_n - H_j = \mathbf{P}_j\mathbf{F}_j\mathbf{P}_j^T = \sum_{i=1}^{n} \frac{\lambda}{\lambda + d_{ji}^2} P_{j,i}P_{j,i}^T$, where $\mathbf{F}_j = diag(\frac{\lambda}{\lambda+d_{j1}^2}, \dots, \frac{\lambda}{\lambda+d_{jn}^2})$.

We then obtain $\mathbf{X}_j^T(\mathbf{I}_n - H_j)\mathbf{Z} = \mathbf{X}_j^T\mathbf{P}_j\mathbf{F}_j\mathbf{P}_j^T\mathbf{Z} = (\mathbf{F}_j^{1/2}\mathbf{P}_j^T\mathbf{X}_j)^T(\mathbf{F}_j^{1/2}\mathbf{P}_j^T\mathbf{Z})$ which is the inner product of $\mathbf{X}_j^\star = F_j^{1/2}\mathbf{P}_j^T\mathbf{X}_j$ and $\mathbf{Z}^\star = F_j^{1/2}\mathbf{P}_j^T\mathbf{Z}$. Let $a_j = \|\mathbf{X}_j^\star\|_2 \cdot \|\mathbf{Z}^\star\|_2$.

The tilted correlation $\hat{\rho}_j(\lambda)$ in (2.4)can be considered as a discrete function of λ, where λ takes values at $\infty = \lambda_0 \geq \lambda_1 \geq \cdots \geq \lambda_\gamma \geq 0$ for a given integer of γ. By choosing different values of λ, the high (most likely spurious) correlation among predictors can be efficiently controlled. Theoretically, this property prevents the irreverent variables from entering the model and thus discriminates relevant and irrelevant variables successfully. The current residual $\mathbf{Z}$ carries the information of the selected variables which save the relevant variables without a big marginal correlation. Dynamically reducing the value of λ and updating the current residual $\mathbf{Z}$, we obtain

a promising high-dimensional variable screening method, and we call it the *dynamic tilted current correlation screening* (DTCCS) method.

Dynamic Tilted Current Correlation for High-dimensional Variable Screening

The key idea of the DTCCS method is to iteratively rank the variables according to the absolute values of the proposed correlation estimator. For each iteration, we initially define a step size d, for instance, let $d = 1, 2, \ldots, \log n$, a large value of d can speed up the algorithm. We choose $d = \sqrt{\frac{p}{n}} \log n$ for illustrations in this section. After γ iterations, we reduce the dimension from high p to a moderate size of $m = \min(\sqrt{\frac{p}{n}} \log n^{\gamma}, n-1)$. Thereafter, the dimension of the covariates is diverging no faster than the sample size. Hence, many classical variable selection and estimation methods (for instance, the LARS method) or the model selection criteria can be implemented easily to obtain the final statistical model.

Suppose we have an active set $\mathcal{M}$ which consists of $(\gamma+1)$ disjoint selected subsets $\mathcal{M}_k$ where $0 \leq k \leq \gamma$. The initial set is the null set $\mathcal{M}_0 = \{\emptyset\}$. For each iteration, we rank the remaining variables by descending the order of the absolute value of $\hat{\rho}_j(\lambda)$ for $j \notin \mathcal{M}$, i.e., $\mathcal{M}_1 = \{j_1, j_2, \ldots, j_d\}$ and $\mathcal{M}_1 \bigcup \mathcal{M}_2 \bigcup \cdots \bigcup \mathcal{M}_{\gamma} = \mathcal{M}$.

When λ is big enough, $\mathbf{F}_j$ is close to $\mathbf{I}$ in (**??**), $\mathbf{I}_n - H_j$ is close to $\mathbf{I}$ in (2.5), and $\hat{\rho}_j$ is close $\mathbf{X}_j^T \mathbf{Y}$ in (2.4), which is the case of sure independence screening. When H_j is close to $\mathbf{0}$, geometrically, $\mathbf{X}_j$ is perpendicular to the space spanned by $\mathbf{X}_k$ for $k \neq j$.

With some decreasing values λ's, say, a knot sequence $\{\lambda_k, k = 1, 2, \ldots, \gamma\}$, each knot λ_k marks the entry of $\mathcal{M}_k$ which is a group of variables to be included in the active set $\mathcal{M}$, $0 \leq k \leq \gamma$.

When λ is small enough, say, λ is close to 0, $\mathbf{F}_j$ is close to $\mathbf{0}$ in (**??**), $\mathbf{I}_n - H_j$ is close to $\mathbf{0}$ in (2.5) and $\hat{\rho}_j$ is close to 0 in (2.4). In this case, H_j is close to $\mathbf{I}$, and geometrically, $\mathbf{X}_j$ is almost in the space spanned by $\mathbf{X}_k$ for $k \neq j$. Also, as $\lambda \to 0$, $H_j = \tilde{\mathbf{X}}_{-j}(\tilde{\mathbf{X}}_{-j}^T \tilde{\mathbf{X}}_{-j})^{-} \tilde{\mathbf{X}}_{-j}^T = \tilde{\mathbf{X}}_{-j} \tilde{\mathbf{X}}_{-j}^T (\tilde{\mathbf{X}}_{-j} \tilde{\mathbf{X}}_{-j}^T)^{-1} = \mathbf{I}$, as shown by Wang and Leng (2016), where A^{-} denotes the Moore-Penrose generalized inverse of matrix A. The relationship between DTCCS and HOLP will be discussed in Theorem 2.5.2.

For the knot selection in LASSO path, Lockhart et al. (2014) suggested to use the absolute value of the marginal correlation to determine the knots of the LASSO path, for example, $\lambda_j = |\mathbf{X}_j^T \mathbf{Y}|$. In our proposed method, we denote the percentage of remaining variables to be δ, $0 \leq \delta \leq 1$. We take $\lambda = \frac{\delta}{1-\delta}$ which approaches ∞ as δ is close to 1. The first step of our DTCCS method is the same as that of the SIS method. The following steps use bounded λ which guarantee the volume of tilting.

Connection to Classical Linear Regression

In this subsection, we discuss the relationship between the new estimator and the least square estimator in the classical linear models (1.2) with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$. We only consider the case of fixed design setting here, then the design matrix in this setting is a deterministic design matrix. Hence, $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I})$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. For $n > p$, let $\mathbf{X}_j^{\perp}$ be the projection of $\mathbf{X}_j$ to the orthogonal complement of the column space of $\tilde{\mathbf{X}}_{-j}$. Note that $\mathbf{X}_j^{\perp} = \mathbf{X}_j$, $j = 1, \ldots, p$, if $\mathbf{X}$ is orthogonal design matrix. For a linear model without interception, the least square estimates $\hat{\beta}_j^{lse}$ and its corresponding 'true' parameter value β_j can be respectively expressed as

$$\hat{\beta}_j^{lse} = ({\mathbf{X}_j^{\perp}}^T \mathbf{Y})/({\mathbf{X}_j^{\perp}}^T \mathbf{X}_j), \text{ and } \beta_j = ({\mathbf{X}_j^{\perp}}^T \boldsymbol{\mu})/({\mathbf{X}_j^{\perp}}^T \mathbf{X}_j) \tag{2.7}$$

where ${\mathbf{X}_j^{\perp}}^T \mathbf{X}_j \neq 0$ for $n > p$ (Berk et al. 2013; Zhang and Zhang 2014). Considering inference for $\hat{\beta}_j^{lse}$ and its target β_j, $\hat{\beta}_j^{lse} \sim N(\beta_j, \sigma^2/\|\mathbf{X}_j^{\perp}\|^2)$. Berk et al. (2013) assumed a valid estimate $\hat{\sigma}^2$ of σ^2 which is independent of all $\hat{\beta}_j^{lse}$, and proposed a post-selection confidence interval from a t-statistic with degree of freedom $(n-p)$,

$$t_j = \frac{(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{X}_j}{\hat{\sigma}\|\mathbf{X}_j^{\perp}\|}. \tag{2.8}$$

In general, for $n > p$, $\mathbf{X}_j^{\perp} = (I - H_j)\mathbf{X}_j$, where $H_j = \tilde{\mathbf{X}}_{-j}(\tilde{\mathbf{X}}_{-j}^T \tilde{\mathbf{X}}_{-j})^{-1} \tilde{\mathbf{X}}_{-j}^T$. $\hat{\beta}_j^{lse}$ is the inner product of $\langle \frac{(I-H_j)\mathbf{X}_j}{\mathbf{X}_j^T(I-H_j)\mathbf{X}_j}, (I - H_j)Y \rangle$. The normalized inner product can be denoted as $\frac{\mathbf{X}_j^T(I-H_j)Y}{\|(I-H_j)\mathbf{X}_j\|_2 \|(I-H_j)Y\|_2}$.

Zhang and Zhang (2014) shows (2.7) is the solution of solving the linear equation ${\mathbf{X}_j^{\perp}}^T(\mathbf{Y} - \beta_j \mathbf{X}_j) = 0$. The vector $\mathbf{X}_j^{\perp}$ was termed the 'score vector' for the least squares

estimation of β_j in this linear equation. For $p > n$, $\mathbf{X}_j^{\perp}$ cannot be considered as a score vector anymore and Zhang and Zhang (2014) suggests the orthogonal constraint of score vector can be relaxed. In our proposed estimator, $\mathbf{P}_j F_j \mathbf{P}_j^T \mathbf{X}_j$ is the score vector to solving ${\mathbf{X}_j}^T \mathbf{P}_j F_j \mathbf{P}_j^T (\mathbf{Y} - \beta_j \mathbf{X}_j) = 0$ for high-dimensional data. The solution is the inner product of $\langle \frac{F_j^{1/2} \mathbf{P}_j^T \mathbf{X}_j}{\mathbf{X}_j^T (\mathbf{P}_j F_j \mathbf{P}_j^T) \mathbf{X}_j}, F_j^{1/2} \mathbf{P}_j^T \mathbf{Y} \rangle$, and the normalized inner product is HDCE.

Final Model Selection after Screening

Fan and Lv (2008) pointed out that all the screening methods for high-dimensional data have high type II error due to the nature of high dimensionality and sparse modeling. A quick and simple remedy is to use a given model selection criterion after finding a sequence of submodels. Classical model selection criteria such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are known to select too many variables than necessary for settings with high dimensional data (Chen and Chen 2008). Kim and Jeon (2016) proposed a unified framework of loss functions for selection consistency which is termed *quadratically supported risks* (QSR). This unified framework includes quadratic loss, Huber loss, quantile loss and logistic loss. For $p_n = O(n^{\alpha})$ and a given subset $\mathcal{M} \subset \{1, \ldots, p_n\}$, a final selected model is determined by

$$\hat{M}_{h_n} = \underset{\mathcal{M} \subset \{1, \ldots, p_n\}}{\arg\min} \{ R_n(\hat{\beta}) + h_n \sigma^2 \, |\mathcal{M}| \}, \tag{2.9}$$

where $R_n(\hat{\beta}) = \frac{\|Y - \mathbf{X}\hat{\beta}\|_2^2}{n}$ is the quadratic loss, and h_n is a sequence of positive numbers termed GIC by Kim and Jeon (2016). Kim and Jeon (2016) showed that different selections of h_n may lead to common model selection criteria, such as AIC, BIC and extended BIC.

By using the DTCCS method, the solution path in $\mathcal{M}$ provides a sequence of submodels with increasing complexities. The final model can be chosen by using the QSR framework with the *risk inflation criterion* (RIC) (Foster and George 1994), which corresponds to the choice of $h_n = \frac{\log p_n}{n}$ for RIC.

2.3.2 Theoretical Results

Conditions, Assumptions and Lemmas

Let us begin with the definitions of orthogonal invariance, spherical symmetry, Stiefel manifold and Sub-Gaussian tail condition.

Definition 2.3.1. *(Orthogonal invariance). Let $\mathcal{O}(n)$ be the set of $n \times n$ orthogonal matrices. An n-dimensional random vector z is said to be* orthogonally invariant *if $Qz \overset{(d)}{=} z$ for any orthogonal matrix $Q \in \mathcal{O}(n)$, where the symbol $\overset{(d)}{=}$ represents equality in distribution.*

Definition 2.3.2. *(Spherical symmetry). A random vector $z \in \mathbb{R}^n$ is said to be* spherically symmetric *around $\mu \in \mathbb{R}^n$ if $z - \mu$ is orthogonally invariant. We denote this as $z \sim \mathbf{S}_n(\mu)$.*

Definition 2.3.3. *(Haar measure). For any orthogonal matrix $Q \in \mathcal{O}(n)$ and an $n \times n$ random matrix $\mathbf{X}$, the measure $\mu(\cdot)$ is called the Haar measure (or the invariant measure) on $\mathcal{O}(n)$ if $\mu(Q\mathbf{X}) = \mu(\mathbf{X}Q) = \mu(\mathbf{X})$.*

Definition 2.3.4. *(Stiefel manifold, Tropp 2012). The Stiefel manifold $V_n(\mathbb{R}^p)$ is the set of all orthonormal n-frames in a p-dimensional Euclidean space. That is $V_n(\mathbb{R}^p) = \{X \in \mathbb{R}^{p \times n} : X^T X = \mathbf{I}_n\}$. The orthogonal group of matrix $\mathcal{O}(n)$ can be considered as a special case of Stiefel manifold which is $V_p(\mathbb{R}^p)$. The Stiefel manifold $V_n(\mathbb{R}^p)$ is invariant under a Haar measure which is uniformly distributed on n-frames in $\mathbb{R}^p$.*

Definition 2.3.5. *(Sub-Gaussian tail condition, Kim and Jeon 2016.)*

In the linear model (1.2), the ϵ_i are independent random variables whose common distribution has a sub-Gaussian tail. That is, there is some $b > 0$ such that for every $t \in \mathbb{R}$, we have $E(e^{t\epsilon_i}) \leq \exp\{b^2 t^2/2\}$, which implies that there exist positive constants c_ϵ and d_ϵ such that

$$P\left(\left|\sum_{i=1}^{n} a_i \epsilon_i\right| > t\right) \leq c_\epsilon \cdot \exp\left(-\frac{d_\epsilon t^2}{\sum_{i=1}^{n} a_i^2}\right) \tag{2.10}$$

for all $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ *and* $t > 0$.

Condition 2.3.1. *(Polynomial high-dimensional).* $p > n$ *and* $p = O(n^\alpha)$ *for some* $\alpha > 0$.

Condition 2.3.2. *(Normality assumption). Assume* $\mathbf{X}$ *follows multivariate normal distribution and* $\epsilon \sim N(0, \sigma^2)$ *with variance* σ^2.

Condition 2.3.3. *(Covariance matrix). Let* $d_\star(A)$ *and* $d^\star(A)$ *represent the smallest and the largest eigenvalues of the positive definite covariance matrix* A *respectively. We assume that for some* $0 \leq \kappa \leq 1$ *and* $c_1 > 0$, *the conditional number of* Σ, $cond(\Sigma) = d^\star(\Sigma)/d_\star(\Sigma) \leq c_1 n^\kappa$.

Condition 2.3.4. *(Tilting parameter). Let* λ *be the tilting parameter introduced in Section 2.3.1,* $\lambda = O(p)$ *for the finite selection of* λ.

Recall the linear model Eq. (1.2) $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of the coefficients of the respective covariates and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise in the model. In this section, $\mathbf{X}$ is the random design matrix with standardized predictors, $\mathbf{x} = (x_1, \ldots, x_p)^T$ is used to denote the random predictor vector, $\Sigma_{p\times p} = cov(\mathbf{x})$ is the covariance matrix of the predictors. Since in model (1.2) we assume all the predictors are standardized, Σ is the correlation matrix. We define

$$\mathbf{B} = \mathbf{X}\Sigma^{-1/2} \text{ and } \mathbf{b} = \Sigma^{-1/2}\mathbf{x}. \tag{2.11}$$

It is easily seen that $\mathbf{b}$ has a spherically symmetric distribution and $cov(\mathbf{b}) = I_p$.

Under Condition 2.3.3 and 2.3.4, the diagonal values of $\mathbf{F}_j$ are bounded by $O(1)$ for different selections of λ's. For each iteration, the only difference is the value of the diagonal elements of $\mathbf{F}_j$. We will show the proof for the first iteration of the DTCCS method,

$$\hat{\rho}_j = \frac{1}{a_j}\mathbf{X}_j^T\mathbf{P}_j\mathbf{F}_j\mathbf{P}_j^T Y = \frac{1}{a_j}e_j^T\mathbf{X}^T\mathbf{P}_j\mathbf{F}_j\mathbf{P}_j^T\mathbf{X}\beta + \frac{1}{a_j}e_j^T\mathbf{X}^T\mathbf{P}_j\mathbf{F}_j\mathbf{P}_j^T\boldsymbol{\epsilon} := \frac{1}{a_j}(\xi_j + \eta_j),$$

where ξ_j is the signal, η_j is the noise, and $e_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is the jth coordinate vector. We know that for $\lambda \neq 0$, a_j is bounded and is of the same order

of $|\xi_j|$. Let $\xi = (\xi_1, \ldots, \xi_p)^T$ and $\eta = (\eta_1, \ldots, \eta_p)^T$. For showing the boundedness of $\|\xi\|$ and $\|\eta\|$, we ease the notation as $\xi_j = e_j^T \Xi_j \beta$, where $\Xi_j = \mathbf{X}^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T \mathbf{X}$.

Lemma 2.3.1. *(Fan and Lv 2008). Let $\mathcal{O}(n)$ be the set of $n \times n$ orthogonal matrices. A singular value decomposition of the $n \times p$ full row rank matrix $\mathbf{B}$ can be expressed as $\mathbf{B} = UDV$, where $U \in \mathcal{O}(n)$, $V \in \mathcal{O}(p)$, and $D = [D_{ij}]$ is an $n \times p$ matrix with $D_{ij} = 0$ for $i \neq j$ and $D_{11} \geq D_{22} \geq \cdots \geq D_{nn} > 0$. Let b_i^T denote the ith row of $\mathbf{B}$ for $i = 1, 2, \ldots, n$. We assume that the b_i^T are independent and orthogonally invariant, then the distribution of $\mathbf{B}$ is also invariant under $\mathcal{O}(p)$, i.e., $\mathbf{B}Q \overset{(d)}{=} \mathbf{B}$ for any $Q \in \mathcal{O}(p)$.*

Proposition 2.3.1. *Assume that the conditional number of Σ, $cond(\Sigma) \leq c_1 n^{\kappa}$ for some constants $c_1 > 0$ and $\kappa \in [0, 1]$ and that $\Sigma_{ii} = 1$ for $i = 1, 2, \ldots, p$, then we have*

$$d_\star(\Sigma) \geq c_1^{-1} n^{-\kappa} \text{ and } d^\star(\Sigma) \leq c_1 n^{\kappa}. \tag{2.12}$$

Proof. Note that $p = tr(\Sigma) = \sum_{i=1}^p d_i$ where d_i, $i = 1, \ldots, p$, are the eigenvalues of Σ, so we obtain that $d_\star(\Sigma) \leq 1$ and $d^\star(\Sigma) \geq 1$. Therefore,

$$d_\star(\Sigma) \geq \frac{1}{cond(\Sigma)} \text{ and } d^\star(\Sigma) \leq cond(\Sigma).$$

□

Let $vec(\mathbf{X})$ be the column vector stacked by all the rows of X. Then $vec(\mathbf{X}) \sim N(\mathbf{0}, I_n \otimes \Sigma)$ and $vec(\mathbf{B}) \sim N(\mathbf{0}, I_n \otimes I_p)$, where $\otimes$ denotes the Kronecker product of two matrices. That is, all the elements of $vec(X)$ are standard normal random variables and all the elements of $vec(\mathbf{B})$ are independent and identically distributed standard normal random variables.

Assumption 2.3.1. *We assume that $var(\mathbf{Y}) = O(1)$ and $\lambda = O(p)$ if λ is finite, and that the true model size $t_0 = c_0 n^{\nu}$ for the sparsity rate $\nu \in [0, 1]$.*

Assumption 2.3.2. *(Concentration Property, Fan and Lv 2008)*

Assume that each entry of the random matrix $\mathbf{B}$ is iid random variables with zero mean and unit variance and that $E|B_{11}|^4 < \infty$. As $n \to \infty$, we assume that $p \to \infty$

and $\frac{n}{p} \to r \in (0,1)$. *Matrix* $\mathbf{B}$ *is said to have the concentration property if there exist constants* c, $c_2 > 1$ *and* $C_1 > 0$ *such that*

$$P\left(d^\star(\frac{1}{\tilde{p}}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T) > c_2 \text{ and } d_\star(\frac{1}{\tilde{p}}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T) < 1/c_2\right) \leq e^{-C_1 n}$$

for any $n \times \tilde{p}$ *submatrix* $\tilde{\mathbf{B}}$ *of* $\mathbf{B}$ *with* $cn < \tilde{p} \leq p$.

Proposition 2.3.2. *(Interlacing inequalities for singular values, Queiró 1987)*

Let $\mathbf{B}$ *be an* $n \times p$ *matrix with rank* $t = \min(n,p)$, *the singular values of* $\mathbf{B}$ *are the square roots of the non-zero eigenvalues of the positive semidefinite matrix* $\mathbf{B}^T\mathbf{B}$ *(or* $\mathbf{B}\mathbf{B}^T$*). The sequence of singular values is* $\sigma_1(\mathbf{B}) \geq \cdots \geq \sigma_t(\mathbf{B}) > 0 = \cdots = 0$. *The relationship between any* $(n-s) \times (p-r)$ *submatrix* $\tilde{\mathbf{B}}$ *and* $\mathbf{B}$ *is*

$$\sigma_k(\mathbf{B}) \geq \sigma_k(\tilde{\mathbf{B}}) \geq \sigma_{k+r+s}(\mathbf{B}) \text{ for all } k \geq 1.$$

Lemma 2.3.2. *For any* $n \times \tilde{p}$ *submatrix* $\tilde{\mathbf{X}}$ *of* $\mathbf{X}$ *with* $n < \tilde{p} \leq p$, *non-zero eigenvalues of* $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ *(or* $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$*) are bounded by* $c_1^{-1}n^{-\kappa}d_\star(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T)$ *and* $c_1 n^\kappa d^\star(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T)$.

Proof. Let $\tilde{\Sigma}$ be a $\tilde{p} \times \tilde{p}$ submatrix of Σ. By Proposition 2.3.1 and 2.3.2,

$c_1^{-1}n^{-\kappa}d_\star(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T)I_n \leq \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \tilde{\mathbf{B}}\tilde{\Sigma}\tilde{\mathbf{B}}^T \leq c_1 n^\kappa d^\star(\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T)I_n$.

□

Proposition 2.3.3. *(Lemma 3 Moderate deviation of Fan and Lv 2008).*

Let $\chi_1^2, \ldots, \chi_n^2$ *be iid* χ_1^2*-distributed random variables. Then,*

(i) for any $\epsilon > 0$, *we have*

$$P\left(\frac{1}{n}\sum_{i=1}^n \chi_i^2 > 1+\epsilon\right) \leq e^{-A_\epsilon n},$$

where $A_\epsilon = [\epsilon - \log(1+\epsilon)]/2 > 0$.

(ii) for any $\epsilon \in (0,1)$, *we have*

$$P\left(\frac{1}{n}\sum_{i=1}^n \chi_i^2 < 1-\epsilon\right) \leq e^{-B_\epsilon n},$$

where $B_\epsilon = [-\epsilon - \log(1-\epsilon)]/2 > 0$.

Proposition 2.3.4. *Let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ denotes the ith natural base in the p dimensional space. Assume that the rows of the $p \times p$ orthogonal matrix $\mathbf{V}$ are random orthonormal p-frames, hence, $\mathbf{V}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^p)$. Let $\tilde{\mathbf{V}}^T$ be the top n rows of $\mathbf{V}$ with $\tilde{V} \in \mathbf{V}_n(\mathbb{R}^p)$. Then for any $C > 0$, there exist c_1', c_2' with $0 < c_1' < 1 < c_2'$, such that*

$$P\left(e_1^T \tilde{\mathbf{V}} \tilde{\mathbf{V}}^T e_1 < c_1' \frac{n}{p} \text{ or } e_1^T \tilde{\mathbf{V}} \tilde{\mathbf{V}}^T e_1 > c_2' \frac{n}{p}\right) \leq 4e^{-Cn}.$$

Proof. $\tilde{\mathbf{V}} = \mathbf{V}^T \begin{pmatrix} I_n \\ \cdots \\ \mathbf{0}_{p-n,n} \end{pmatrix}_{p \times n}$, and its transpose $\tilde{\mathbf{V}}^T = \begin{pmatrix} I_n & \mathbf{0}_{n,p-n} \end{pmatrix}_{n \times p} \mathbf{V}$.

Ve_1 is the first column of $\mathbf{V}$ and is uniformly distributed on a unit sphere. Let $\omega_i, i = 1, 2, \ldots, p$, be independent and identically distributed random variable from standard normal distribution, we have

$$\mathbf{V}e_1 \overset{(d)}{=} \left(\sqrt{\sum_{j=1}^{p} \omega_j^2}\right)^{-1/2} (\omega_1, \omega_2, \ldots, \omega_p)^T$$

and

$$\tilde{\mathbf{V}}^T e_1 \overset{(d)}{=} \left(\sqrt{\sum_{j=1}^{p} \omega_j^2}\right)^{-1/2} (\omega_1, \omega_2, \ldots, \omega_n)^T .$$

Hence $e_1 \tilde{\mathbf{V}} \tilde{\mathbf{V}}^T e_1 \overset{(d)}{=} \frac{\omega_1^2 + \cdots + \omega_n^2}{\omega_1^2 + \cdots + \omega_p^2}$, which is a random variable following a beta distribution with parameter $n/2$ and $(p-n)/2$.

By Proposition 2.3.3, we know that for any $C > 0$, there exists some $\epsilon_1, \epsilon_4 \in (0, 1)$, $\epsilon_2, \epsilon_3 > 0$ such that

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \omega_i^2 < 1 - \epsilon_1\right) \leq e^{-Cn}, \; P\left(\frac{1}{p}\sum_{i=1}^{p} \omega_i^2 > 1 + \epsilon_2\right) \leq e^{-Cp} < e^{-Cn},$$

and

$$P\left(\frac{1}{n}\sum_{i=1}^{n} \omega_i^2 > 1 + \epsilon_3\right) \leq e^{-Cn}, \; P\left(\frac{1}{p}\sum_{i=1}^{p} \omega_i^2 < 1 - \epsilon_4\right) \leq e^{-Cp} < e^{-Cn}.$$

Let $c_1' = \frac{1-\epsilon_1}{1+\epsilon_2}$ and $c_2' = \frac{1+\epsilon_3}{1-\epsilon_4}$. By Bonferroni's inequality, we have

$$P\left(e_1^T \tilde{\mathbf{V}} \tilde{\mathbf{V}}^T e_1 < c_1' \frac{n}{p} \text{ or } e_1^T \tilde{\mathbf{V}} \tilde{\mathbf{V}}^T e_1 > c_2' \frac{n}{p}\right) \leq 4e^{-Cn}.$$

□

Lemma 2.3.3. *Suppose Assumption 2.3.1 and 2.3.2 hold. Then for $C > 0$ and for any unit norm vector v, there exist constants c_3 and c_4 with $0 < c_3 < 1 < c_4$ such that*

$$P\left(|e_j^T \Xi_j v| < c_3 n^{1-\kappa/2} \text{ or } |e_j^T \Xi_j v| > c_4 n^{1+\kappa/2}\right) \leq 4e^{-Cn}.$$

Proof. Recall $\Xi_j = \mathbf{X}^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T \mathbf{X}$, $\mathbf{X} = \mathbf{B}\Sigma^{1/2}$ and $\mathbf{F}_j = \begin{pmatrix} \frac{\lambda}{\lambda + d_{j1}^2} & & \\ & \ddots & \\ & & \frac{\lambda}{\lambda + d_{jn}^2} \end{pmatrix}$.

There exist two $q \times q$ orthogonal matrices Q_1 and Q_2 that respectively rotate $\Sigma^{1/2} e_j$ and $\Sigma^{1/2} v$ to the same direction of e_1, i.e., $\Sigma^{1/2} e_j = \|\Sigma^{1/2} e_j\|_2 Q_1 e_1$ and $\Sigma^{1/2} v = \|\Sigma^{1/2} v\|_2 Q_2 e_1$. Then we have

$$|e_j^T \Xi_j v| = \|\Sigma^{1/2} v\| \cdot \|\Sigma^{1/2} e_j\| \cdot e_1^T Q_1^T \mathbf{B}^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T \mathbf{B} Q_2 e_1. \tag{2.13}$$

By Lemma 2.3.1, rows of $\mathbf{B}$ are independent and orthogonally invariant, then the distribution of $\mathbf{B}$ is also invariant under $\mathcal{O}(p)$, i.e., $\mathbf{B}Q \overset{(d)}{=} \mathbf{B}$ for any $Q \in \mathcal{O}(p)$. Rewrite $\mathbf{B} = U\, diag(D_{11}, \ldots, D_{nn}) \tilde{V}^T$, where $\tilde{V}^T = (I_n, \mathbf{0}_{n,p-n})_{n \times p} V$ and $\tilde{V} \in V_n(R^p)$. It is obvious that $p \cdot [d_\star(\frac{1}{p}\mathbf{B}\mathbf{B}^T)] I_n \leq diag(D_{11}^2, \ldots, D_{nn}^2) \leq p \cdot [d^\star(\frac{1}{p}\mathbf{B}\mathbf{B}^T)] I_n$.

For the norm of the vector $\|\Sigma^{1/2} v\|$ and $\|\Sigma^{1/2} e_j\|$, we have

$$d_\star(\Sigma) \leq v^T \Sigma v = \|\Sigma^{1/2} v\|^2 \leq d^\star(\Sigma) \text{ and } e_j^T \Sigma e_j = \|\Sigma^{1/2} e_j\|^2 = 1.$$

By Assumption 2.3.1, 2.3.2, Proposition 2.3.2 and Lemma 2.3.2, the diagonal of $\mathbf{F}_j$ is bounded by $O(1)$ and 1.

By Proposition 2.3.1 and Assumption 2.3.2, we have $c_1^{-1} n^{-\kappa} \leq d_\star(\Sigma) \leq d^\star(\Sigma) \leq c_1 n^\kappa$ and $P\left(d^\star(\frac{1}{\tilde{p}}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T) > c_2 \text{ and } d_\star(\frac{1}{\tilde{p}}\tilde{\mathbf{B}}\tilde{\mathbf{B}}^T) < 1/c_2\right) \leq e^{-C_1 n}$. Along with Proposition 2.3.4 and Bonferroni's inequality, we obtain

$$P\left(|e_j^T \Xi_j v| < c_3 n^{1-\kappa/2} \text{ or } |e_j^T \Xi_j v| > c_4 n^{1+\kappa/2}\right) \leq 4e^{-Cn}.$$

□

Lemma 2.3.4. *If Assumption 2.3.1 and 2.3.2 hold, then for any $C > 0$, there exists some $c_5, c_6 > 0$ such that for any $j = 1, 2, \ldots, p$,*

$$P\left(|e_j^T \Xi_j \beta| < c_5 n^{1-\kappa}\right) \leq O(e^{-Cn}),$$

and

$$P\left(|e_j^T \Xi_j \beta| > c_6 n^{1+\kappa+\nu/2}\right) \leq O(e^{-Cn}).$$

Proof. Let $\beta = \frac{\beta}{\|\beta\|_2} \cdot \|\beta\|_2$. Apply the result of Lemma 2.3.3, for $v = \frac{\beta}{\|\beta\|_2}$,

$$P\left(|e_j^T \Xi_j \frac{\beta}{\|\beta\|_2}| < c_3 n^{1-\kappa/2}\right) \leq O(e^{-Cn}).$$

With probability at least $1 - O(e^{-Cn})$, $|e_j^T \Xi_j \frac{\beta}{\|\beta\|_2}| \geq c_3 n^{1-\kappa/2}$. By Assumption 2.3.1 and Proposition 2.3.1, $var(\mathbf{Y}) = O(1)$ and $d^\star(\Sigma) \leq c_1 n^\kappa$. Then we have $c_1 n^\kappa \|\beta\|_2^2 \geq \|\beta\|_2^2 \, d^\star(\Sigma) \geq \beta^T \Sigma \beta = var(Y) - \sigma^2 \geq c_7$ for some constant c_7. Therefore, with probability at least $1 - O(e^{-Cn})$, $\|\beta\|_2 \geq c_7' n^{-\kappa/2}$ and $e_j^T \Xi_j \beta \geq c_5 n^{1-\kappa}$ for some constants c_7' and c_5 respectively. Hence, $P\left(\left|e_j^T \Xi_j \beta\right| < c_5 n^{1-\kappa}\right) \leq O(e^{-Cn})$.

Apply the result of Lemma 2.3.3, for $v = e_i$, $i = 1, 2, \ldots, p$,

$$P\left(|e_j^T \Xi_j e_i| < c_3 n^{1-\kappa/2} \text{ or } |e_j^T \Xi_j e_i| > c_4 n^{1+\kappa/2}\right) \leq 4e^{-Cn}.$$

We know that the true model size $t_0 = c_0 n^\nu$ from Assumption 2.3.1 and $c_1^{-1} n^{-\kappa} \|\beta\|_2^2 \leq \|\beta\|_2^2 \, d_\star(\Sigma) \leq \beta^T \Sigma \beta = var(Y) - \sigma^2$,

$$\begin{aligned} |e_j^T \Xi_j \beta| &= \left|\sum_{i \in \mathcal{T}} e_j^T \Xi_j e_i \beta_i\right| \leq \sum_{i \in \mathcal{T}} \{|e_j^T \Xi_j e_i| \cdot |\beta_i|\} \\ &\leq \sqrt{\sum_{i \in \mathcal{T}} |e_j^T \Xi_j e_i|^2} \cdot \|\boldsymbol{\beta}\|_2 \leq c_6 n^{1+\kappa+\nu/2}, \end{aligned} \tag{2.14}$$

with probability at least $1 - O(e^{-Cn})$. The first inequality in (2.14) is from the Jensen's inequality and the second one is from the Cauchy-Schwarz inequality.

Hence, $P\left(|e_j^T \Xi_j \beta| > c_6 n^{1+\kappa+\nu/2}\right) \leq O(e^{-Cn})$. □

Lemma 2.3.5. *Suppose Condition 2.3.2, and Assumption 2.3.1, 2.3.2 hold and let ν be the sparsity rate, i.e., $t_0 = c_0 n^\nu$ is the true model size. Assume $1+2\kappa > \alpha - \nu$ where α is defined in Condition 2.3.1. Then for any $C > 0$, there exists some constants c_8 and $c_8' > 0$ such that for any $j = 1, 2, \ldots, p$,*

$$P\left(|\eta_j| > c_8 n^{1+\kappa+\nu/2}\right) \leq c_\epsilon \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\}.$$

Also, there exists some small positive c_9, $c_9 = o(n^{\kappa-1})$, *such that for any* $j = 1, 2, \ldots, p$,

$$P\left(|\eta_j| < c_9 n^{1-\kappa}\right) \leq O(e^{-Cn}).$$

Proof. Define $\eta_j = a\epsilon$, where $a = e_j^T \mathbf{X}^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T$ and $\epsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$ where ϵ_i is the ith element of ϵ.

For the norm square of a, we have, with probability at least $1 - O(e^{-C_1 n})$,

$$\begin{aligned} \|e_j^T \mathbf{X}^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T\|_2^2 &= e_j^T \Sigma^{1/2} \mathbf{B}^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T \mathbf{B} \Sigma^{1/2} e_j \\ &\leq p \cdot \|\Sigma^{1/2} e_j\|_2^2 \cdot [d^\star(\frac{1}{p}\mathbf{B}\mathbf{B}^T)] \\ &= p \cdot [d^\star(\frac{1}{p}\mathbf{B}\mathbf{B}^T)]. \end{aligned} \tag{2.15}$$

By Assumption 2.3.2, the lower bound of $\|e_j^T \mathbf{X}^T \mathbf{P}_j \mathbf{F}_j \mathbf{P}_j^T\|_2^2$ is of the same order as the upper bound.

According to the sub-Gaussian tail assumption, we have $P\left(|a\epsilon| > t\right) \leq c_\epsilon \cdot \exp\left(-\frac{d_\epsilon t^2}{\|a\|^2}\right)$. By choosing $t = c_8 n^{1+\kappa+\nu/2}$,

$$P\left(|a\epsilon| > c_8 n^{1+\kappa+\nu/2}\right) \leq c_\epsilon \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\}, \tag{2.16}$$

where $c_8' \propto c_2 c_8^2 d_\epsilon$.

Since $\epsilon_i \sim N(0, \sigma^2)$, then $a\epsilon = \sum_{i=1}^n a_i \epsilon_i \sim N(0, \|a\|^2 \sigma^2)$. By the property of Normal distributions, $P\left(|a\epsilon| < c_9 n^{1-\kappa}\right) \to 0$ as $c_9 n^{1-\kappa} \to 0$. Thus,

$$P\left(|\eta_j| < c_9 n^{1-\kappa}\right) \leq O(e^{-Cn}).$$

□

Finally, we combine the results obtained in Lemma 2.3.4 and 2.3.5, together with Bonferroni's inequality, for some constants c_5', c_6' , $C > 0$,

$$P\left(\min_{j \in \mathcal{T}} |\rho_j| < c_5' n^{1-\kappa} \text{ or } \|\boldsymbol{\rho}\|_2^2 > c_6' n^{2(1+\kappa)+\nu+\alpha}\right) \leq O(\exp(-C_1 n)).$$

Main theorems

Theorem 2.3.1. *(Accuracy of DTCCS).*

Assume that Assumption 2.3.1 and 2.3.2 hold and that there exist positive constants c_8', c_ϵ and C_1 defined previously. Then there exists $\theta_n \in (0,1)$ such that

$$P\left(\mathcal{T} \subset \mathcal{M}_{\theta_n}\right) = 1 - t_0 c_\epsilon \cdot \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\} - O(t_0 \cdot \exp(-C_1 n)).$$

Proof. Applying Lemmas 2.3.4 and 2.3.5 to all $j \in \mathcal{T}$, for $t_0 = c_0 n^\nu$, we have

$$P\left(\min_{j\in\mathcal{T}} |\xi_j| < c_5' n^{1-\kappa}\right) = O(t_0 \cdot \exp(-C_1 n)),$$

and

$$P\left(\max_{j\in\mathcal{T}} |\eta_j| > c_8 n^{1+\kappa+\nu/2}\right) \leq t_0 c_\epsilon \cdot \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\}.$$

If we choose $\theta_n \in (c_0' n^{2\kappa+\nu/2}, 1)$, which is a rate between $\frac{c_8 n^{1+\kappa+\nu/2}}{c_5' n^{1-\kappa}}$ and 1, then we have

$$\begin{aligned} P\left(\min_{j\in\mathcal{T}} |\hat{\rho}_j| < \theta_n\right) &= P\left(\min_{j\in\mathcal{T}} |\xi_j + \eta_j| < a_j \theta_n\right) \\ &\leq P\left(\min_{j\in\mathcal{T}} |\xi_j| < c_5'' n^{1-\kappa}\right) + P\left(\max_{j\in\mathcal{T}} |\eta_j| > c_8'' n^{1+\kappa+\nu/2}\right) \\ &\leq O(t_0 \cdot \exp(-C_1 n)) + t_0 c_\epsilon \cdot \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\}. \end{aligned}$$

The first inequality holds since a_j is on the same order of $|\xi_j|$ for $\lambda \neq 0$ and the fact that if one event implies another, it has a smaller probability. The second inequality follows from Lemma 2.3.4 and 2.3.5. The detail of the first inequality is

$$\begin{aligned} \min_{j\in\mathcal{T}} |\xi_j + \eta_j| < a_j \theta_n &\Rightarrow \min_{j\in\mathcal{T}} |\xi_j| - \min_{j\in\mathcal{T}} |\eta_j| < a_j \theta_n \\ &\Rightarrow \min_{j\in\mathcal{T}} |\xi_j| - \max_{j\in\mathcal{T}} |\eta_j| < a_j \theta_n \\ &\Rightarrow \min_{j\in\mathcal{T}} |\xi_j| < M \text{ or } \max_{j\in\mathcal{T}} |\eta_j| > m, \end{aligned}$$

where M, m can reach $c_5'' n^{1-\kappa}$ and $c_8'' n^{1+\kappa+\nu/2}$ respectively.

Hence,

$$P\left(\mathcal{T} \subset \mathcal{M}_{\theta_n}\right) = 1 - t_0 c_\epsilon \cdot \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\} - O(t_0 \cdot \exp(-C_1 n)).$$

□

Theorem 2.3.2. *(Asymptotic sure screening).*

If Condition 2.3.1-2.3.4, Assumption 2.3.1, 2.3.2, Lemma 2.3.1-2.3.5 hold and $1 + 2\kappa > \alpha - \nu$*, then*

$$P(\mathcal{T} \subset \mathcal{M}_{\theta_n}) \to 1 \text{ as } n \to \infty, \tag{2.17}$$

i.e., the asymptotic sure screening property holds for DTCCS.

Proof. Apply Theorem 2.3.1, $t_0 \cdot \exp(-C_1 n) \leq (c_0 n)/(e^{C_1 n}) \to 0$ as $n \to \infty$.

Since $1+2\kappa > \alpha-\nu$, we have $\exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\} \to 0$ as $n \to \infty$. This completes the proof of Theorem 2.3.2.

□

Theorem 2.3.3. *(Screening consistency).*

Assume that with a large probability, $\log(p - t_0) = o(\min\{Cn, c_8' n^{2(1+\kappa)+\nu-\alpha}\})$ *for* C *and* c_8' *defined in Lemmas 2.3.4 and 2.3.5. Then with a large probability, we have*

$$P(\min_{j \in \mathcal{T}} |\hat{\rho}_j| \geq \theta_n \geq \max_{j \notin \mathcal{T}} |\hat{\rho}_j|) \to 1 \text{ as } n \to \infty, \tag{2.18}$$

where θ_n *is defined in Theorem 2.3.1.*

Proof. The same as Theorem 2.3.1,

$$P\left(\max_{j \notin \mathcal{T}} |\xi_j| > c_6' n^{1+\kappa+\nu/2}\right) \leq (p - t_0) \cdot O(e^{-Cn}),$$

and

$$P\left(\max_{j \notin \mathcal{T}} |\eta_j| > c_8' n^{1+\kappa+\nu/2}\right) \leq (p - t_0) \cdot \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\} + (p - t_0) \cdot \exp(-C_1 n).$$

Now, if θ_n is chosen as the same as in the Theorem 2.3.1, then

$$P\left(\min_{j \in \mathcal{T}} |\hat{\rho}_j| > \theta_n\right) \leq (p - t_0) \cdot \exp\{-c_8' n^{2(1+\kappa)+\nu-\alpha}\} + O((p - t_0) \cdot \exp(-C_1 n)).$$

Then, for $\log(p-t_0) = o(\min\{Cn, c_8' n^{2(1+\kappa)+\nu-\alpha}\})$ and combining with Theorem 2.3.1, we have

$$P(\min_{j \in \mathcal{T}} |\hat{\rho}_j| \geq \theta_n \geq \max_{j \notin \mathcal{T}} |\hat{\rho}_j|) \to 1 \text{ as } n \to \infty. \tag{2.19}$$

□

One practical issue for variable screening is how to determine the size of the submodel. As shown in the theory, as long as the size of the submodel is larger than the true model, the DTCCS method preserves the relevant predictors with a large probability. Thus, for $t_0 < n$, the solution path with increasing complexities can be transferred to the next stage of final model selection. The less computationally demanding option for the final model selection is to use a criterion under the QSR framework (Kim and Jeon 2016). Kim and Jeon (2016) showed the solution path that includes the true model converges to 1 by using the QSR framework.

2.4 Numerical Studies

Extensive simulation studies are conducted to assess the performance of the proposed DTCCS method with comparisons to widely used methods in the literature, such as the ISIS method (Fan and Lv 2008), the tilting approach (Cho and Fryzlewicz 2012) and the HOLP algorithm (Wang and Leng 2016). For implementation, we make use of the existing R package **SIS** for the ISIS procedure and **tilting** for the tilting method. For the SIS method, we use the marginal correlation to rank variables, and for the HOLP procedure, we use ridge-type HOLP with a submodel size n. Similar to the numerical criterion used in Cho and Fryzlewicz (2012) and Wang and Leng (2016), the screening accuracy of each method after some replications is defined as the proportion $P(\mathcal{T} \subset \mathcal{M}_s)$ where $\mathcal{M}_s$ denotes the submodel after each screening. We evaluate the methods by the frequencies of the selected models which contains all the variables of the true model. The screening accuracy of each method is reported as the proportion $P(\mathcal{T} \subset \mathcal{M}_s)$ in Table 2.1-2.3.

2.4.1 Simulation Studies

Generally, a strong correlation among the predictors and/or a small signal-to-noise ratio create difficulty in high dimensional variable screening. To assess the performance of the proposed method, we examine three scenarios. In the first scenario we highlight the advantage of the proposed DTCCS method in overcoming issues associated with

strong correlation among predictors, the second scenario examines the ability of the DTCCS method for dealing with collinearity, and the third scenario demonstrates the advantage of DTCCS in parsimonious interpretation. 200 replications of simulation are run for each scenario.

Scenario I: Compound Symmetry Structure of Σ

For the first scenario, we use model (1.2) with true $\boldsymbol{\beta} = (5, 5, 5, 0, \ldots, 0)^T$. In this model, $X_1, \ldots, X_p$ are p predictors and $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ is the noise that is independent of the predictors. In this simulation, a sample of $(X_1, \ldots, X_p)$ with size n was drawn from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with covariance matrix $\Sigma = (1 - \rho)I_p + \rho \mathbf{1}\mathbf{1}^T$, where $\mathbf{1} = (1, \ldots, 1)^T$. 16 models are generated by using $n = 50$, or 70, $p = 100$ or 1000, $\rho = 0$, 0.1, 0.5 or 0.9, respectively. For each model, three different values of σ^2 are chosen to obtain different SNR values defined as $SNR = \frac{\beta^T \Sigma \beta}{\sigma^2}$. Three levels of SNR are considered as 10, 20, and 30. This scenario modifies Example I of Fan and Lv (2008) with smaller values of SNR. Since many screening methods perform fairly well in a very high SNR setting and almost equally poorly in low SNR settings, we deliberately choose small SNR values here, for example, $SNR = 10$, and report the results in Table 2.1.

Table 2.1 shows that when the signal-to-noise ratio is low or the data are highly correlated, the DTCCS and HOLP methods outperform the ISIS and tilting approaches. All methods perform well when signal is strong or the data are weakly correlated.

Scenario II: Strong Correlation among the Predictors

In this scenario, we use model (1.2) with true $\boldsymbol{\beta} = (5, 5, 5, -15\sqrt{\rho}, 0, \ldots, 0)^T$. The predictors $X_1, \ldots, X_p$ and the noise ϵ are generated the same as in the first scenario, with the covariance matrix for the predictors being $\Sigma = (\sigma_{ij})_{p \times p}$ which has the same entries as in the first scenario except for the $4th$ row and the $4th$ column. For the $4th$ row and the $4th$ column, we replace $(\rho, \rho, \rho, 1, \rho, \ldots, \rho)^T$ with $(\rho_M, \rho_M, \rho_M, 1, 1 - \rho_M, \ldots, 1 - \rho_M)^T$ where ρ_M is the correlation of multicollinearity. Function *make.positive.definite* in package **corpcor** is used to guarantee a positive

Table 2.1: Screening Accuracy for Scenario I

			$p = 100$				$p = 1000$			
n	SNR	Method	ρ=0	ρ=0.1	ρ=0.5	ρ=0.9	ρ=0	ρ=0.1	ρ=0.5	ρ=0.9
50	10	DTCCS	1.000	1.000	0.995	0.885	0.980	0.995	0.965	0.290
		ISIS	1.000	0.980	0.890	0.030	0.800	0.860	0.220	0.000
		Tilting	1.000	1.000	0.920	0.030	0.980	0.990	0.590	0.000
		HOLP	1.000	1.000	1.000	0.930	0.990	1.000	0.960	0.240
	20	DTCCS	1.000	1.000	1.000	0.985	0.985	0.995	0.995	0.635
		ISIS	1.000	1.000	0.915	0.230	0.890	0.875	0.535	0.005
		Tilting	1.000	1.000	0.990	0.140	1.000	0.995	0.920	0.000
		HOLP	1.000	1.000	0.995	0.985	0.995	1.000	0.990	0.560
	30	DTCCS	1.000	1.000	1.000	1.000	0.990	0.995	0.995	0.795
		ISIS	1.000	1.000	1.000	0.450	0.900	0.880	0.630	0.010
		Tilting	1.000	1.000	1.000	0.310	1.000	1.000	0.980	0.040
		HOLP	1.000	1.000	1.000	0.980	0.990	1.000	1.000	0.760
70	10	DTCCS	1.000	1.000	1.000	0.975	1.000	1.000	1.000	0.585
		ISIS	1.000	1.000	1.000	0.130	1.000	0.990	0.770	0.000
		Tilting	1.000	1.000	0.990	0.010	1.000	1.000	0.870	0.000
		HOLP	1.000	1.000	1.000	0.990	1.000	1.000	1.000	0.540
	20	DTCCS	1.000	1.000	1.000	0.995	1.000	1.000	1.000	0.870
		ISIS	1.000	1.000	1.000	0.610	1.000	0.995	0.940	0.010
		Tilting	1.000	1.000	1.000	0.330	1.000	1.000	0.995	0.015
		HOLP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.825
	30	DTCCS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.940
		ISIS	1.000	1.000	1.000	0.850	1.000	1.000	0.940	0.170
		Tilting	1.000	1.000	1.000	0.700	1.000	1.000	1.000	0.170
		HOLP	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.920

definite covariance matrix. In this example, X_4 is uncorrelated with the response Y or irrelevant predictors but is strongly correlated with important predictors. Multicollinearity arises when ρ_M is close to 1. This example modifies Example II of Fan and Lv (2008) with greater values of ρ. ρ is set as 0.5, 0.6, 0.7, 0.8 or 0.9 in order to examine the difficulty induced by the strong correlation among the predictors.

Twenty models are generated by using $n = 50$ or 70 with $p = 100$ or 1000 and different ρ. The results are reported in Table 2.2. With the presence of collinearity, we found that the ISIS and tilting methods are not stable in dealing with collinearity, and the HOLP approach does not even work. However, the DTCCS method perform well. After comparing with the results of SIS, Tilting, HOLP and to the best of our knowledge, the proposed DTCCS method seems to be the only effective screening method to handle data with extreme multicollinearity.

Table 2.2: Screening Accuracy for Scenario II

			$\rho_M = 0.9$					$\rho_M = 1.0$				
p	n	Method	ρ=0.5	ρ=0.6	ρ=0.7	ρ=0.8	ρ=0.9	ρ=0.5	ρ=0.6	ρ=0.7	ρ=0.8	ρ=0.9
100	50	DTCCS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		ISIS	0.020	0.160	0.355	0.755	0.935	0.025	0.130	0.350	0.725	0.910
		Tilting	0.590	0.705	0.770	0.785	0.595	0.790	0.885	0.815	0.82	0.650
		HOLP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	70	DTCCS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		ISIS	0.215	0.435	0.745	0.960	0.995	0.185	0.515	0.860	0.945	0.995
		Tilting	0.880	0.915	0.945	0.940	0.855	0.935	0.950	0.965	0.955	0.935
		HOLP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1000	50	DTCCS	0.620	0.945	1.000	1.000	1.000	0.625	0.970	1.000	1.000	1.000
		ISIS	0.000	0.000	0.005	0.020	0.005	0.000	0.000	0.010	0.020	0.005
		Tilting	0.130	0.255	0.295	0.130	0.020	0.185	0.310	0.350	0.135	0.030
		HOLP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	70	DTCCS	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		ISIS	0.000	0.005	0.080	0.370	0.550	0.000	0.000	0.055	0.370	0.505
		Tilting	0.275	0.510	0.385	0.250	0.075	0.465	0.530	0.420	0.340	0.115
		HOLP	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Scenario III: Auto-Regressive Correlation

In the third scenario, we use model (1.2) with true $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^T$. The predictors $X_1, \ldots, X_p$ and the noise ϵ are again generated the same as in the first scenario, but having different covariance matrix for the predictors. The covariance matrix Σ has entries $\sigma_{ii} = 1$, $i = 1, \ldots, p$ and $\sigma_{ij} = \rho^{|i-j|}$, $i \neq j$. This example is modified from Example 1 of Tibshirani (1996) with ρ set at 0.5, 0.7 or 0.9 and SNR taken values at 10, 20 or 30. We use a two-stage procedure to show parsimonious interpretation of the $DTCCS$ method. After the variable screening is finished, RIC under the QSR framework (termed QRIC here) is applied to obtain a final model. We report the screening accuracy rate (SA) and the final model size (MS) for DTCCS+QRIC, SIS+QRIC and HOLP+QRIC respectively in Table 2.3. The $DTCCS$ method together with QRIC is always able to select a parsimonious model with almost perfect accuracy rate even when the SNR value is small or the data are highly correlated. For this scenario, the SIS method, the $HOLP$ approach and the $DTCCS$ method are all good in parsimonious interpretation by using QRIC. Through these three scenarios, we conclude that DTCCS is an efficient and effective variable screening algorithm in high-dimensional screening and sparse modeling.

Table 2.3: Screening Accuracy and Final Model Size for Scenario III

n	SNR	Method	$p = 100$						$p = 1000$					
			$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$		$\rho = 0.5$		$\rho = 0.7$		$\rho = 0.9$	
			SA[1]	MS[2]	SA	MS	SA	MS	SA	MS	SA	MS	SA	MS
50	10	DTCCS	1.000	4.425	1.000	4.600	0.930	4.730	0.935	6.580	0.985	5.285	0.900	5.525
		SIS	0.995	5.290	1.000	6.960	0.950	9.420	0.945	6.160	0.990	4.960	0.930	5.885
		HOLP	1.000	5.555	1.000	8.105	0.975	13.790	0.930	6.095	0.985	5.400	0.920	6.320
	20	DTCCS	1.000	3.975	1.000	4.400	0.990	4.730	0.955	5.570	0.990	4.495	0.970	4.805
		SIS	0.995	4.030	1.000	4.400	0.980	5.075	0.950	5.285	0.990	4.350	0.970	4.790
		HOLP	1.000	3.325	1.000	3.940	0.980	5.155	0.935	5.190	0.990	4.345	0.945	4.705
	30	DTCCS	1.000	3.880	1.000	4.250	0.990	4.700	0.965	5.255	0.990	4.300	0.995	4.765
		SIS	1.000	4.035	1.000	4.250	0.990	4.810	0.960	5.250	0.990	4.300	0.995	4.765
		HOLP	1.000	3.150	1.000	3.565	0.990	4.465	0.925	4.580	0.990	4.245	0.960	4.605
70	10	DTCCS	1.000	3.765	1.000	4.405	1.000	4.775	0.990	4.720	1.000	5.190	0.985	5.980
		SIS	1.000	4.565	1.000	6.090	1.000	10.310	0.985	4.525	1.000	5.270	0.985	6.240
		HOLP	1.000	5.625	1.000	9.390	1.000	15.230	0.990	4.665	1.000	5.515	0.980	7.375
	20	DTCCS	1.000	3.495	1.000	4.230	1.000	4.735	0.990	3.955	1.000	4.290	1.000	4.935
		SIS	1.000	3.505	1.000	4.175	1.000	5.000	0.990	3.915	1.000	4.300	1.000	9.945
		HOLP	1.000	3.215	1.000	3.365	1.000	5.070	0.990	3.715	1.000	4.085	0.995	4.940
	30	DTCCS	1.000	3.395	1.000	4.095	1.000	4.785	0.995	3.900	1.000	4.260	1.000	4.875
		SIS	1.000	3.405	1.000	4.105	1.000	4.855	0.990	3.815	1.000	4.260	1.000	4.875
		HOLP	1.000	3.025	1.000	3.245	1.000	4.480	0.990	3.600	1.000	4.035	0.995	4.780

[1] SA: screening accuracy rate; [2] MS: final model size.

2.4.2 Real Data Analysis

To illustrate the proposed method, we apply DTCCS method to the leukemia data which were reported by Golub et al. (1999). The complete dataset is available from `http://portals.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?paper_id=43`. Part of this microarray data can be found in the R package **plsgenomics** with name *leukemia* which has 3051 genes for 38 leukemia patients. Among the genes under this study, the expression level on gene $CST3$ exhibited the most significant difference for different types of leukemia (Sakhinia et al. 2005). The $CST3$ gene is believed to be linked to only a small number of genes in the leukemia study (Fang and Grzymala-Busse 2006; Tang et al. 2009). Hence, we consider $CST3$ as the effect of leukemia and take it as the continuous response in linear model (1.2). Our goal of this data analysis is to find other genes (3050 in total) whose expressions are correlated with that of gene $CST3$. We firstly apply the variable screening method, and then build the final model using QRIC criteria. For variable screening method, we consider the SIS method, the tilting algorithm, and the HOLP procedure from the literature together with the proposed DTCCS method for the purpose of comparison. Leave one out (LOO) technique is considered such that each observation in the sample is used once as the validation data. We obtain the variables after *screening+QRIC* procedure on the training set and then obtain the OLS estimator of those variables via a linear regression. To evaluate the prediction accuracy, square error $(Y - Y_i)^2$, $i = 1, \ldots, n$, is recorded for each validation observation. In Table 2.4, we report the means and the standard deviation (SD) of the square errors for prediction and the mean and median of selected model sizes from n training sets.

It can be seen from Table 2.4 that models selected by the proposed $DTCCS$ method and the SIS method have smaller cross-validation error than those selected by $HOLP$ and $Tilting$, which justifies that the proposed $DTCCS$ screening method keeps the useful variables in the screening procedure, while $HOLP$ or $Tilting$ may screen out some response relevant variables.

Table 2.4: Data Analysis of Leukemia Data (LOOCV)

Method	Mean square errors	SD square errors	Model size (mean)	Model size (median)
DTCCS	0.8257	1.4419	0.8257	2.0000
SIS	0.8257	1.4419	0.8257	2.0000
Tilting	1.9702	2.4890	1.9702	1.0000
HOLP	1.3410	1.7389	1.3410	2.0000

We also apply the *DTCCS* method, in contrast to the *SIS*, *Tilting*, *HOLP* approaches, to obtain a final model from the full data by first applying the screening methods and then obtaining the final model using QRIC criteria. Table 2.5 reports the variables (gene ID) selected in the final model using different approaches. The mean square error (MSE) and R^2 obtained after applying an OLS estimator to the final selected variables are also reported. We see that the proposed DTCCS method and the SIS method share the same MSE and R^2 and outperform the tilting algorithm and HOLP procedure.

Table 2.5: Final Models for Leukemia Full Data using Different Methods

Method	Variables (gene ID)	MSE	R^2
DTCCS	D88422, X62055	0.5878	0.7454
SIS	D88422, X62055	0.5878	0.7454
Tilting	HG1078	1.7558	0.2395
HOLP	L05624, U72509, D83920	0.7617	0.6701

2.5 Deterministic Extensions

The high-dimensional problems with random covariates have been studied for decades, but they are not well developed for the scenario when the covariates are from a deterministic design matrix. In this section, we extend DTCCS to the deterministic

design matrix. A key component in DTCCS is the choice of the tuning parameters λ's (the amount of tilting). The choice of λ's is connected with the selection of the step size d. The principle of the selection criterion has not been discussed for the random matrix design. We generally give a predetermined value of d and a sequence of λ's for different purposes of the data analysis, for instance, the screening accuracy or the numerical efficiency. Roughly speaking, small selection of d leads to the screening accuracy while big selection of d leads to the numerical efficiency. Different selections of d's and λ's would greatly influence the performance of the DTCCS for the random design. Under the deterministic design, we will discuss a fixed selection of λ's by a minimax procedure. Under the methodology of DTCCS, spurious correlation among predictors can be eliminated or minimized before forming a screening ranking. Our proposed method has the appeal in several aspects for deterministic design. It can retain the important predictors which have small marginal correlations with the response, and meanwhile, exclude unimportant predictors which have large correlation with the response. Thresholded regression is potentially feasible to determine the solution path of this extension, but it is still difficult to practice at this moment due to the computational burden. A practical counterpart is simply reducing the dimensionality from p to a moderate size, say n.

2.5.1 High-dimensional Deterministic Design Matrix

The sparse modeling means that the number of the relevant covariates is much smaller than n in a high-dimensional design matrix. High-dimensional but sparse vectors are commonly seen in large dataset. Zhao and Yu (2006) defined the sparseness for model selection: a model is sparse if only few regression coefficients are nonzero and those nonzero coefficients are uniformly bounded away from zero at a certain rate. Sparseness of $\boldsymbol{\beta}$ roughly guarantees that the model is identifiable (Candes and Tao 2007). Zhang and Huang (2008) gave a more general concept of sparseness: a model is sparse if most coefficients are small and the absolute sum of these small coefficients is below a certain level. Under this general sparsity assumption, variable selection is no longer separating nonzero and zero coefficient estimates, but determining a cut-off

threshold value of $\hat{t}$, $j = 1, \ldots, p$, in the sense that all coefficient estimates above $\hat{t}$ are preserved in the selected model with high probability. s_n denotes the generalized sparsity and its constraint for the property of L_p-consistent has been discussed in recent literature, see details in Meinshausen and Yu (2009) and Bickel et al. (2009).

The large design matrices can be viewed in two different ways: a probabilistic one and a nonprobabilistic one. In the probabilistic view, the design matrix is random and the random matrix theory has been discussed for decades. Most recent articles at the intersection of random matrix theory and high-dimensional statistics are focused on the concentration property (Lv 2013). In the seminal work of Fan and Lv (2008), the concentration property of a random matrix is the key to establish the sure screening property which means that this screening method keeps all important variables in the reduced feature space with asymptotic probability one. Fan and Lv (2008) proved that the concentration property holds when the design matrix is generated from Gaussian distribution and Lv (2013) proved that the concentration property holds when the design matrix is from a wide class of elliptical distributions. By using the concentration property, variable screening methods including SIS, HOLP and DTCCS have been developed.

Due to the identifiability of the high-dimensional regression parameter vector, estimation and variable selection/screening problems with deterministic design matrix are very different from those in the case with random design matrix. Deterministic design matrix did not attract enough attention as the random design matrix in traditional statistics. However, deterministic design matrix is more common in the era of high-dimensional statistics, such as modern biological research data and quantum phenomena data. The most popular deep learning methods, *convolutional neural network* (CNN, Rumelhart et al. 1985, Hinton and Salakhutdinov 2006) and *capsule network* (CN, Sabour et al. 2017), both use the deterministic design with a forward structure. Some recent articles discussed the deterministic design matrix in the high-dimensional context. Shao and Deng (2012) used an approach which focus on the projection of the regression parameter vector onto the row space generated by the deterministic design matrix. Lv (2013) derived general bounds on dimensionality

with some distance constraint on sparse models. Zhang and Zhang (2014) derived confidence intervals of high-dimensional regression coefficients, by using the flexible score vector with the residual from the sparse linear regression under deterministic design.

Consistent with the common procedure in high-dimensional deterministic design matrix (for instance, Zhang and Huang 2008, Zhang and Zhang 2014, Fan et al. 2018), we now standardize the response vector $\mathbf{Y}$ using the transformation $\mathbf{Y} - E(\mathbf{Y})$ and standardize the covariate column vectors $\mathbf{X}_j$ by the transformation $\sqrt{n}\,\mathbf{X}_j/\|\mathbf{X}_j\|_2$. Hence, all covariates are standardized to have an equal finite norm. Define the full model as $\mathcal{F} = \{1, \ldots, p\}$. Let (j) be the index of order statistics $|\beta_{(1)}| \geq \cdots \geq |\beta_{(p)}|$. Assume $\sum_{j=q+1}^{p} |\beta_{(j)}| \leq C_0$ where C_0 is a constant. Hence, there exists an index set $A_0 \subset \{1, \ldots, p\}$ such that $\#\{j \in \mathcal{F} : j \notin A_0\} = q$. Under this condition, there exists at most q 'large' coefficients and the rest 'small' coefficients are negligible. Let $A_1 = A_0^c$ be the 'true' parameter set and $\hat{A}_1$ be the selected sparse set. $\mathbf{X}_j$ is referred to as a relevant (or irrelevant) predictor if $j \in A_1$ (or $j \in A_0$). Let $k = |\hat{A}_1|$ denote the cardinality of $\hat{A}_1$ which is the model size. Hence, $|F| = p$, $|A_1| = q$.

2.5.2 Inferential Methods

Recall that in the linear regression model (1.2) with deterministic design matrix, each column of $\mathbf{X}$ is standardized and assumed to have a norm $\sqrt{n}$. The ϵ_i's are assumed to be independent and identically distributed (iid) random noise following a normal distribution $N(0,\, \sigma^2)$ with variance $\sigma^2 < \infty$.

In this section, we make the link between HDCE and the least square estimator of the classical linear models and develop DTCCS method to the deterministic design matrix.

DTCCS for High-dimensional Variable Screening with Deterministic Design Matrix

We first discuss the connection between HCDE and other classical parameter estimator. For $n > p$, let $\mathbf{X}_j^{\perp}$ be the projection of $\mathbf{X}_j$ to the orthogonal complement of the column space of $\tilde{\mathbf{X}}_{-j}$. For a linear model without interaction, the least square estimator for the jth regression coefficient can be expressed as

$$\hat{\beta}_j^{lse} = ({\mathbf{X}_j^{\perp}}^T Y)/({\mathbf{X}_j^{\perp}}^T \mathbf{X}_j), \tag{2.20}$$

for $n > p$, the scalar ${\mathbf{X}_j^{\perp}}^T \mathbf{X}_j \neq 0$, and $\mathbf{X}_j^{\perp} = (I - H_j)\mathbf{X}_j$ with $H_j = \tilde{\mathbf{X}}_{-j}(\tilde{\mathbf{X}}_{-j}^T \tilde{\mathbf{X}}_{-j})^{-1} \tilde{\mathbf{X}}_{-j}^T$ (Zhang and Zhang 2014). $\hat{\beta}_j^{lse}$ is the inner product $\langle \frac{(I-H_j)\mathbf{X}_j}{\mathbf{X}_j^T(I-H_j)\mathbf{X}_j}, (I - H_j)Y \rangle$. The normalized inner product can be denoted as $\frac{\mathbf{X}_j^T(I-H_j)Y}{\|(I-H_j)\mathbf{X}_j\|_2 \|(I-H_j)Y\|_2}$.

For the high-dimensional case $p > n$, $\mathbf{X}_j^{\perp}$ cannot be easily obtained anymore and Zhang and Zhang (2014) suggested to relax the orthogonal constraint of the vector $\mathbf{X}_j^{\perp}$ and proposed $\mathbf{X}_j^{\star}$ to solving ${\mathbf{X}_j^{\star}}^T(Y - \beta_j \mathbf{X}_j) = 0$ where $\mathbf{X}_j^{\star}$ is the residual vector from LASSO when regressing $\mathbf{X}_j$ against $\tilde{\mathbf{X}}_{-j}$. For the nonzero vector z_j that is not orthogonal to $\mathbf{X}_j$, the corresponding univariate linear regression estimator satisfies

$$\hat{\beta}_j^{(lin)} = (z_j^T Y)/(z_j^T \mathbf{X}_j) = \beta_j + \frac{\sum_{k \neq j} z_j^T \mathbf{X}_k \beta_k}{z_j^T \mathbf{X}_j} + \frac{z_j^T \epsilon}{z_j^T \mathbf{X}_j}. \tag{2.21}$$

Different from traditional linear regression with $n > p$, the bias is unavoidable in Eq. (2.21) since it is impossible to have a scalar $z_j^T \mathbf{X}_k = 0$ for all $k \neq j$ with nonzero vector z_j and note that $\boldsymbol{\beta}$ is generalized sparse. In Eq. (2.21), every nonzero $z_j^T \mathbf{X}_k$, $k \neq j$ linearly contributes the bias of β_j. To reduce the bias brought by all $\mathbf{X}_k$'s, $k \neq j$, Zhang and Zhang (2014) proposed a *low dimensional projection estimator* (LDPE) which uses a non-linear initial estimator $\hat{\beta}^{(init)}$ to be a bias correction:

$$\begin{aligned} \hat{\beta}_j &= \hat{\beta}_j^{(lin)} - \frac{\sum_{k \neq j} z_j^T \mathbf{X}_k \hat{\beta}_k^{(init)}}{z_j^T \mathbf{X}_j} \\ &= \frac{z_j^T Y}{z_j^T \mathbf{X}_j} - \frac{\sum_{k \neq j} z_j^T \mathbf{X}_k \hat{\beta}_k^{(init)}}{z_j^T \mathbf{X}_j}. \end{aligned} \tag{2.22}$$

We know that $Y = \sum_{j=1}^{p} \mathbf{X}_j \beta_j + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is the $n \times 1$ error vector. Combining Eq. (2.21) and Eq. (2.22), the bias can be decomposed to a noise term and a term

of approximation errors:

$$\begin{aligned}\left|\hat{\beta}_j - \beta_j\right| &= \left|\frac{z_j^T\boldsymbol{\epsilon}}{z_j^T\mathbf{X}_j} + \frac{\sum_{k\neq j} z_j^T\mathbf{X}_k(\beta_k - \hat{\beta}_k^{(init)})}{z_j^T\mathbf{X}_j}\right| \\ &\leq \tau_j \cdot \left(\frac{|z_j^T\boldsymbol{\epsilon}|}{\|z_j\|_2} + \|\zeta_j\|_\infty \cdot \|\beta - \hat{\beta}^{(init)}\|_1\right), \end{aligned} \tag{2.23}$$

where $\tau_j = \frac{\|z_j\|_2}{|z_j^T\mathbf{X}_j|}$ and vector $\zeta_j = \frac{z_j^T\tilde{\mathbf{X}}_{-j}}{\|z_j\|_2}$.

Since the nonzero vector z_j depends on the deterministic matrix $\mathbf{X}$ only, $\frac{z_j^T\boldsymbol{\epsilon}}{\|z_j\|_2} \sim N(0,\sigma^2)$. Hence, τ_j can be considered as the noise factor in Eq. (2.23). ζ_j can be considered as the bias factor since $\|\zeta_j\|_\infty \cdot \|\beta - \hat{\beta}^{init}\|_1$ controls the approximation error in Eq. (2.23). The Dantzig selector $\hat{\beta}_D$ can be used as the non-linear initial estimator and the boundedness of $\|\beta - \hat{\beta}_D\|_1$ will be discussed in Theorem 2.5.1. For getting the asymptotic normality and efficiency of estimation, we need ζ_j has a small infinity norm, $\|\beta - \hat{\beta}_D\|_1$ is bounded from above and τ_j is very small, that is

$$\|\zeta_j\|_\infty \cdot \frac{\|\beta - \hat{\beta}_D\|_1}{\sigma} = o(1) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\tau_j\sigma} \overset{(d)}{\approx} N(0,1), \tag{2.24}$$

where the symbol $\overset{(d)}{\approx}$ represents approximately identical in distribution.

In the proposed estimator HDCE, $z_j(\lambda) = \mathbf{P}_j F_j(\lambda)\mathbf{P}_j^T\mathbf{X}_j$ is an alternative residual vector for solving ${\mathbf{X}_j}^T\mathbf{P}_j F_j(\lambda)\mathbf{P}_j^T(Y - \beta_j\mathbf{X}_j) = 0$ for high-dimensional data. The solution is the inner product $\langle \frac{F_j^{1/2}(\lambda)\mathbf{P}_j^T\mathbf{X}_j}{\mathbf{X}_j^T(\mathbf{P}_j F_j(\lambda)\mathbf{P}_j^T)\mathbf{X}_j}, F_j^{1/2}(\lambda)\mathbf{P}_j^T Y\rangle$, and the normalized inner product is HDCE which can perform high-dimensional variable screening for deterministic design matrix.

An unnormalized version of HDCE with a bias correction can be naturally considered as a high-dimensional parameter estimator $\hat{\boldsymbol{\beta}}$. For getting the asymptotic normality of $\hat{\beta}_j(\lambda_j)$, $j = 1,\ldots,p$ we suggest a minimax procedure to determine the value of λ_j in a big but finite range, i.e. $\lambda_j \in (0, C_1)$.

$$\begin{aligned}\hat{\lambda}_j &= \underset{\lambda_j\in(0,C_1)}{\arg\min}\{\tau_j \cdot \|\zeta_j\|_\infty\} \\ &= \underset{\lambda_j\in(0,C_1)}{\arg\min}\left\{\left\|\frac{z_j^T(\lambda_j)\tilde{\mathbf{X}}_{-j}}{z_j^T(\lambda_j)\mathbf{X}_j}\right\|_\infty\right\}. \end{aligned} \tag{2.25}$$

After the minimax procedure, we obtain a general sparse vector $\hat{\boldsymbol{\beta}}$. Small but not exactly zero components of $\hat{\boldsymbol{\beta}}$ do not contribute much in estimation but add variability. $\hat{\boldsymbol{\beta}}$ can be viewed as a generalized sparse vector and can separate the relevant and irrelevant predictor variables efficiently by either choosing a predetermined subset number, such as n or using a threshold value to determine how small is 'negligible'.

Thresholded Regression on DTCCS

Recall penalized regression Eq. (1.7), $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|Y - \mathbf{X}\boldsymbol{\beta}\|_2^2 + J_\lambda(|\boldsymbol{\beta}|)$. Under the orthonormal setting, the hard thresholding rule takes the penalty function $J_\lambda(|\beta|) = \lambda^2 - (|\beta| - \lambda)^2 \mathbb{I}(|\beta| < \lambda)$ and obtain the hard thresholding function $\hat{\beta}_j^{(thr)} = \hat{\beta}_j \cdot \mathbb{I}(|\hat{\beta}_j| > \lambda)$ where $\hat{\beta}_j$ is the jth usual least squares estimate. The soft thresholding rule takes the L_1 penalty function to obtain the soft thresholding function $\hat{\beta}_j^{(thr)} = sgn(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$ which is the well known LASSO solution. The SCAD penalty brings a continuous thresholding function which connects the hard and soft thresholding functions (Fan and Li 2001).

Recently, Zhang and Huang (2008), Shao and Deng (2012), Zhang and Zhang (2014) and Zheng et al. (2014) performed thresholded regression by using different thresholding rules. Shao and Deng (2012) proposed thresholded ridge regression to discriminate large and small elements of $\boldsymbol{\beta}$ and pointed out that the generalized sparse model is identifiable if and only if the estimator is lies in a set having a one-to-one correspondence with $row(\mathbf{X})$. One advantage of using $row(\mathbf{X})$ in high-dimensional statistics is that the dimension of $row(\mathbf{X})$ is at most n which is much smaller than p. Wang and Leng (2016)'s HOLP method projects $\boldsymbol{\beta}$ onto $row(\mathbf{X})$ to obtain $\hat{\boldsymbol{\beta}}_{holp}$ which consists of many negligible components. Hence, $\hat{\boldsymbol{\beta}}_{holp}$ can also be viewed as a sparse vector and can separate the relevant and irrelevant predictor variables efficiently by a predetermined subset number, such as n or $\log(n)$.

Using a predetermined subset number is an efficient way to do variable screening, but a more computationally expensive thresholded DTCCS is also theoretically feasible. Thus, we would like to carry out a hard thresholding procedure as defined in Zhang and Huang (2008) and Zheng et al. (2014), that is, the negligible components

of $\hat{\beta}$ are forced to be 0. Hence, a threshold a_n is required for discriminating the components of $\hat{\boldsymbol{\beta}}$. Let $\hat{\beta}_j$ be the jth component of $\hat{\boldsymbol{\beta}}$, $j = 1, \ldots, p$. We use an indicator function to map $\hat{\boldsymbol{\beta}}$ to $\tilde{\boldsymbol{\beta}}$ whose jth component $\tilde{\beta}_j = \hat{\beta}_j$ if $|\hat{\beta}_j| > a_n$ and $\tilde{\beta}_j = 0$ if $|\hat{\beta}_j| \leq a_n$, where

$$a_n = C_2 n^{-\gamma},\ 0 < \gamma \leq 1/2,\ C_2 > 0, \tag{2.26}$$

is the thresholding value with constants C_2 and γ. This thresholding stage performs a variable selection procedure to keep the $\hat{\beta}_j$'s when they are greater than the thresholding value, and force the negligible components to be zero. To apply thresholding, we need to select the value of C_1 in Eq. (2.25) and C_2 in Eq. (2.26) and set γ in Eq. (2.26) be a fixed number in $(0, 1/2]$. Similar to many high-dimensional problems, C_1 and C_2 can be viewed as the tuning parameters. Let $\Psi(C)$ be the average prediction mean squared error when $C = (C_1, C_2)$ is used in λ and a_n. It is possible to use a data-driven method of find values of tuning parameters by minimizing $\Psi(C)$,

$$\hat{\Psi}(C) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T \tilde{\beta}^{(i)}(C))^2, \tag{2.27}$$

where $\tilde{\beta}^{(i)}(C)$ is the thresholded estimator of $\hat{\beta}^{(i)}(C)$ which is based on the dataset with $(\mathbf{x}_i^T, y_i)$ removed, $i = 1, \ldots, n$. This thresholded DTCCS is connecting to the classical leave-one-out cross validation (LOOCV), but LOOCV is almost not executable due to high-dimensional computation burden. Limited by current computational ability, we mainly focus on the DTCCS of the deterministic design matrix by selecting a predetermined subset number.

Conditions and Lemmas

In this subsection, we present regularity conditions and main lemmas.

Condition 2.5.1. *(Polynomial high-dimensional). $p > n$ and $p_n = O(n^\alpha)$ for some $\alpha > 0$.*

Condition 2.5.2. *(Generalized Sparsity). The generalized sparsity s_n satisfies $s_n \geq q$, $s_n = o(\sqrt{n})$ and $\frac{s_n^2 \log p_n}{n} \to 0$ as $n \to \infty$. The positive integer s_n plays the role of an upper bound on the generalized sparsity of a vector of coefficients $\boldsymbol{\beta}$.*

Condition 2.5.3. *(Sub-Gaussian tail condition, Kim and Jeon 2016.). In the linear model (1.2), the ϵ_i are independent random variables whose common distribution has a sub-Gaussian tail. That is, there is some $b > 0$ such that for every $t \in \mathbb{R}$, we have $E(e^{t\epsilon_i}) \leq \exp\{b^2t^2/2\}$, which implies that there exist positive constants c_ϵ and d_ϵ such that*

$$P\left(\left|\sum_{i=1}^{n} a_i\epsilon_i\right| > t\right) \leq c_\epsilon \cdot \exp\left(-\frac{d_\epsilon t^2}{\sum_{i=1}^{n} a_i^2}\right) \tag{2.28}$$

for all $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ and $t > 0$.

Condition 2.5.4. *(Tilting parameter). Let λ be the tilting parameter introduced in Section 2.5.2, $\lambda = O(p)$ for the finite selection of λ.*

Definition 2.5.1. *(Dantzig selector, Candes and Tao 2007). The Dantzig selector for linear model (1.2) can be formulated as the solution to the following convex program,*

$$\hat{\boldsymbol{\beta}}_D = \underset{\boldsymbol{\beta}\in\Lambda}{\arg\min} \|\hat{\boldsymbol{\beta}}\|_1, \tag{2.29}$$

where where Λ is the set of all $\boldsymbol{\beta}$'s which satisfies the Dantzig constraint:

$$\left\|\frac{1}{n}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right\|_\infty \leq r. \tag{2.30}$$

Let $\delta = \hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}$ and c_δ be a positive constant. Let δ_{A_1} denote the vector in $\mathbb{R}^p$ that has the same coordinates as δ on A_1 and zero coordinates on the complement of A_1 which is A_0. Bickel et al. (2009) proved that $\|\delta_{A_0}\|_1 \leq c_\delta\|\delta_{A_1}\|_1$ with a suggestion $c_\delta = 1$ for Dantzig selector when p is large.

Definition 2.5.2. *(Rayleigh–Ritz ratio). For a Gram matrix $\mathbf{M} = \frac{1}{n}\mathbf{X}^T\mathbf{X}$ and nonzero vector V, the Rayleigh–Ritz ratio is defined as:*

$$R(\mathbf{M}, V) = \frac{V^T\mathbf{M}V}{V^TV}, \tag{2.31}$$

and the Rayleigh–Ritz ratio is bounded by the maximum and minimum eigenvalues of $\mathbf{M}$. *Discussing the value of Rayleigh–Ritz ratio is equivalent to study the value of* $\frac{\|\mathbf{X}V\|_2}{\sqrt{n}\|V\|_2}$.

The Rayleigh–Ritz ratio and the spiritually similar conditions have been widely used in sparse modeling in recent articles. Zhang and Huang (2008) defined the *sparse Riesz condition* (SRC) which limits the range of the eigenvalues of the Gram matrix given by the subdesign matrix of a fixed number of covariates. Bickel et al. (2009) introduced the *restricted eigenvalue condition* which uses the subvector of V in the denominator of Eq. (2.31). Zheng et al. (2014) proposed a weaker condition termed *robust spark* which set a lower positive bound by using a Gram matrix given by the subdesign matrix in Eq. (2.31). In this section, we are considering the *restricted eigenvalue condition.*

Condition 2.5.5. *(Restricted Eigenvalue (RE) Condition, Bickel et al. 2009; Cai et al. 2017).*

For p_n, s_n *defined in Condition (2.5.1) and (2.5.2), and a positive number* c_δ *introduced in Definition 2.5.1, the following condition holds:*

$$\kappa(s_n, c_\delta) \stackrel{\triangle}{=} \min_{\substack{|A_1|\le s_n,\\ \delta\in\mathbb{R}^p,\, \delta\neq\mathbf{0},\\ \|\delta_{A_0}\|_1\le c_\delta\|\delta_{A_1}\|_1}} \frac{\|\mathbf{X}\delta\|_2}{\sqrt{n}\|\delta_{A_1}\|_2} > 0. \tag{2.32}$$

For $p > n$, the Gram matrix $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is degenerate which means $\min_{\delta\in\mathbb{R}^p,\,\delta\neq\mathbf{0}} \frac{\|\mathbf{X}\delta\|_2}{\sqrt{n}\|\delta\|_2} = 0$. Under the restriction defined in Condition (2.5.4), $\|\delta_{A_1}\|_2 < \|\delta\|_2$. Hence, there exists positive $\min_{\substack{|A_1|\le s_n,\\ \delta\in\mathbb{R}^p,\,\delta\neq\mathbf{0},\\ \|\delta_{A_0}\|_1\le c_\delta\|\delta_{A_1}\|_1}} \frac{\|\mathbf{X}\delta\|_2}{\sqrt{n}\|\delta_{A_1}\|_2} > \min_{\delta\in\mathbb{R}^p,\,\delta\neq\mathbf{0}} \frac{\|\mathbf{X}\delta\|_2}{\sqrt{n}\|\delta\|_2}$.

Lemma 2.5.1. *For the linear regression model (1.2),* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, *where* $\boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$ *and* $\mathbf{X}$ *is a full row rank deterministic matrix. Let* $\boldsymbol{\beta}_D\,(\in \mathbb{R}^p)$ *satisfy the Dantzig constraint (2.30) with* $r = c\sigma\sqrt{\frac{\log p}{n}}$ *where* $c > \sqrt{2}$. *Let* $\delta = \hat{\boldsymbol{\beta}}_D - \boldsymbol{\beta}$. *Then, with probability of at least* $1 - 2p^{1-c^2/2}$, *we have*

$$\left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}\delta\right\|_\infty \le 2r. \tag{2.33}$$

Proof. Consider the event

$$\mathcal{B} = \left\{ \left\| \frac{1}{n} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right\|_\infty \leq r \right\} = \left\{ \left\| \frac{1}{n} \mathbf{X}^T \boldsymbol{\epsilon} \right\|_\infty \leq r \right\} = \bigcap_{j=1}^{p} \{ |\frac{1}{n} \mathbf{X}_j^T \boldsymbol{\epsilon}| \leq r \}. \quad (2.34)$$

Applying the sub-Gaussian tail condition for standard normal distribution, we find that the probability of the complementary event $\mathcal{B}^c$ satisfies

$$\begin{aligned} P(\mathcal{B}^c) &\leq \sum_{j=1}^{p} P\{ |\frac{1}{n} \mathbf{X}_j^T \boldsymbol{\epsilon}| > r \} \leq p \cdot P\{ |\eta| \geq \frac{\sqrt{n} r}{\sigma} \} \\ &\leq 2p \cdot \exp\left(-\frac{n r^2}{2\sigma^2} \right) = 2p \cdot \exp\left(-\frac{c^2 \log p}{2} \right) = 2p^{1 - c^2/2}, \end{aligned} \quad (2.35)$$

where $\eta = \frac{\mathbf{X}_j^T \boldsymbol{\epsilon}}{\sqrt{n}\sigma} \sim N(0,1)$ for $j = 1, \ldots, p$. Hence, $P(\mathcal{B}) \geq 1 - 2p^{1-c^2/2}$.

Let $\hat{\boldsymbol{\beta}}_D$ be the solution of the Dantzig selector, it satisfies the event with very large probability

$$\mathcal{D} = \left\{ \left\| \frac{1}{n} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_D) \right\|_\infty \leq r \right\}. \quad (2.36)$$

Since event $\mathcal{B}$ together with $\mathcal{D}$ implies $\mathcal{E} = \left\{ \left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} \delta \right\|_\infty \leq 2r \right\}$, we conclude with probability of at least $1 - 2p^{1-c^2/2}$, we have $\left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} \delta \right\|_\infty \leq 2r$. □

Definition 2.5.3. *(Selection Consistency).*

Let A_1 be the set of indices of 'large' components of $\boldsymbol{\beta}$, and let $\hat{A}_1$ be the set of indices of components of $\boldsymbol{\beta}$ selected using a variable selection method. The variable selection method or $\hat{A}_1$ is said to be selection consistent if and only if

$$\lim_{n \to \infty} P(A_1 \subseteq \hat{A}_1) = 1. \quad (2.37)$$

Shao and Deng (2012) pointed out that for deterministic matrix $\mathbf{X}$, the selection consistency (2.37) is generally not achievable if $\boldsymbol{\beta}$ is not identifiable and they propose a lemma which related to the row space of $\mathbf{X}$ to reveal the identifiable $\boldsymbol{\beta}$'s.

Lemma 2.5.2. *(Identifiability of $\boldsymbol{\beta}$, Shao and Deng 2012).*

For a full row rank $n \times p$ deterministic matrix $\mathbf{X}$ defined in model (1.2), $p > n$ and the rank of $\mathbf{X}$ is n. Performing a singular value decomposition of $\mathbf{X}$,

$$\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{Q}^T = \mathbf{P}\mathbf{D}\tilde{\mathbf{Q}}^T \mathbf{D}_1^{-1}, \quad (2.38)$$

where $\mathbf{P}$ *is an* $n \times n$ *matrix satisfying* $\mathbf{P}^T\mathbf{P} = \mathbf{I}_n$, $\mathbf{Q}$ *is a* $p \times n$ *matrix satisfying* $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n$, $\mathbf{I}_n$ *denotes the* $n \times n$ *identity matrix,* $\mathbf{D}$ *is an* $n \times n$ *diagonal matrix of full rank and* $\mathbf{D}_1$ *is an* $n \times n$ *diagonal matrix with all positive entries.* $\mathbf{D}_1\mathbf{X}$ *is each row of* $\mathbf{X}$ *multiplying the corresponding diagonal entry in* $\mathbf{D}_1$. *Hence,* $\mathbf{D}_1\mathbf{X}$ *and* $\mathbf{X}$ *have the same basis for the row space. Let* $\mathbf{Q}_\perp$ *be a* $p \times (p-n)$ *matrix such that* $\mathbf{Q}^T\mathbf{Q}_\perp = \mathbf{0}$ *where* $\mathbf{0}$ *is a* $n \times (p-n)$ *matrix of all* 0*'s. Under this design matrix,* $\boldsymbol{\beta}$ *is identifiable if and only if there exists a known function* ϕ *from* $\mathbb{R}^n$ *to* $\mathbb{R}^{p-n}$ *such that*

$$\mathcal{G} = \{\boldsymbol{\beta} : \boldsymbol{\beta} = \mathbf{Q}\xi + \mathbf{Q}_\perp\phi(\upsilon),\, \upsilon \in \mathbb{R}^n\}\,, \tag{2.39}$$

which means identifiable $\boldsymbol{\beta}$*'s must be in a set having a one-to-one correspondence with the row space* $\text{row}(\mathbf{X}) = \{\mathbf{Q}\upsilon,\, \upsilon \in \mathbb{R}^n\}$).

Shao and Deng (2012) and Wang and Leng (2016) showed ridge estimator, thresholded ridge and HOLP, ridge HOLP are always in $row(\mathbf{X})$. In the next section, we will show that low dimensional projection estimator (LDPE) of DTCCS is in a set having a one-to-one correspondence with the row space $row(\mathbf{X}) = \{\mathbf{Q}\upsilon,\, \upsilon \in \mathbb{R}^n\}$) by using the relationship between HOLP and DTCCS.

Main Theorems

Theorem 2.5.1. *Let* ϵ_i, $i = 1, \ldots, n$, *be independent* $N(0, \sigma^2)$ *random variables with finite* $\sigma^2 > 0$, *let deterministic design matrix* $\mathbf{X}$ *with equal* L_2 *norm* $\sqrt{n}$ *for each column. Then, all the diagonal elements of the Gram matrix* $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ *are* 1. *Let Condition (2.5.1)- (2.5.4) be satisfied. Consider the Dantzig selector* $\hat{\boldsymbol{\beta}}_D$ *defined by Definition (2.5.1) with* $r = c\sigma\sqrt{\frac{\log p}{n}}$ *where* $c > \sqrt{2}$. τ_j *and* ζ_j *are defined in Eq. (2.23). Then with probability at least* $1 - 2p^{1-c^2/2}$, *the low dimensional projection estimator (LDPE) for DTCCS has asymptotic Normal distribution*

$$\frac{\hat{\beta}_j - \beta_j}{\tau_j\sigma} \overset{(d)}{\approx} N(0, 1). \tag{2.40}$$

Proof. Using Lemma 2.5.1, with probability at least $1 - 2p^{1-c^2/2}$, $\left\|\frac{1}{n}\mathbf{X}^T\mathbf{X}\delta\right\|_\infty \leq 2r$.

Together with restricted eigenvalue condition 2.5.4, we have

$$\begin{aligned}\kappa^2(s_n, c_\delta)\|\delta_{A_1}\|_2^2 \leq \frac{1}{n}\|\mathbf{X}\delta\|_2^2 &= \frac{1}{n}\delta^T\mathbf{X}^T\mathbf{X}\delta \\ &\leq \frac{1}{n}\|\mathbf{X}^T\mathbf{X}\delta\|_\infty\|\delta\|_1 \\ &\leq 2r(\|\delta_{A_1}\|_1 + \|\delta_{A_0}\|_1) \\ &\leq 2(1+c_\delta)r\|\delta_{A_1}\|_1 \\ &\leq 2(1+c_\delta)r\sqrt{s_n}\|\delta_{A_1}\|_2. \end{aligned} \tag{2.41}$$

From Eq. (2.41),

$$\|\delta_{A_1}\|_2 \leq 2(1+c_\delta)r\sqrt{s_n}/\kappa^2(s_n, c_\delta), \text{ and } \frac{1}{n}\|\mathbf{X}\delta\|_2^2 \leq 4(1+c_\delta)^2r^2s_n/\kappa^2(s_n, c_\delta).$$

Since $\|\delta_{A_1}\|_0 \leq s_n$, using the relationship of L_1 and L_2 norm, we have

$$\|\delta\|_1 \leq (1+c_\delta)\|\delta_{A_1}\|_1 \leq (1+c_\delta)\sqrt{s_n}\|\delta_{A_1}\|_2 \leq 2(1+c_\delta)^2rs_n/\kappa^2(s_n, c_\delta). \tag{2.42}$$

The LDPE of DTCCS with $\hat{\boldsymbol{\beta}}_D$ as the bias correction can be bounded as

$$\begin{aligned}\left|\hat{\beta}_j - \beta_j\right| &\leq \tau_j \cdot \left(\frac{|z_j^T\boldsymbol{\epsilon}|}{\|z_j\|_2} + \|\zeta_j\|_\infty \cdot \|\beta - \hat{\beta}_D\|_1\right) \\ &\leq \tau_j \cdot \left(\frac{|z_j^T\boldsymbol{\epsilon}|}{\|z_j\|_2} + \|\zeta_j\|_\infty \cdot \frac{2(1+c_\delta)^2s_n}{\kappa^2(s_n, c_\delta)} \cdot c\sigma\sqrt{\frac{\log p}{n}}\right). \end{aligned} \tag{2.43}$$

Using the minimax procedure defined in Section 2.5.2 and generalized sparsity s_n defined in Condition (2.5.2),

$$\|\zeta_j\|_\infty \cdot \frac{\|\beta - \hat{\beta}_D\|_1}{\sigma} = o(1).$$

Hence, with large probability, $\left|\hat{\beta}_j - \beta_j\right| \approx \tau_j \cdot \frac{|z_j^T\boldsymbol{\epsilon}|}{\|z_j\|_2}$ which is

$$\frac{\hat{\beta}_j - \beta_j}{\tau_j\sigma} \overset{(d)}{\approx} N(0,1),$$

where the symbol $\overset{(d)}{\approx}$ represents approximately identical in distribution. □

Theorem 2.5.2. *(Identifiability of HDCE). The* high-dimensional correlation estimator *(HDCE) is in a set having a one-to-one corresponding with row space of* $\mathbf{X}$*, i.e., the HDCE is identifiable.*

Proof. Recall HDCE with tuning parameter $\lambda = 1$,

$$\begin{aligned}\hat{\rho}_j(\lambda) &= \frac{1}{a_j}\mathbf{X}_j^T(\mathbf{I}_n - H_j)\mathbf{Y} \\ &= \frac{1}{a_j}\mathbf{X}_j^T(\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T + \mathbf{I}_n)^{-1}\mathbf{Y}.\end{aligned}$$

Recall ridge-HOLP with tuning parameter $\lambda = 1$,

$$\hat{\boldsymbol{\beta}}_{holp} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}\mathbf{Y}.$$

$\mathbf{X}(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}\mathbf{X}$ is the projection matrix for ridge-HOLP with diagonal entries $h_{jj} = \mathbf{X}_j^T(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}\mathbf{X}_j$ for $j = 1, \ldots, p$. Let $\theta_j = \mathbf{X}_j^T(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}\mathbf{X}\boldsymbol{\beta}$ and θ_j is in the row space of $\mathbf{X}$.

For $j = 1, \ldots, p$, $\mathbf{X}\mathbf{X}^T + \mathbf{I}_n = \tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T + \mathbf{X}_j\mathbf{X}_j^T + \mathbf{I}_n$, apply Sherman-Morrison formula (Sherman and Morrison 1950):

$$\begin{aligned}(\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T + \mathbf{I}_n)^{-1} &= \left[\left(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n\right)\left\{\mathbf{I}_n - (\mathbf{X}\mathbf{X}^T + \mathbf{I})^{-1}\mathbf{X}_j\mathbf{X}_j^T\right\}\right]^{-1} \\ &= \left(\mathbf{I}_n + \frac{(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}\mathbf{X}_j\mathbf{X}_j^T}{1 - \mathbf{X}_j^T(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}\mathbf{X}_j}\right)\left(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n\right)^{-1} \\ &= \left(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n\right)^{-1} + \frac{(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}\mathbf{X}_j\mathbf{X}_j^T(\mathbf{X}\mathbf{X}^T + \mathbf{I}_n)^{-1}}{1 - h_{jj}}.\end{aligned}$$

Hence, $\mathbf{X}_j^T(\tilde{\mathbf{X}}_{-j}\tilde{\mathbf{X}}_{-j}^T + \mathbf{I}_n)^{-1}\mathbf{X}\beta = \theta_j \cdot \frac{1}{1-h_{jj}}$ which means HDCE is in a set having a one-to-one corresponding with row space of $\mathbf{X}$, i.e., the HDCE is identifiable by Lemma 2.5.2. □

Through Theorem 2.5.1 and 2.5.2 together with Theorems 2.3.1-2.3.3, we conclude that DTCCS for the deterministic design matrix is reliable and follows sure screening and consistency properties.

2.5.3 Simulation Studies

In this section, we set $n = 100$ and $p = 1000$. We firstly generate initial design matrix $\mathbf{G} = (\mathbf{G}_1, \ldots, \mathbf{G}_p) \in \mathbb{R}^{n\times p}$. Although $\mathbf{G}$ is assumed to be a deterministic design matrix, we can only simulate it from a certain known distribution such as the

normal distribution. Given a particular $\rho \in (-1, 1)$, simulate the rows of $\mathbf{G}$ from $N(0, \Sigma)$ with $\Sigma = (1-\rho)I_p + \rho \mathbf{1}\mathbf{1}^T$. The standardized deterministic design matrix is $\mathbf{X}$ with the jth column $\mathbf{X}_j = \sqrt{n}\,\mathbf{G}_j / \|\mathbf{G}_j\|_2$. $(\mathbf{X}, \mathbf{Y})$ is defined in Eq. (1.2) with $\sigma = 1$. Given a particular $\alpha > 1$, let $\beta_j = 3\sqrt{(2/n)\log(p)}$, $j = 1, 500, 1000$, and $\beta_j = 3\sqrt{(2/n)\log(p)}/\alpha^j$ for all other j. From Figure 2.1, we found that for different values of ρ's, the majority bulk of $\frac{\beta}{\tau}$ are on or near the QQ-line which verifies the result of Theorem 2.5.1.

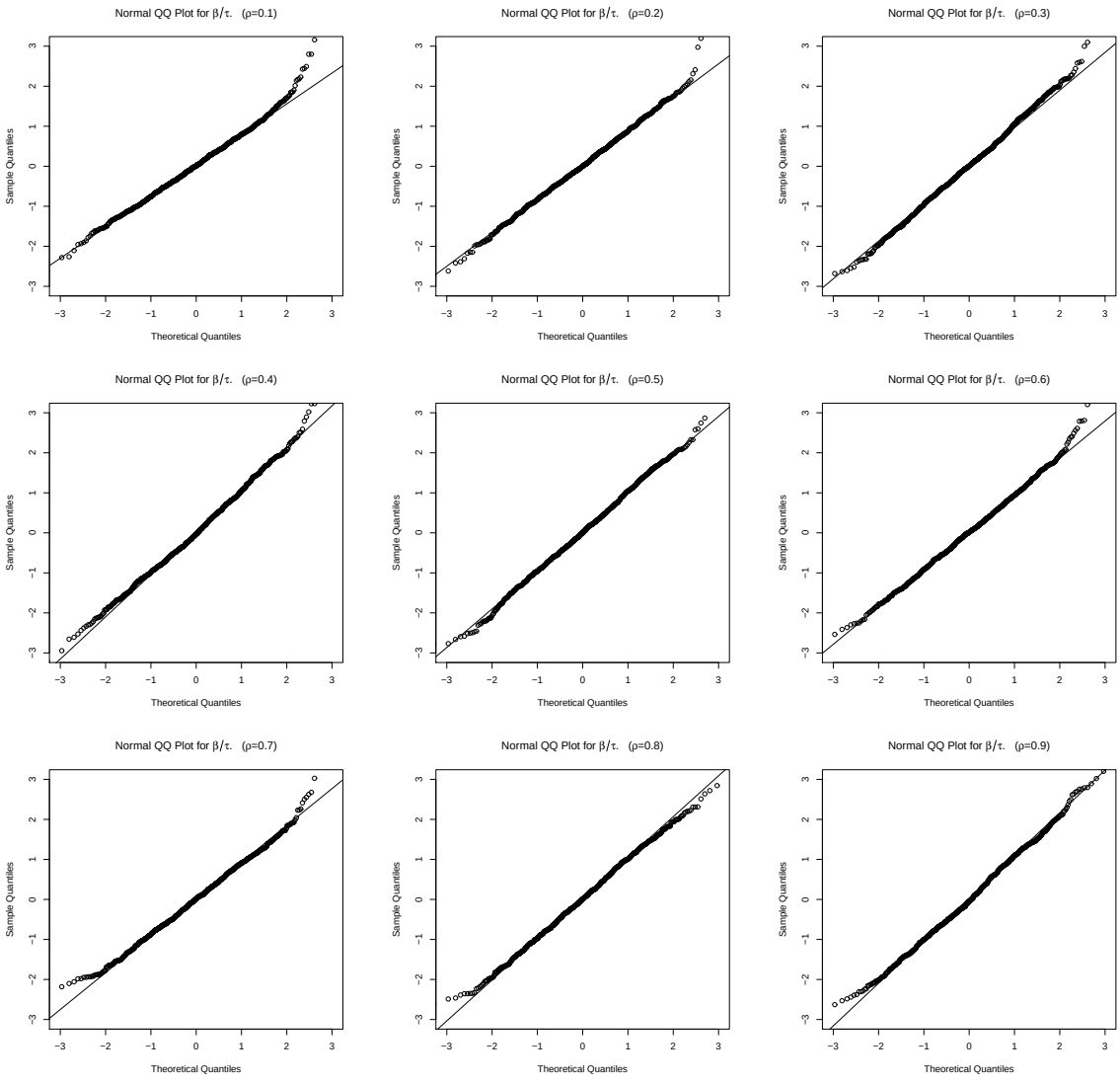

Figure 2.1: Normal QQ Plot for β/τ with $\rho = 0.1, \ldots, 0.9$.

2.6 Discussion

In this chapter, we propose a new estimator HDCE for measuring the correlation between candidate variables and the response (or current residual), and a simple and efficient variable screening method, DTCCS, which is developed based on the new correlation measurement. The proposed method is justified theoretically and numerically for high dimensional variable screening/selection. Comparing with the seminal screening method SIS, the DTCCS method does not require the marginal correlation assumption and can successfully screen covariates with high spurious correlation. Comparing with iterative screening methods such as Tilting, the proposed DTCCS method can provide more accurate screening results and is less computationally expensive. Comparing with the most recent remarkable HOLP approach, the DTCCS method works much better when the multicollinearity can not be identified and removed from the data. Extensive simulation studies show that the performance of the DTCCS method is competitive and reliable. The DTCCS method enjoys nice properties of successful variable screening and computational efficiency; it is especially appealing for handling data which are highly correlated or multicollinearity exists.

A natural extension of the DTCCS method is to make use of the residual vector from other methods other than the ridge when regressing $\mathbf{X}_j$ against all other variables $\tilde{\mathbf{X}}_{-j}$, such as LASSO, SCAD and etc., for identifying accurate relationships among them. Exploring the DTCCS method under the deterministic design and the post-selection inference are other valuable directions. A thorough numerical study is expected to be explored after the breakthrough of the computing ability.

Chapter 3

Analysis Challenges for High Dimensional Influence Measure

The main objective of this chapter is to develop new methodologies in high dimensional influence measure and discuss the connection to other widely used influence measure methods for both low and high dimensional data. For high-dimensional data, classical methods designed for the low dimensional case either perform poorly or are no longer applicable. In general, the test statistic for the influence measure is a function of a distance between the parameter estimator of the complete dataset and the leave-one-out dataset. This distance captures the effect of individual observations relative to a specific positive definite matrix. Cook (1977) propose the classical Cook's distance which measures each individual observation's influence by using the difference of least squares regression coefficient estimate $\hat{\boldsymbol{\beta}}$ of the full dataset and $\hat{\boldsymbol{\beta}}_{(-i)}$ of the ith deleted dataset relative to the positive definite matrix $\mathbf{M} = \mathbf{X}^T\mathbf{X}$. Note that Cook's distance is proportional to $\|\mathbf{M}^{1/2}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})\|_2^2$ and $\mathbf{M}$ is proportional to the precision matrix (the inverse of the covariance matrix) of $\hat{\boldsymbol{\beta}}$. For $i = 1, \ldots, n$, the real limiting distribution of the Cook's distance is complicated, but a scaling of this statistics follows approximately central F-distribution (Cook 1977). Due to the computational cost associated with large number of covariates as well as the problem of statistical inference accuracy and algorithm stability, traditional influence detection methods such as the Cook's distance do not work well on high-dimensional datasets. Also,

the high dimensional nature of dataset is likely to amplify the potential observation's impact on the analysis. Zhao et al. (2013) suggested to use the marginal correlation between the response and the predictor variables $\hat{\boldsymbol{\rho}}$ as a high-dimensional counterpart of least squares regression coefficient estimate $\hat{\boldsymbol{\beta}}$. *High-dimensional influence measure* (HIM, Zhao et al. 2013) measures the Euclidean norm (squared distance) between the marginal correlation estimate based on all n observations, $\hat{\boldsymbol{\rho}}$, and the estimate obtained by deleting the ith point, say $\hat{\boldsymbol{\rho}}^{(i)}$. Many other test statistics on influence measure are based on an estimator of $\|\mathbf{M}^{1/2}(T_n - T_{n-1})\|_2^2$ for a given positive definite matrix $\mathbf{M}$ and a given estimator T of the full dataset and the deleted dataset respectively, see detail in Cook and Sanford (1980). We shall call these test statistics 'sum-of-squares type statistics' as they aim to estimate the squared Euclidean norm $\|\mathbf{M}^{1/2}(T_n - T_{n-1})\|_2^2$. Although sum-of-squares type statistics are widely used in hypothesis tests, many conditions which are required in hypothesis tests are no longer met in the high dimensional sparse setting. Cai et al. (2014) pointed out that the test based on the sum-of-squares type statistics are not powerful to distinguish between the null and the alternative hypothesis, and proposed test statistics of extreme value distribution (EVD) type. In this chapter, to measure high-dimensional influence, we first propose an EVD type statistic which is based on a linear transformation of $(T_n - T_{n-1})$ by the precision matrix Ω of T. Suppose for the moment that the precision matrix $\Omega = \Sigma^{-1}$ is known. This new statistic is theoretically powerful against sparse alternatives in the high dimensional setting under dependence. However, in most cases Ω is unknown and thus needs to be estimated. When Ω is known to be sparse, the *constrained l_1-minimization for inverse matrix estimation* (CLIME, Cai et al. 2014) can estimate Ω directly but it is time-consuming. To get another efficient method for high-dimensional influence measure, we propose a testing statistic from the perspective of the robustness of design. Similar to the kernel idea in machine learning, a transformation of the influence function (IF) of marginal correlation is used to calculate the inner product on the new high-dimensional sphere. This inner product is measuring the *Hellinger distance* (HD) of two discrete probability mass functions which are transformed from the marginal correlations between the respec-

tive variables or quantities of interest. This construction gives detecting power to flag the observations that have unusual effect on high-dimensional models. The second method will be illustrated theoretically and numerically. Note that the design matrix is assumed to be full row rank deterministic throughout this chapter and some notations will be redefined in this chapter.

3.1 Introduction

In classical regression, an observation is influential if the estimates of $\hat{\boldsymbol{\beta}}$ change substantially when this observation is omitted. The presence of influential observations could lead to distorted analysis and misleading interpretations in statistical analysis. In formulating linear models $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, measuring the changes in the individual estimates is the classical approach of looking at the effect of the observed value on the model. In the work of Cook (Cook 1977, 1979), the influence of the ith observed value is measured by using squared distance between the estimated regression coefficient of the dataset with and without the ith observed value relative to a specific geometry (that is, the plane spanned by the explanatory variables, such as $\mathbf{X}^T\mathbf{X}$). Intuitively, if an observed value has influence on the model, this distance is expected to be large. The derivation of Cook's Distance is based on the above idea.

Without loss of generality (WLOG), we delete the first row of the design matrix $\mathbf{X}$. Write $\mathbf{X} = \begin{pmatrix} x_1^T \\ \mathbf{X}_{(-1)} \end{pmatrix}$, $\mathbf{X}_{(-1)}^T\mathbf{X}_{(-1)} = \mathbf{X}^T\mathbf{X} - x_1x_1^T$ and similarly $\mathbf{Y} = \begin{pmatrix} y_1 \\ \mathbf{Y}_{(-1)} \end{pmatrix}$. Let $a = (\mathbf{X}^T\mathbf{X})^{-1}x_1$, $b = x_1$. By applying Sherman-Morrison-Woodbury formula $(\mathbf{I} - ab^T)^{-1} = \mathbf{I} + \frac{ab^T}{1-b^Ta}$ (Sherman and Morrison 1950), we obtain

$$\begin{aligned}
(\mathbf{X}_{(-1)}^T\mathbf{X}_{(-1)})^{-1} &= [\mathbf{X}^T\mathbf{X}(\mathbf{I} - (\mathbf{X}^T\mathbf{X})^{-1}x_1x_1^T)]^{-1} \\
&= (\mathbf{I} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}x_1x_1^T}{1 - x_1^T(\mathbf{X}^T\mathbf{X})^{-1}x_1})(\mathbf{X}^T\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}x_1x_1^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - h_{11}}.
\end{aligned}$$

Similarly $\mathbf{X}_{(-1)}^T\mathbf{Y}_{(-1)} = \mathbf{X}^T\mathbf{Y} - x_1y_1$, so that the regression coefficients computed

from the reduced sample are:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(-1)} &= [\mathbf{X}_{(-1)}^T\mathbf{X}_{(-1)}]^{-1}\mathbf{X}_{(-1)}^T\mathbf{Y}_{(-1)} \\ &= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}x_1}{1-h_{11}}e_1,\end{aligned}$$

where $e_1 = y_1 - \mathbf{x}_1^T\hat{\boldsymbol{\beta}}$; and in general

$$\hat{\boldsymbol{\beta}}_{(-i)} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}^T\mathbf{X})^{-1}x_i}{1-h_{ii}}e_i,$$

where $e_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$.

Then Cook's distance is

$$\begin{aligned}D_i &= \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})^T\mathbf{X}^T\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)})}{p\hat{\sigma}^2} \\ &= (\frac{e_i}{1-h_{ii}})^2\,\frac{x_i^T(\mathbf{X}^T\mathbf{X})^{-1}x_i}{p\hat{\sigma}^2} \\ &= \frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})}\,\frac{h_{ii}}{p(1-h_{ii})} \\ &= \hat{r}_i^2\frac{h_{ii}}{p(1-h_{ii})}, \qquad (3.1)\end{aligned}$$

where $\hat{r}_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$, $i = 1, \ldots, n$. Note that this statistic does not follow the $F(p, n-p)$ distribution. However, a F-distributed statistic with degrees of freedom p and $(n-p)$ can be approximately employed to conduct the hypothesis testing. The magnitude of D_i is assessed by comparing it with $F_\alpha(p, n-p)$ where α is the significance level. A large value of D_i, for instance, $D_i \geq F_{0.5}(p, n-p)$, indicates that deleting the *ith* observation would move $\hat{\boldsymbol{\beta}}_{(-i)}$ to the boundary of an approximate 50% or more confidence region for $\boldsymbol{\beta}$ based on the complete dataset (Cook 1977). The Cook's Distance is not effective when it is used in high dimensional dataset. In view of this, Zhao et al. (2013) proposed a method to address the above problem, which is termed *high-dimensional influence measure* (HIM). In the work of Zhao et al. (2013), a proposition of a novel high-dimensional influence measure for regressions with the number of explanatory variables far exceeding the number of observations is made. In their work, they considered the distance between the estimated marginal correlation of the response and individual predictors of the original dataset, say $\hat{\boldsymbol{\rho}}$, and that of the response and the individual predictors of the dataset with the single observed value

deleted, say $\hat{\boldsymbol{\rho}}^{(k)}$ where (k) corresponds to the kth observed value deleted. Inference is then conducted on this measure by deriving its asymptotic distribution which is shown to follow a chi-Squared distribution. The resulting conclusion based on this inference is then used to determine the influence an observation has on the model under construction. Even though HIM performs appreciably well, there are some drawbacks:

1. The performance of the method depends on the robustness of the estimate of mean and variance.

2. Since standardization is not employed in each leave-one-out step, the estimates of the marginal correlations are not bounded between -1 and 1.

3. The intractability involved in the analysis of high dimensional datasets that renders the Euclidean norm is not preferable in applications involving high dimensional data mining.

4. High dimensional datasets mostly induce high correlation among predictors. These correlations are mostly 'spurious', hence marginal correlations based on this statistical phenomenon may not yield reliable result, and therefore it require some adjustment.

To overcome the limitations faced by HIM and the deficiency of the test based on the sum-of-squares type statistics, we discuss the influence diagnosis measure from the perspectives of test statistics of extreme-value-distribution (EVD) type and of robustness of design type.

3.2 High-dimensional Influence Measure Based on EVD Statistics

In traditional linear regression, ordinary least squares (OLS) projects the response $\mathbf{Y}$ onto the linear space $col(\mathbf{X})$ spanned by the column vectors of $\mathbf{X}$. Shao and Deng (2012) used an approach to project the regression vector $\boldsymbol{\beta}$ onto the row space

$row(\mathbf{X})$ and showed that high-dimensional estimate is identifiable if and only if it lies in a set having a one-to-one correspondence with $row(\mathbf{X})$. One advantage to use $row(\mathbf{X})$ in high-dimensional screening is that the dimension of $row(\mathbf{X})$ is at most n for $p > n$. For $p > n$, ridge regression estimator and HOLP (Wang and Leng 2016) are in $row(\mathbf{X})$.

In model (1.2), $\mathbf{X}$ is considered to be a full rank deterministic design matrix whose dimension p is larger than n, hence $\mathbf{X}$ has full row rank n and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$. From the singular value decomposition,

$$\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{Q}^T, \tag{3.2}$$

where $\mathbf{P}$ is an $n \times n$ matrix satisfying $\mathbf{P}^T\mathbf{P} = \mathbf{I}_n$, $\mathbf{Q}$ is a $p \times n$ matrix satisfying $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_n$, $\mathbf{D} = diag(D_{11}, \ldots, D_{nn})$ is an $n \times n$ full rank diagonal matrix with $D_{11} \geq D_{22} \geq \cdots \geq D_{nn} > 0$. Let $\mathbf{D}^{-1} = diag(1/D_{11}, \ldots, 1/D_{nn})$.

Recall HOLP by using singular value decomposition,

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{holp} &= \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}Y, \\ &= \mathbf{Q}\mathbf{D}^{-1}\mathbf{P}^T Y. \end{aligned} \tag{3.3}$$

Since we are using deterministic design matrix in this chapter, the covariance matrix of $\hat{\boldsymbol{\beta}}_{holp}$,

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{holp} &= \mathbf{Q}\mathbf{D}^{-1}\mathbf{P}^T \cdot Var(Y) \cdot \mathbf{P}\mathbf{D}^{-1}\mathbf{Q}^T, \\ &= \mathbf{Q}\mathbf{D}^{-2}\mathbf{Q}^T\sigma^2. \end{aligned} \tag{3.4}$$

Combining Eq. (3.3) and (3.4), we obtain $\hat{\boldsymbol{\beta}}_{holp} \sim N(\mathbf{Q}\mathbf{Q}^T\boldsymbol{\beta}, \mathbf{Q}\mathbf{D}^{-2}\mathbf{Q}^T\sigma^2)$. The estimation of the precision matrix $\hat{\Omega} = \hat{\boldsymbol{\Sigma}}_{holp}^{-1}$ can be determined by *CLIME*.

Using Leave-One-Out technique, denote $\hat{\boldsymbol{\beta}}_{holp}^{(i)}$, $i = 1, \ldots, n$, the HOLP solution from the reduced dataset. If there is no influential observation and $n \to \infty$, $\hat{\boldsymbol{\beta}}_{holp}^{(i)}$ is believed to be identical to $\hat{\boldsymbol{\beta}}_{holp}$. We show the distribution of $\hat{\boldsymbol{\beta}}_{holp}^{(i)}$ when $n \to \infty$.

Proposition 3.2.1. *Let* $\mathbf{X}$ *be a full row rank deterministic design matrix* ($n < p, rank(\mathbf{X}) = n$) *and* $\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{Q}^T$ *where* $\mathbf{P}$ *is uniformly distributed on the Stiefel*

manifold $V_p(\mathbb{R}^p)$, $\mathbf{Q}$ *is uniformly distributed on the Stiefel manifold* $V_n(\mathbb{R}^p)$ *and* $\mathbf{D} = diag(D_{11}, \ldots, D_{nn})$ *is an* $n \times n$ *diagonal matrix with* $D_{11} \geq D_{22} \geq \cdots \geq D_{nn} > 0$. *If there is no influential observation and* $n \to \infty$,

$$\hat{\boldsymbol{\beta}}^{(i)}_{holp} \sim N(\mathbf{Q}\mathbf{Q}^T\boldsymbol{\beta}, \mathbf{Q}\mathbf{D}^{-2}\mathbf{Q}^T\sigma^2). \tag{3.5}$$

Proof. Let p_i^T is the *ith* row of $\mathbf{P}$, $(n-1) \times n$ matrix $\mathbf{P}_{(-i)}$ consists the remaining $(n-1)$ rows of $\mathbf{P}$, $\mathbf{P}^T_{(-i)} \in V_{n-1}(\mathbb{R}^n)$. Hence $\mathbf{P}_{(-i)}\mathbf{P}^T_{(-i)} = \mathbf{I}_{n-1}$.

Without loss of generality, we use $i = 1$ in this proof.

$\mathbf{P} = \begin{pmatrix} p_1^T \\ \cdots \\ \mathbf{P}_{(-1)} \end{pmatrix}_{n\times n}$, and its transpose $\mathbf{P}^T = \begin{pmatrix} p_1 & \mathbf{P}^T_{(-1)} \end{pmatrix}_{n\times n}$. Using the same $\mathbf{D}$ and $\mathbf{Q}$ as defined in Eq. (3.2), $\mathbf{X}_{(-1)} = \mathbf{P}_{(-1)}\mathbf{D}\mathbf{Q}^T$, $\hat{\boldsymbol{\beta}}^{(1)}_{holp} = \mathbf{Q}\mathbf{D}\mathbf{P}^T_{(-1)}\mathbf{P}_{(-1)}\mathbf{D}^{-2}\mathbf{P}^T_{(-1)}Y$ and $E(\hat{\boldsymbol{\beta}}^{(1)}_{holp}) = \mathbf{Q}\mathbf{D}\mathbf{P}^T_{(-1)}\mathbf{P}_{(-1)}\mathbf{D}^{-2}\mathbf{P}^T_{(-1)}\mathbf{P}_{(-1)}\mathbf{D}\mathbf{Q}^T\boldsymbol{\beta}$. The proof is left to show $\mathbf{P}^T_{(-1)}\mathbf{P}_{(-1)} \to \mathbf{I}_n$ as $n \to \infty$.

$\mathbf{P}e_1$ is the first column of $\mathbf{P}$ and is uniformly distributed on a unit sphere S^{n-1}. Let $\{w_i, i = 1, 2, \ldots, n\}$ be i.i.d random variable from standard normal distribution, we have $\mathbf{P}e_1 \overset{(d)}{=} \left(\sqrt{\sum_{j=1}^n w_j^2}\right)^{-1/2} (w_1, w_2, \ldots, w_n)^T$ and

$$\mathbf{P}_{(-1)}e_1 \overset{(d)}{=} \left(\sqrt{\sum_{j=1}^n w_j^2}\right)^{-1/2} (w_1, w_2, \ldots, w_{n-1})^T.$$

Hence, the first diagonal element of $\mathbf{P}^T_{(-1)}\mathbf{P}_{(-1)}$, $e_1^T\mathbf{P}^T_{(-1)}\mathbf{P}_{(-1)}e_1 \overset{(d)}{=} \frac{w_1^2+\cdots+w_{n-1}^2}{w_1^2+\cdots+w_n^2}$, is a random variable which follows a beta distribution with parameter $(n-1)/2$ and $1/2$.

From Lemma 3 Moderate deviation of Fan and Lv (2008), we know that for any $C > 0$, there exists some $\epsilon_1, \epsilon_4 \in (0, 1)$, $\epsilon_2, \epsilon_3 > 0$ such that

$$P\left(\frac{1}{n-1}\sum_{i=1}^{n-1} w_i^2 < 1 - \epsilon_1\right) \leq e^{-C(n-1)},\ P\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 > 1 + \epsilon_2\right) \leq e^{-Cn} < e^{-C(n-1)}.$$

and

$$P\left(\frac{1}{n-1}\sum_{i=1}^{n-1} w_i^2 > 1 + \epsilon_3\right) \leq e^{-C(n-1)},\ P\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 < 1 - \epsilon_4\right) \leq e^{-Cn} < e^{-C(n-1)},$$

Let $c_1' = \frac{1-\epsilon_1}{1+\epsilon_2}$, $c_2' = \frac{1+\epsilon_3}{1-\epsilon_4}$ and by Bonferroni's inequalities, we have

$$P\left(e_1^T\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}e_1 < c_1'\frac{n-1}{n} \text{ or } e_1^T\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}e_1 > c_2'\frac{n-1}{n}\right) \leq 4e^{-C(n-1)}.$$

Then for any $C > 0$, there exist c_1'', c_2'' with $0 < c_1'' < 1 < c_2''$, such that

$$P\left(e_1^T\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}e_1 < c_1'' \text{ or } e_1^T\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}e_1 > c_2''\right) \to 0 \text{ as } n \to \infty. \tag{3.6}$$

Since $\mathbf{P}_{(-1)}^T \in V_{n-1}(\mathbb{R}^n)$, $\mathbf{P}_{(-1)}\mathbf{P}_{(-1)}^T = \mathbf{I}_{n-1}$. The trace of $\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}$ is the same as that of $\mathbf{P}_{(-1)}\mathbf{P}_{(-1)}^T$, hence $tr(\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}) = n-1$. Combining with the result of Eq. (3.6), we conclude the expectation of the diagonal elements tend to 1 as $n \to \infty$.

Since the Frobenius norm $\|\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)} - \mathbf{I}_n\|_F = 1$ and $\mathbf{P}$ is uniformly distributed on the Stiefel manifold $V_p(\mathbb{R}^p)$, we obtain the sparseness of the off-diagonal elements. That completes the proof of this proposition. □

To determine the degree of influence the ith observation has on the HOLP estimate, the sum-of-square type statistic $\|\mathbf{M}^{1/2}(\hat{\boldsymbol{\beta}}_{holp} - \hat{\boldsymbol{\beta}}_{holp}^{(i)})\|_2^2$ and the extreme-value-distribution (EVD) type statistic $\|\mathbf{M}(\hat{\boldsymbol{\beta}}_{holp} - \hat{\boldsymbol{\beta}}_{holp}^{(i)})\|_\infty$ are two approaches to measure the influence measure. In this section, we follow the preference of Cai et al. (2014) to discuss the EVD type statistics.

For a given invertible $p \times p$ matrix $\mathbf{M}$, denote the vector $E = \mathbf{M}(\boldsymbol{\beta}_{holp} - \boldsymbol{\beta}_{holp}^{(i)}) = (E_1, E_2, \ldots, E_p)^T$. Let $\mathbf{b} = (b_{11}, \ldots, b_{pp})^T$ be the diagonal of the covariance matrix of $\mathbf{M}\boldsymbol{\beta}_{holp}$. We propose to test the null hypothesis 'H_0 : the ith observation is not influential' on the basis of the test statistic

$$D_{\mathbf{M}} = \max_{1\leq j\leq p} \frac{E_j^2}{b_{jj}}. \tag{3.7}$$

Let $\Omega = \Sigma^{-1}$ be the precision matrix of $\boldsymbol{\beta}_{holp}$ and assume Ω is known. A nature choice of $\mathbf{M}$ is $\Omega^{1/2}$ since the components of $\Omega^{1/2}\hat{\boldsymbol{\beta}}_{holp}$ and $\Omega^{1/2}\hat{\boldsymbol{\beta}}_{holp}^{(i)}$ are i.i.d. random variables following normal distribution, $\Omega^{1/2}\hat{\boldsymbol{\beta}}_{holp} \sim N(\mathbf{Q}\mathbf{Q}^T\boldsymbol{\beta}, I_p)$ and $\Omega^{1/2}\hat{\boldsymbol{\beta}}_{holp}^{(i)} \sim N(\mathbf{Q}\mathbf{D}\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}\mathbf{D}^{-2}\mathbf{P}_{(-1)}^T\mathbf{P}_{(-1)}\mathbf{D}\mathbf{Q}^T\boldsymbol{\beta}, I_p)$. By selecting $\mathbf{M} = \Omega^{1/2}$, the test statistic for the ith influence measure is

$$D_{\Omega^{1/2}} = \max_{1\leq j\leq p} \frac{\{[\Omega^{1/2}(\hat{\boldsymbol{\beta}}_{holp} - \hat{\boldsymbol{\beta}}_{holp}^{(i)})]^{\circ 2}\}_j}{b_{jj}} = \max_{1\leq j\leq p} \frac{E_j^2}{b_{jj}}, \tag{3.8}$$

where $^{\circ 2}$ is the Hadamard square of a vector. Eq. (3.8) is termed the *extreme value distribution for high-dimensional influence measure* (EVD-HIM) for $\mathbf{M} = \Omega^{1/2}$.

Under H_0, E_j, $j = 1, 2 \ldots, p$, follows normal difference distribution with mean 0 and variance b_{jj}. Hence, a scaled $\frac{E_j^2}{b_{jj}}$, $j = 1, 2 \ldots, p$, follows standard chi-squared distribution. To evaluate the test statistics, we start with the required 'concentration condition' for the high-dimensional covariance matrix.

Condition 3.2.1. *$C_0^{-1} \leq d_\star(\Sigma) \leq d^\star(\Sigma) \leq C_0$ for some constant $C_0 > 0$.*

Theorem 3.2.1. *Let the test statistics $D_{\Omega^{1/2}}$ be defined as in Eq. (3.8) and Condition 3.2.1 holds. Let $a_0 = o(\log(p))$ be a prespecified constant which is proportional to the correlation of $\Omega^{1/2}\boldsymbol{\beta}_{holp}$ and $\Omega^{1/2}\boldsymbol{\beta}_{holp}^{(i)}$. Let constant $b_p = 2\log p - \log(\log p) - \log \pi$ $(p = 2, 3, \ldots)$, then $\frac{a_0 D_{\Omega^{1/2}} - b_p}{2}$ has a nondegenerate limit distribution as $p \to \infty$, i.e.,*

$$\lim_{p\to\infty} P(\frac{a_0 D_{\Omega^{1/2}} - b_p}{2} \leq t) = e^{-e^{-t}}. \tag{3.9}$$

Proof. Let χ_i^2, $i = 1, \ldots, p$, be a series of iid standard chi-squared random variables. Since $b_p = 2\log p - \log(\log p) - \log \pi$ $(p = 2, 3, \ldots)$ and $\frac{b_p}{a_0} \to \infty$, $P(\frac{a_o \chi_i^2 - b_p}{2} > t) \sim e^{-t}/p$ as $p \to \infty$, which is

$$\lim_{p\to\infty} p \cdot (1 - F(\frac{b_p + 2t}{a_0})) = e^{-t}. \tag{3.10}$$

We want to find the limiting distribution $L(t) = \lim_{p\to\infty} F^p(\frac{b_p+2t}{a_0})$ which is equivalent to $\log L(t) = \lim_{p\to\infty} p \cdot \log(F(\frac{b_p+2t}{a_0}))$. We know that from De Haan and Ferreira (2007), $\lim_{p\to\infty} \frac{-\log F(\cdot)}{1-F(\cdot)} = 1$. Combine this result with (3.10), $e^{-t} = -\log L(t)$. Hence, $L(t) = e^{-e^{-t}}$. □

Theorem 3.2.1 shows the asymptotic null distribution of $D_{\Omega^{1/2}}$. On the basis of the limiting null distribution, the asymptotic $\alpha-$level test can be defined as follows:

$$\Psi_\alpha(\Omega^{1/2}) = I[D_{\Omega^{1/2}} \geq \frac{b_p + 2q_\alpha}{a_0}], \tag{3.11}$$

where q_α is the $(1-\alpha)$-quantile of the the Gumbel distribution (Standard Extreme Value distribution Type-I) with the cumulative distribution function $G(t) = \exp\{-\exp(-t)\}$, i.e. $q_\alpha = -\log[-\log(1-\alpha)]$.

The null hypothesis H_0 is rejected if and only if $\Psi(\cdot) = 1$ at test level α. The $\alpha-$level test can be defined as $T_\alpha = \{\Psi_\alpha : P_{H_0}(\Psi_\alpha = 1) \leq \alpha\}$. We will show that the EVD-HIM's ability of detecting the influential observation is getting more powerful as we increase the value of p in Section 3.4.1.

3.3 High-dimensional Influence Measure Based on Robustness of Design

In the previous section, we propose the EVD-statistics which is theoretically feasible in the high dimensional setting with known precision Ω. However, Ω is unknown and difficult to estimate in most cases. In this section, we discuss the influence measure from another perspective which does not need to compute the precision matrix. We start with the notion of the sample correlation.

The Sample Analogue of Marginal Correlation

Standardizing the predictors and centering the response are the common procedure to solve high-dimensional problems, see details in Efron et al. (2004), Fan and Lv (2008), Wang (2009), Cho and Fryzlewicz (2012), Zhao et al. (2013) Wang and Leng (2016) and Fan et al. (2018). Zhao et al. (2013) standardized all regressors $\mathbf{X}_j$ (columns of $\mathbf{X} \in \mathbb{R}^{n\times p}$) and $\mathbf{Y}$ ($\in \mathbb{R}^n$) to find the asymptotic properties of HIM. In this section, unit length scaling is used to standardize all regressors and the response as well.

Firstly, let $\mathbf{S} = (s_{ij})$ denotes the sample covariance matrix such that

$$(n-1)\mathbf{S} = \mathbf{X}^T(I_n - J_n)\mathbf{X}, \tag{3.12}$$

where $\mathbf{1}_n = (1, \ldots, 1)^T$ and $J_n = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ is symmetric and idempotent. Hence, $I_n - J_n$ is symmetric and idempotent.

Hence, the design matrix after standardization is

$$\mathbf{K}_{n\times p} = \begin{pmatrix} k_1^T \\ \vdots \\ k_n^T \end{pmatrix} = \begin{pmatrix} & \vdots & \\ \frac{x_{i1}-\bar{x}_1}{\sqrt{s_{11}}} & \ldots & \frac{x_{ip}-\bar{x}_p}{\sqrt{s_{pp}}} \\ & \vdots & \end{pmatrix}, \tag{3.13}$$

where $k_i = (\frac{x_{i1}-\bar{x}_1}{\sqrt{s_{11}}}, \ldots, \frac{x_{ip}-\bar{x}_p}{\sqrt{s_{pp}}})^T$, $i = 1, \ldots, n$. We have the sample correlation matrix as

$$R = \frac{1}{n-1}\mathbf{K}^T\mathbf{K} = \begin{pmatrix} 1 & \hat{\rho}_{12} & \hat{\rho}_{13} & \ldots & \hat{\rho}_{1p} \\ \hat{\rho}_{12} & 1 & \hat{\rho}_{23} & \ldots & \hat{\rho}_{2p} \\ \hat{\rho}_{13} & \hat{\rho}_{23} & 1 & \ldots & \hat{\rho}_{3p} \\ \vdots & \vdots & \vdots & & \vdots \\ \hat{\rho}_{1p} & \hat{\rho}_{2p} & \hat{\rho}_{3p} & \ldots & 1 \end{pmatrix}, \tag{3.14}$$

and R can take the form

$$R = D_S^{-1/2}\mathbf{S}D_S^{-1/2}, \tag{3.15}$$

where $D_S = diag(s_{11}, \ldots, s_{pp})$ is the diagonal matrix of sample variances. From Eq. (3.3), (3.14) and (3.15), we have $\mathbf{K}$ as

$$\mathbf{K} = (I_n - J_n)\mathbf{X}D_S^{-1/2}. \tag{3.16}$$

Denote the sample variance of $\mathbf{Y}$ as SS_y and define

$$SS_y = \frac{1}{n}\mathbf{Y}^T(I_n - J_n)\mathbf{Y}.$$

Similarly $Y_0 = SS_y^{-1/2}(I_n - J_n)\mathbf{Y}$ is the standardized values of $\mathbf{Y}$, so that the linear square estimate for regression coefficients of Eq. (1.2) is

$$\hat{\boldsymbol{\beta}} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^TY_0 = R^{-1}\begin{pmatrix} \hat{\rho}_{1Y} \\ \hat{\rho}_{2Y} \\ \vdots \\ \hat{\rho}_{pY} \end{pmatrix}, \tag{3.17}$$

where $\hat{\rho}_{jY} = \frac{\sum_{u=1}^{n}(x_{uj}-\bar{x}_j)(y_u-\bar{y})}{(S_{jj}SS_Y)^{1/2}} = \frac{n\,S_{jY}}{(S_{jj}SS_Y)^{1/2}}$.

In the presence of multicollinearity, the matrix $\mathbf{K}^T\mathbf{K}$ becomes singular or nearly singular. In this case, $\hat{\boldsymbol{\beta}} = (\mathbf{K}^T\mathbf{K})^{-1}\mathbf{K}^TY_0$ does not exist. However, $\mathbf{K}^TY_0$ is possible to calculate. This implies that the correlation between $\mathbf{X}$ and $\mathbf{Y}$ can be ascertained.

By employing the leave-one-out technique, let $G^{(k)} = [\mathbf{K}_{(-k)}]^TY_{0(-k)}$, where $\mathbf{K}_{(-k)}$ and $Y_{0(-k)}$ are the matrices obtained by deleting the kth row from $\mathbf{K}$ and Y_0.

The Hellinger Distance Setting

Distance or divergence measures are importance in statistics and machine learning. One of the most important deep learning techniques, *convolutional neural network* (CNN, Hinton and Salakhutdinov 2006) uses cross entropy as the criteria to minimize the loss function, where the cross entropy is derived from *Kullback-Leibler divergence*. There are articles regarding the distance or divergence measures in theoretical or applied statistics. Among them, the minimum Hellinger distance (MHD, Beran 1977) is believed to be one of the most popular approach for independent and identically distributed (iid) continuous random variables in parametric or nonparametric models. MHD estimators have been shown to have excellent robust properties in parametric models such as the resistance to outliers and robustness with respect to model misspecification (Beran 1977; Donoho and Liu 1988). Since the original work of Beran, MHD estimators have been developed in the literature for various setups and models including discrete random variables, parametric mixture models, semiparametric models, nonparametric models and etc. Recent developments in this area and some important references can be found in the recent articles of Tang and Karunamuni (2013), Karunamuni et al. (2015).

Let F and G denote two probability measures that are absolutely continuous with respect to a dominating probability measure μ, denote the densities as $f = \frac{dF}{d\mu}$ and $g = \frac{dG}{d\mu}$, respectively, the squared Hellinger distance $D_H(F, G)$ between two probability measures F and G can be expressed as a standard calculus integral

$$D_H^2(F, G) = \frac{1}{2} \int [\sqrt{f} - \sqrt{g}]^2 d\mu, \tag{3.18}$$

and the choice of μ does not affect the value of $D_H(F, G)$ (Shorack 2017). Let F_j and $G_j^{(k)}$, $j = 1, \ldots, p$, be the marginal correlation of jth predictor and the response from the whole and kth deleted dataset respectively. We apply the dot-product kernel idea to measure the distance of two $p \times 1$ absolute correlation vectors $F = \{|F_1|, |F_2|, \ldots, |F_p|\}$ and $G = \{|G_1^{(k)}|, |G_2^{(k)}|, \ldots, |G_p^{(k)}|\}$. Let $\boldsymbol{\theta}$ be the absolute value

of the marginal correlation, we use the following transformation

$$\phi(\boldsymbol{\theta}) = \frac{1}{p}\left\{ \begin{pmatrix} \mathbf{0}_p \\ \cdots \\ \mathbf{1}_p \end{pmatrix}_{2p\times 1} + \begin{pmatrix} \mathbf{I}_p \\ \cdots \\ -\mathbf{I}_p \end{pmatrix}_{2p\times p} \boldsymbol{\theta}_{p\times 1} \right\}, \tag{3.19}$$

where $\boldsymbol{\theta} \in \mathbb{R}^p$, $\mathbf{0}_p$, $\mathbf{1}_p$ are two $p \times 1$ vectors of 0's and 1's respectively. After the transformation, two probability mass functions (pmfs) $\tilde{F}$ and $\tilde{G}^{(k)}$ are constructed from F and $G^{(k)}$,

$$\tilde{F} = \phi(F) = \frac{1}{p}\{|F_1|, |F_2|, \ldots, |F_p|,\, 1-|F_1|, 1-|F_2|, \ldots, 1-|F_p|\}$$

and

$$\tilde{G}^{(k)} = \phi(G) = \frac{1}{p}\{|G_1^{(k)}|, |G_2^{(k)}|, \ldots, |G_p^{(k)}|,\, 1-|G_1^{(k)}|, 1-|G_2^{(k)}|, \ldots, |1-|G_p^{(k)}|\}.$$

Based on the pmfs above, we have the Squared Hellinger distance between $\tilde{F}$ and $\tilde{G}^{(k)}$,

$$\begin{aligned} D_H^2(\tilde{F}, \tilde{G}^{(k)}) &= \left\| \sqrt{\tilde{F}} - \sqrt{\tilde{G}^{(k)}} \right\|_2^2 \\ &= \frac{1}{2p}\sum_{j=1}^{p} \left\{ \left(\sqrt{|F_j|} - \sqrt{|G_j^{(k)}|} \right)^2 + \left(\sqrt{1-|F_j|} - \sqrt{1-|G_j^{(k)}|} \right)^2 \right\} \\ &= 1 - h^k, \end{aligned} \tag{3.20}$$

where

$$h^k = \frac{1}{p}\sum_{j=1}^{p} \left(\sqrt{|F_j G_j^{(k)}|} + \sqrt{(1-|F_j|)(1-|G_j^{(k)}|)} \right). \tag{3.21}$$

Considering F as the baseline distribution since the 'true' distribution of the marginal correlations is unknown. We check the Hellinger distance of the transformed pmfs $\tilde{F}$ and $\tilde{G}^{(k)}$. If the Hellinger distance between $\tilde{F}$ and $\tilde{G}^{(k)}$ is negligible, then the kth observation is not flagged as influential. Otherwise, the kth observation is a reasonable candidate of an influential observation. Therefore, h^k measures the closeness of $\tilde{F}$ and $\tilde{G}^{(k)}$. This implies that large values of h^k indicates that the observation in question may not be influential. Conversely, small values of h^k indicates that the observation in question may have a potential influence on the model. We term h^k as the Hellinger distance for high-dimensional influence measure (HD-HIM).

The Population Analogue of IF based on HD

From the sum-of-squares type test statistic for the influence measure as follows,

$$D_i(M, c, \Upsilon_i) = \frac{\Upsilon_i^T M \Upsilon_i}{c}. \tag{3.22}$$

$D_i(M, c, \Upsilon_i)$ comprises of three components: the influence function Υ of a specific parameter estimator, a matrix M which captures the space span by the explanatory variables and a scalar c. (3.22) is applicable to low dimensional datasets, for instance, the Cook's Distance. The IF is the regression coefficients in this case. However, in high dimensional datasets seting, the regression coefficients can not easily be found or may not be estimable but marginal correlations between the explanatory and the response variables can always be calculated. (3.22) can then be extended to high dimensions by choosing IF to be the marginal correlation. Following the similar spirit, techniques of the correlation learning are widely used as the alternative approach of the traditional regression coefficients, see details in Section 1.3.

Similar to the construction of influence function for regression coefficients, we need an appropriate functional T defined on the joint distribution, F of the $(p+1)$-vector of the x and y with

$$E_F\left\{\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix}(\mathbf{x}^T, y)\right\} = \begin{pmatrix} E_F(\mathbf{x}\mathbf{x}^T) & E_F(\mathbf{x}y)\\ E_F(y\mathbf{x}^T) & E_F(yy)\end{pmatrix}. \tag{3.23}$$

So that by standardizing the variables $\mathbf{x}$ and y we have

$$E_F\left\{\begin{pmatrix}\mathbf{x}\\ y\end{pmatrix}(\mathbf{x}^T, y)\right\} = \begin{pmatrix} \mathbf{I}_p & \Gamma(F)\\ \Gamma^T(F) & 1\end{pmatrix}, \tag{3.24}$$

where $\mathbf{I}_p$ is a p-dimensional identity matrix. Functional T on F is constructed from the feature transformation ϕ in Eq. (3.19) which is based on the marginal correlation between the response y and explanatory variables x,

$$T(F) = \left\{\frac{1}{p}\left[\begin{pmatrix}\mathbf{0}_p\\ \cdots\\ \mathbf{1}_p\end{pmatrix} + \begin{pmatrix}\mathbf{I}_p\\ \cdots\\ -\mathbf{I}_p\end{pmatrix}|\Gamma(F)|\right]\right\}^{\circ\frac{1}{2}}, \tag{3.25}$$

where $(\cdot)^{\circ\frac{1}{2}}$ is the Hadamard positive square root of $(\cdot)$ and $\mathbf{0}_p$ and $\mathbf{1}_p$ are vectors of zeros and ones respectively. Note that $\Gamma(F)$ is a p-dimensional vector of absolute

marginal correlations of $\mathbf{x}$ and y. To apply the functional T to the derivation of the influence function (IF), we define the following:

$$\Upsilon_i = T_n(F) - T_{n-1}(F), \tag{3.26}$$

where subscript n corresponds to the sample size. Note that HD-HIM in Eq. (3.21) coincides with the dot-product kernel of $T(F)$ and $T(G^{(k)})$.

3.3.1 Inference of Proposed Method

Asymptotic Properties

In this section, our interest is to discuss the asymptotic property of the proposed test statistic. We start to present the regularity conditions and definitions.

Condition 3.3.1. *(Polynomial high-dimensional). $p > n$ and $p = O(n^\alpha)$ for some $\alpha > 0$.*

Condition 3.3.2. *(Normality assumption). Assume $\mathbf{X}$ follows the matrix normal distributions and $\epsilon_i \sim N(0, \sigma^2)$ with variance σ^2 for $i = 1, \ldots, n$.*

Definition 3.3.1. *(Blended weight chi-squared disparity (BWCS) and blended weight Hellinger distance (BWHD), (Lindsay, 1994, 2004))*

Lindsay (1994) defined the BWCS and BWHD between two p-dimensional discrete mass function $\mathcal{F}$ and $\mathcal{G}$ respectively as

$$BWCS(\alpha) = \sum_{i=1}^{p} \frac{(\mathcal{F}_i - \mathcal{G}_i)^2}{\alpha\mathcal{F}_i + (1-\alpha)\mathcal{G}_i},\ \alpha \in [0,1], \tag{3.27}$$

and

$$BWHD(\alpha) = \sum_{i=1}^{p} \frac{(\mathcal{F}_i - \mathcal{G}_i)^2}{[\alpha\sqrt{\mathcal{F}_i} + (1-\alpha)\sqrt{\mathcal{G}_i}]^2},\ \alpha \in [0,1]. \tag{3.28}$$

α in Eq. (3.30) and (3.31) adjusts the weight of $\mathcal{F}$ and $\mathcal{G}$. α equals 0 and 1 corresponds to Pearson's chi-square and Neyman's chi-square respectively. For $\alpha = \frac{1}{2}$ in Eq. (3.30), it is symmetric chi-square which is a squared distance satisfying the triangle inequality (Le Cam 1986, Ch. 4). Also, we would need to multiply $BWCS(\alpha)$ by p in order to obtain the usual chi-squared test statistics (Lindsay 2004).

Lemma 3.3.1. *(Theorem 2.7 of Van der Vaart 1998)*

If A_n converges in distribution to A and the difference between A_n and B_n converges in probability to zero, then B_n also converges in distribution to A,

$$|A_n - B_n| \xrightarrow{p} 0, \text{ and } A_n \xrightarrow{d} A \quad \Rightarrow \quad B_n \xrightarrow{d} A. \tag{3.29}$$

Theorem 3.3.1. *Assume that Condition 3.3.1 and Condition 3.3.2 hold and $BWCS(\alpha)$ and $BWHD(\alpha)$ are defined in Definition 3.3.1 with $\mathcal{F}_i = (f_i, 1 - f_i)^T$ and $\mathcal{G}_i = (g_i, 1 - g_i)^T$. Suppose there is no influential point and $p \to \infty$, the asymptotic distribution of $p \cdot BWHD(\frac{1}{2})$ approximately converge to the chi-squared distribution with degree of freedom of* 1.

Proof. Let the elementwise disparities of BWCS and BWHD between $\mathcal{F}$ and $\mathcal{G}$ be

$$BWCS_p(\alpha) = \sum_{i=1}^{p} \frac{(\mathcal{F}_i - \mathcal{G}_i)^2}{\alpha\mathcal{F}_i + (1-\alpha)\mathcal{G}_i}, \ \alpha \in [0, 1], \tag{3.30}$$

and

$$BWHD_p(\alpha) = \sum_{i=1}^{p} \frac{(\mathcal{F}_i - \mathcal{G}_i)^2}{[\alpha\sqrt{\mathcal{F}_i} + (1-\alpha)\sqrt{\mathcal{G}_i}]^2}, \ \alpha \in [0, 1]. \tag{3.31}$$

Note that $BWHD_p(\frac{1}{2})$ and $BWCS_p(\frac{1}{2})$ can be considered as the two sequences of random variables since they are both constructed from the probability mass functions.

From Lemma 3.3.1, the proof is left to show

$$d_p = \|BWHD(\frac{1}{2}) - BWCS(\frac{1}{2})\| \xrightarrow{p} 0. \tag{3.32}$$

$\tilde{F}$ and $\tilde{G}^{(k)}$ are defined in Eq. (3.20).

$$
\begin{aligned}
h^k &= \sum_{j=1}^{p}\left\{\sqrt{\frac{|F_jG_j^{(k)}|}{p^2}}+\sqrt{\frac{(1-|F_j|)(1-|G_j^{(k)}|)}{p^2}}\right\} \\
&= \frac{1}{2}\sum_{j=1}^{p}\{\sqrt{\left(\frac{|F_j|}{p}+\frac{|G_j^{(k)}|}{p}\right)^2-\left(\frac{|F_j|}{p}-\frac{|G_j^{(k)}|}{p}\right)^2} \\
&\quad +\sqrt{\left(\frac{1-|F_j|}{p}+\frac{1-|G_j^{(k)}|}{p}\right)^2-\left(\frac{1-|F_j|}{p}-\frac{1-|G_j^{(k)}|}{p}\right)^2}\} \\
&= \frac{1}{2}\sum_{j=1}^{p}\left(\frac{|F_j|}{p}+\frac{|G_j^{(k)}|}{p}\right)\left\{1-\left(\frac{|F_j|-|G_j^{(k)}|}{|F_j|+|G_j^{(k)}|}\right)^2\right\}^{1/2} \\
&\quad +\frac{1}{2}\sum_{j=1}^{p}\left(\frac{1-|F_j|}{p}+\frac{1-|G_j^{(k)}|}{p}\right)\left\{1-\left(\frac{|F_j|-|G_j^{(k)}|}{2-|F_j|-|G_j^{(k)}|}\right)^2\right\}^{1/2} \\
&= \frac{1}{2}\sum_{j=1}^{p}\left(\frac{|F_j|}{p}+\frac{|G_j^{(k)}|}{p}\right)\left\{1-\frac{1}{2}\left(\frac{|F_j|-|G_j^{(k)}|}{|F_j|+|G_j^{(k)}|}\right)^2-\frac{1}{8}\left(\frac{|F_j|-|G_j^{(k)}|}{|F_j|+|G_j^{(k)}|}\right)^4-\dots\right\} \\
&+ \frac{1}{2}\sum_{j=1}^{p}\left(\frac{1-|F_j|}{p}+\frac{1-|G_j^{(k)}|}{p}\right)\left\{1-\frac{1}{2}\left(\frac{|F_j|-|G_j^{(k)}|}{2-|F_j|-|G_j^{(k)}|}\right)^2-\frac{1}{8}\left(\frac{|F_j|-|G_j^{(k)}|}{2-|F_j|-|G_j^{(k)}|}\right)^4-\dots\right\} \\
&= \frac{1}{2}\sum_{j=1}^{p}\left\{\left(\frac{|F_j|}{p}+\frac{|G_j^{(k)}|}{p}\right)-\frac{1}{4p}\sum_{j=1}^{p}\frac{\left(|F_j|-|G_j^{(k)}|\right)^2}{|F_j|+|G_j^{(k)}|}\right\} \\
&\quad +\frac{1}{2}\sum_{j=1}^{p}\left\{\left(\frac{1-|F_j|}{p}+\frac{1-|G_j^{(k)}|}{p}\right)-\frac{1}{4p}\sum_{j=1}^{p}\frac{(|F_j|-|G_j^{(k)}|)^2}{2-|F_j|-|G_j^{(k)}|}\right\}-\epsilon \\
&= 1-\frac{1}{2p}\sum_{j=1}^{p}\frac{(|F_j|-|G_j^{(k)}|)^2}{(|F_j|+|G_j^{(k)}|)(2-|F_j|-|G_j^{(k)}|)}-\epsilon,
\end{aligned}
$$

where $\epsilon = o(1)$. Since $BWHD(\frac{1}{2}) = 8(1-h^k) \approx \frac{4}{p}\sum_{j=1}^{p}\frac{(|F_j|-|G_j^{(k)}|)^2}{(|F_j|+|G_j^{(k)}|)(2-|F_j|-|G_j^{(k)}|)}$. In this case, $BWCS(\frac{1}{2}) = \frac{4}{p}\sum_{j=1}^{p}\frac{(|F_j|-|G_j^{(k)}|)^2}{(|F_j|+|G_j^{(k)}|)(2-|F_j|-|G_j^{(k)}|)}$. As $p\to\infty$, $d_p = \|BWHD(\frac{1}{2}) - BWCS(\frac{1}{2})\| \xrightarrow{p} 0$. Hence, the asymptotic distribution of $p\cdot BWHD(\frac{1}{2})$ approximately converge to the chi-squared distribution with degree of freedom of 1. □

Lindsay (2004) suggests multiplying $BWCS(\alpha)$ by its dimension in order to obtain the usual chi-squared test statistics. In this case, $p\cdot BWHD(\frac{1}{2})$ is approximately the usual chi-squared test statistics which is the same result of Ch. 17.2 (page 558) of

Le Cam (1986). Hence, $4pBWHD(\frac{1}{2})$ behaves in the same manner as the usual chi-squared test statistics for $p \to \infty$. An important implication of Theorem 3.3.1 is that the test statistic approximately follows the chi-square distribution with degree 1. Thus, $p-$value can be obtained for high dimensional influence diagnosis by given the test statistic. Specifically, for the hypothesis test 'H_0 : the kth observation is not influential versus its alternative', the p-value is $P(\chi^2(1) > 4pBWHD(\frac{1}{2}))$, see detail in the next subsection.

Hypothesis Testing

Since the absolute value of correlation is bounded by 0 and 1, we construct two probability mass functions (pmfs), one for that of the entire sample set and the other for the deleted sample. For simplicity, $\boldsymbol{\theta}$ denotes the absolute value of the marginal correlation and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T \in \mathbb{C}_0$, where $\mathbb{C}_0$ is a p-dimensional real vector.

Let $\phi : \mathbb{C}_0 \to \mathbb{M}_0$ be a function transforming $\mathbb{C}_0$ to a $2p \times 1$ vector of pmfs in $\mathbb{M}_0$. $\mathbb{M}_0$ is given as

$$\mathbb{M}_0 = \{\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{2p})^T : \pi_i \in [0,1],\, i = 1, \ldots, 2p,\, \sum_{i=1}^{2p} \pi_i = 1\}. \tag{3.33}$$

The mapping of ϕ implies that for every element $\boldsymbol{\pi}$ in $\mathbb{M}_0$ there exists an element $\boldsymbol{\theta}$ in $\mathbb{C}_0$ such that $\phi(\boldsymbol{\theta}) = \boldsymbol{\pi}$.

Now, the null hypothesis can be expressed in two ways,

$$H_0 : \boldsymbol{\theta} \in \mathbb{C}_0 \text{ or } H_0 : \boldsymbol{\pi} \in \mathbb{M}_0. \tag{3.34}$$

The hypothesis testing of $\boldsymbol{\theta} \in \mathbb{C}_0$ and $\boldsymbol{\pi} \in \mathbb{M}_0$ are equivalent, see reason in Appendix A5 of Read and Cressie (1988). Based on our proposed method, we can specify function ϕ as follows:

$$\phi(\boldsymbol{\theta}) = \frac{1}{p}\left\{ \begin{pmatrix} \mathbf{0}_p \\ \hdashline \mathbf{1}_p \end{pmatrix} + \begin{pmatrix} \mathbf{I}_p \\ \hdashline -\mathbf{I}_p \end{pmatrix} \boldsymbol{\theta} \right\}. \tag{3.35}$$

Eq. (3.34) implies that to reject H_0 means the difference between the pmfs $\tilde{F}$ and $\tilde{G}^{(k)}$ is significant. Conversely, failing to reject H_0 means the difference between the

pmfs F and $G^{(k)}$ is not significant. In this case, the observation in question is not influential. Let $\boldsymbol{\theta}^*$ be the 'true' test statistics, F is given as $\phi(\boldsymbol{\theta}^*) = \boldsymbol{\pi}^*$ and that of $G^{(k)}$ as $\phi(\boldsymbol{\theta}) = \boldsymbol{\pi}$.

In general, a critical value will be calculated by $4p[\phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}) - \phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}^\star)]^T[\phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}) - \phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}^\star)]$, and a p-value can be obtained by using the asymptotical properties $P(\chi_1^2 > 4p[\phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}) - \phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}^\star)]^T[\phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}) - \phi^{\circ\frac{1}{2}}(\boldsymbol{\theta}^\star)])$. This means, to test a hypothesis as whether an observation is influential or not, we first calculate a test statistic based on this based on this particular observation and then use this statistic as the critical value to obtain the p-value for the test.

3.4 Numerical Studies

To test the proposed methods, we make use of most popular high-dimensional screening method, SIS, to check the selection consistency with or without influential observation. Also, we employed coverage probability (CP) to capture the frequency of the true variable screening rate. The proposed method, HD-HIM, is used to detect the observations that are influential in the simulated dataset.

The *Least Absolute Shrinkage and Selection Operator* (LASSO) for parameter estimation is employed after the detected influential observations are deleted from the dataset. To investigate how good the estimates for the regression coefficients $\boldsymbol{\beta}$ ascertained by the use of LASSO is, we find the respective associated errors between the estimated $\hat{\boldsymbol{\beta}}$ and the true parameter $\boldsymbol{\beta}_{\mathcal{T}}$. In this report, the error is defined as $ERR = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{\mathcal{T}}\|$.

In order to choose the appropriate variables for the model, we conducted a validation process to check whether the result based on LASSO is improved or not. We did this by first assuming a set consisting of the true variables for the model (say A) and then construct a false positive rate (FPR) as well as a true positive rate (TPR) based on the selection of the β's. The following mathematical relation is used to

ascertain the above rates,

$$\begin{cases} FPR = \frac{FP}{FP+TN}, \\ TPR = \frac{TP}{TP+FN}. \end{cases} \tag{3.36}$$

Where FP, TN, TP and FN are false positive, true negative, true positive and false negative respectively and are given as

$$\begin{cases} FP = \sum_{j \notin A} P(\{\hat{\beta}_j \neq 0\}) \\ TN = \sum_{j \notin A} P(\{\hat{\beta}_j = 0\}) \\ TP = \sum_{j \in A} P(\{\hat{\beta}_j \neq 0\} \\ FN = \sum_{j \in A} P(\{\hat{\beta}_j = 0\}) \end{cases} \tag{3.37}$$

Eq. (3.36) determines the probability of falsely excluding variable(s) in A. To account for the proportion of the number of observations that are correctly deleted, say n_{Tp}, we find the ratio/proportion of n_{Tp} to the number of influential observations among the entire number of observations, say n_{inf}. This proportion is termed 'Power of Detection Influence' (Power).

3.4.1 Simulation Study of EVD-HIM

In this subsection, we will give a short simulation example to show the EVD-HIM is good in the high-dimensional influence detection. Due to computational difficulty, the estimation of the precision matrix $\hat{\Omega} = \hat{\mathbf{\Sigma}}_{holp}^{-1}$ can be difficult to obtain even with the recent R package **CLIME** (Cai et al. 2011). In this simulation, we are using a pseudoinverse $\hat{\Omega}_{holp} = \mathbf{Q}\mathbf{D}^2\mathbf{Q}^T\sigma^{-2}$. It is easy to obtain for full row rank $\mathbf{X}$ but $\hat{\Omega}_{holp}\hat{\mathbf{\Sigma}}_{holp} = \mathbf{Q}\mathbf{Q}^T \neq \mathbf{I}_p$ for $\mathbf{Q} \in V_n(\mathbb{R}^p)$. Another computational burden is to calculate $\hat{\Omega}^{1/2}$. For computing efficiency, we use $\mathbf{M} = \hat{\Omega}_{holp}$ in Eq. (3.7). Cai et al. (2014) illustrated D_Ω and $D_{\Omega_{holp}^{1/2}}$ has the same extreme value distribution and the power $\Psi(\Omega)$ uniformly dominates those of $\Psi(\Omega^{1/2})$. These two parts of adjustment will make the proposed EVD-HIM a feasible numerical method under current computing capability.

We use model (1.2) with true $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^T$. In this model, $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are p predictors and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is the noise that is independent of the predictors.

In this simulation, a sample of $(\mathbf{X}_1, \ldots, \mathbf{X}_p)$ with size n was drawn from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with covariance matrix $\Sigma = \rho^{|i-j|}$. Original design matrix and response variables are generated by using $n = 100$ and $p = 500$ or 1000, $\rho = 0.5$. But we manually add influence to the first K observations, i.e., $\mathbf{x}_i^{new} = \kappa * \mathbf{x}_i$, $j = 1, \ldots, K$. Here, we set $\kappa = 20$, $K = 1, 3, 5$. This scenario modifies Example I of Tibshirani (1996) with a fixed $\sigma^2 = 1$. After 100 replication, we report the coverage probability (CP) of true β's, ERR, FPR and the 'Power of the influence detection' (Power).

Table 3.1: Influence Detection of EVD-HIM

(n,p)	K	CP(β1)	CP(β2)	CP(β_5)	ERR	FDR	Power
(100,500)	1	0.99	1.00	0.98	0.66	0.04	0.96
	3	0.99	0.98	0.99	0.83	0.03	0.94
	5	0.99	0.99	0.97	0.97	0.03	0.95
(100,1000)	1	1.00	1.00	0.98	0.70	0.02	0.98
	3	0.99	0.97	0.99	0.80	0.02	0.95
	5	1.00	0.99	0.99	0.93	0.02	0.96

From Table 3.1, we found that the proposed method, EVD-HIM, is applicable in solving the problem of high-dimensional influence measure. The results of $(100, 500)$ and $(100, 1000)$ are both very good, and the 'Power' of the influence detection is slightly better in the case of $(100, 1000)$ than the case of $(100, 500)$. In this simulation, we used an approximate way to obtain the precision matrix and get a promising results. After the numerical development of calculating the precision matrix efficiently, a more complete simulation studies of the EVD-HIM can be easily conducted.

3.4.2 Simulation Study of HD-HIM

For this simulation, we set the sample size $n = 100$, and the number of explanatory variables $p = 1000$. $K\%$ $(K = 1, 2, \ldots, 5)$ of the total observations is set as influential

so that $\tilde{n} = K$. We consider the model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where $\mathbf{X}$ is multivariate normal with $cov(X_{ij}, X_{ij'}) = 0.5^{|j-j'|}$. ϵ follows the multivariate standard normal distribution, and $\beta = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^T$. We simulated $n = 100$ i.i.d. observations from this model. Next, we reset the first $\tilde{n} = K$ data observations as coming from another model,

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\tilde{\beta} + \epsilon.$$

In our study, three cases are considered in the generation of influential points: When

1. the regression coefficients,

2. covariates,

3. both the regression coefficients and covariates

are subjected to different levels of changes. Let κ be the parameter that dictates the magnitude of the influential points such that $\kappa = 0$ implies that influential point(s) is/are not present in the dataset. We used $\kappa = 0$, 0.4, 0.8, 1.2 and 1.6 in the experiment. Let now consider the cases above:

Case 1: The regression coefficients are subjected to changes

For $i = 1, \ldots, \tilde{n}$, and $\tilde{X}_i = X_i$, we have $\tilde{\beta} = (3, 1.5, \kappa, \kappa, 2, \kappa, \ldots, \kappa)^T$. So that, the influential observations are generated according to $\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\tilde{\beta} + \kappa\mathbf{X}\gamma + \epsilon$, where $\gamma = (0, 0, 1, 1, 0, 1, 1, \ldots, 1)^T$. In this case, the responses of the influential observations are contaminated by a random perturbation $\kappa\mathbf{X}\gamma$. Consequently, the corresponding responses admit a different pattern, whereas the predictors of influential observations follow the same distribution as the rest. Table 3.2-3.6 shows the simulation results for case 1. They show how the performance of the proposed method against HIM. Figure 3.1 shows that plot of the powers against the contaminated rate for both HIM and HD-HIM. The graph shows that the HD-HIM performed better than HIM for contamination rate within 5%.

Table 3.2: Simulation results for case 1 with $K = 1$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9850	0.8700	0.7700	0.6850
	CP of β_2	1.0000	0.9650	0.8150	0.6950	0.5750
	CP of β_5	1.0000	0.9150	0.7050	0.4950	0.4200
LASSO	ERR	0.4604	0.9584	1.5883	2.0674	2.3924
	FPR	0.0181	0.0173	0.0158	0.0135	0.0106
HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_5	1.0000	1.0000	0.9950	1.0000	1.0000
HIM+LASSO	ERR	0.4683	0.5868	0.5338	0.5002	0.4946
	FPR	0.0179	0.0187	0.0170	0.0178	0.0172
	POWER	–	0.0000	0.0000	0.0000	0.0000
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	0.9950	1.0000	0.9900	0.9850
	CP of β_5	1.0000	0.9950	1.0000	1.0000	1.0000
HD-HIM+LASSO	ERR	1.0625	1.0035	0.8358	0.7406	0.6791
	FPR	0.0231	0.0225	0.0223	0.0200	0.0187
	POWER	–	0.8050	0.9000	0.9100	0.9400

Table 3.3: Simulation results for case 1 with $K = 2$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9750	0.7950	0.5800	0.5000
	CP of β_2	1.0000	0.9250	0.6850	0.5150	0.4050
	CP of β_5	1.0000	0.7850	0.4550	0.2850	0.2200
LASSO	ERR	0.4604	1.2517	2.1586	2.7593	3.1131
	FPR	0.0181	0.0157	0.0144	0.0112	0.0084
HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	1.0000	1.0000	0.9950	1.0000
	CP of β_5	1.0000	0.9950	0.9900	0.9950	1.0000
HIM+LASSO	ERR	0.4683	0.6813	0.5953	0.5415	0.5204
	FPR	0.0179	0.0170	0.0166	0.0169	0.0174
	POWER	–	0.3425	0.4500	0.4800	0.4900
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	1.0000	0.9950	1.0000	1.0000
	CP of β_5	1.0000	0.9950	1.0000	0.9950	1.0000
HD-HIM+LASSO	ERR	1.0625	0.9597	0.7555	0.6410	0.5811
	FPR	0.0231	0.0208	0.0198	0.0187	0.0175
	POWER	–	0.7900	0.8650	0.8900	0.9050

Table 3.4: Simulation results for case 1 with $K = 3$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9650	0.7050	0.5000	0.3550
	CP of β_2	1.0000	0.9000	0.5500	0.3550	0.2250
	CP of β_5	1.0000	0.7250	0.3100	0.1300	0.0750
LASSO	ERR	0.4604	1.5705	2.6517	3.2189	3.4897
	FPR	0.0181	0.0150	0.0113	0.0076	0.0048
HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_5	1.0000	0.9750	0.9900	0.9950	1.0000
HIM+LASSO	ERR	0.4683	0.7561	0.6463	0.5847	0.5365
	FPR	0.0179	0.0158	0.0163	0.0168	0.0159
	POWER	–	0.4367	0.6033	0.6417	0.6550
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	0.9900	1.0000	1.0000	1.0000
	CP of β_5	1.0000	0.9950	1.0000	0.9950	0.9850
HD-HIM+LASSO	ERR	1.0625	0.9474	0.6869	0.6009	0.5781
	FPR	0.0231	0.0194	0.0184	0.0171	0.0159
	POWER	–	0.7617	0.8533	0.8717	0.8817

Table 3.5: Simulation results for case 1 with $K = 4$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9400	0.6200	0.3600	0.2350
	CP of β_2	1.0000	0.8500	0.4450	0.2550	0.1350
	CP of β_5	1.0000	0.6100	0.2300	0.0800	0.0400
LASSO	ERR	0.4604	1.9348	3.0961	3.6563	3.9806
	FPR	0.0181	0.0187	0.0123	0.0086	0.0065
HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_5	1.0000	0.9850	0.9900	0.9950	1.0000
HIM+LASSO	ERR	0.4683	0.8240	0.6768	0.6289	0.5700
	FPR	0.0179	0.0177	0.0169	0.0185	0.0182
	POWER	–	0.4938	0.6738	0.7150	0.7338
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	0.9900	1.0000	1.0000	0.9950
	CP of β_5	1.0000	0.9900	0.9900	0.9900	0.9750
HD-HIM+LASSO	ERR	1.0625	0.8984	0.6771	0.6454	0.6698
	FPR	0.0231	0.0206	0.0178	0.0186	0.0176
	POWER	–	0.7588	0.8250	0.8375	0.8425

Table 3.6: Simulation results for case 1 with $K = 5$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9200	0.5350	0.2600	0.1650
	CP of β_2	1.0000	0.8100	0.3350	0.1600	0.0700
	CP of β_5	1.0000	0.5050	0.1750	0.0500	0.0300
LASSO	ERR	0.4604	2.1555	3.4358	3.9497	4.2881
	FPR	0.0181	0.0182	0.0136	0.0084	0.0069
HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	0.9950	0.9850	0.9950	1.0000
	CP of β_5	1.0000	0.9750	0.9800	0.9850	1.0000
HIM+LASSO	ERR	0.4683	0.9155	0.7321	0.6712	0.6117
	FPR	0.0179	0.0177	0.0189	0.0185	0.0175
	POWER	–	0.4950	0.7040	0.7520	0.7750
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	0.9950	1.0000	1.0000	0.9950
	CP of β_5	1.0000	0.9950	0.9950	0.9750	0.9550
HD-HIM+LASSO	ERR	1.0625	0.9053	0.7102	0.7044	0.7617
	FPR	0.0231	0.0191	0.0181	0.0164	0.0170
	POWER	–	0.7430	0.7970	0.8130	0.8160

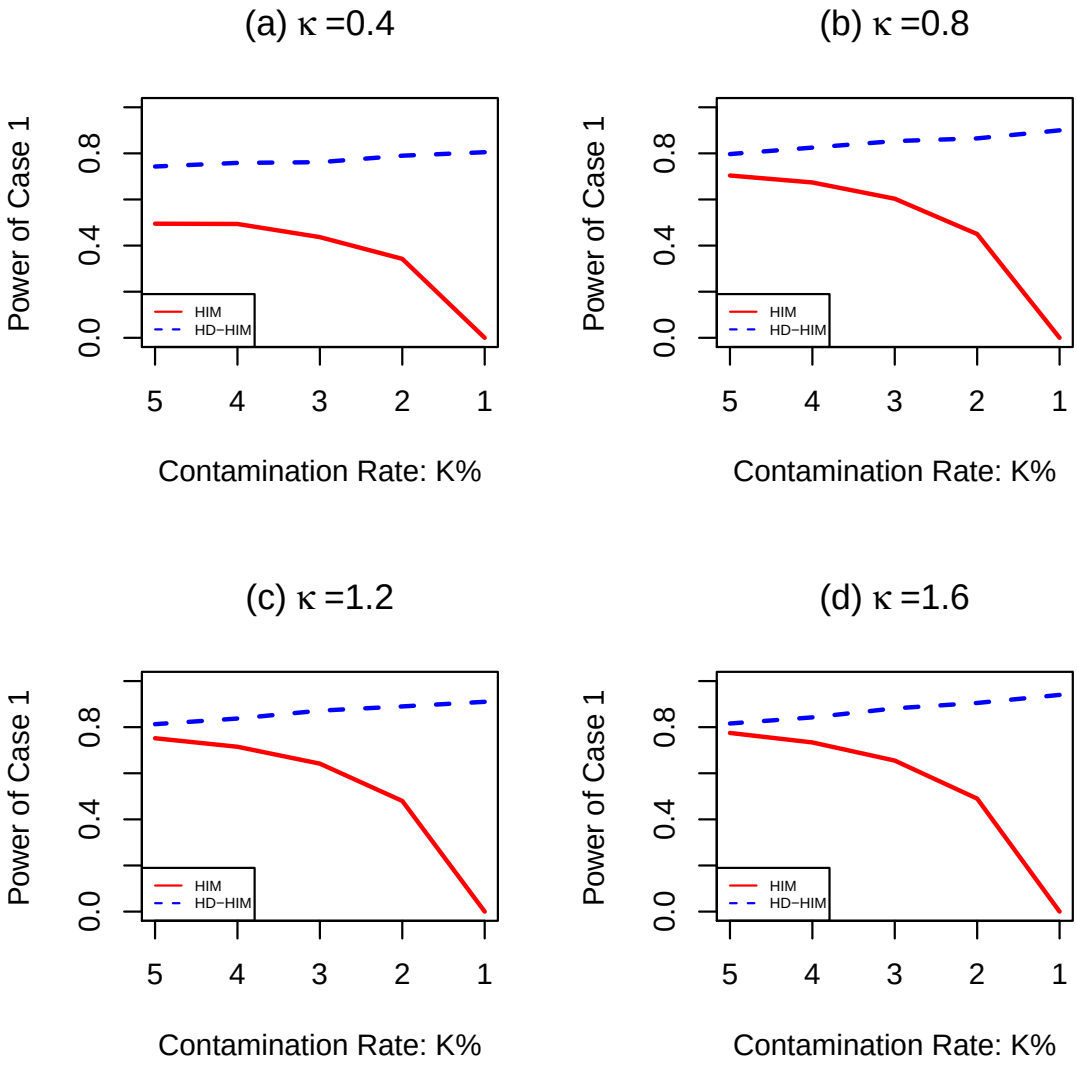

Figure 3.1: Power comparison between HIM and HD-HIM of case 1

Case 2: Covariates are subjected to changes

In this case, we assume that explanatory variables are subjected to changes with the response variable remaining unchange. We set $\tilde{X}_{ij} = X_{ij} + 30\kappa I_{\{i \in S\}}$, while $\tilde{Y}_n = Y_i$, for $i = 1, \ldots, \tilde{n}$, and $j = 1, \ldots, p$. In other words, a set S of explanatory variables admit a different pattern, and its magnitude is controlled by the scalar κ. We examined $S = \{1, \ldots, K\}$, and in this case, the first K observations are considered as the leverage.

Table 3.7-3.11 show the simulation results of case 2. Similarly to the table results for case 1, the HD-HIM is compared to the HIM. Figure 3.2 shows plot of power against the contamination rate. It is shown that the powers ascertained from the HD-HIM perform better than those of HIM for contamination rate within 3%.

Table 3.7: Simulation results for case 2 with $K = 1$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9850	0.9700	0.9750	0.9750
	CP of β_2	1.0000	0.7300	0.7450	0.7500	0.7650
	CP of β_5	1.0000	0.4450	0.4800	0.5150	0.5150
LASSO	ERR	0.4604	3.8289	3.8633	3.7785	3.6731
	FPR	0.0181	0.0005	0.0016	0.0048	0.0091
HIM+SIS	CP of β_1	1.0000	1.0000	0.9850	0.9750	0.9600
	CP of β_2	1.0000	0.9500	0.9550	0.9500	0.8950
	CP of β_5	1.0000	0.9250	0.9300	0.9300	0.8950
HIM+LASSO	ERR	0.4683	1.1723	1.0271	1.0662	1.1925
	FPR	0.0180	0.0147	0.0161	0.0184	0.0194
	POWER	–	0.0000	0.0000	0.0000	0.0000
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_5	1.0000	1.0000	1.0000	1.0000	1.0000
HD-HIM+LASSO	ERR	1.0625	0.6302	0.4938	0.4674	0.4643
	FPR	0.0231	0.0177	0.0181	0.0171	0.0170
	POWER	–	1.0000	1.0000	1.0000	1.0000

Table 3.8: Simulation results for case 2 with $K = 2$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9800	0.9650	0.9500	0.9600
	CP of β_2	1.0000	0.7250	0.7400	0.7650	0.8100
	CP of β_5	1.0000	0.4350	0.5400	0.5250	0.5500
LASSO	ERR	0.4604	1.0650	1.0645	1.0879	1.1351
	FPR	0.0181	0.0911	0.0980	0.1004	0.1034
HIM+SIS	CP of β_1	1.0000	0.9900	0.9500	0.9000	0.8800
	CP of β_2	1.0000	0.8950	0.9150	0.8600	0.8000
	CP of β_5	1.0000	0.8050	0.8600	0.7700	0.7600
HIM+LASSO	ERR	0.4683	1.5783	1.3813	1.5002	1.6536
	FPR	0.0179	0.0182	0.0174	0.0189	0.0200
	POWER	–	0.4675	0.4925	0.4950	0.4950
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	0.9950	0.9950
	CP of β_2	1.0000	0.9300	0.9550	0.9200	0.9200
	CP of β_5	1.0000	0.8650	0.8850	0.8600	0.8800
HD-HIM+LASSO	ERR	1.0625	1.1718	1.2675	1.3671	1.3496
	FPR	0.0231	0.0161	0.0138	0.0137	0.0148
	POWER	–	0.9025	0.8800	0.8650	0.8600

Table 3.9: Simulation results for case 2 with $K = 3$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9550	0.9500	0.9250	0.9350
	CP of β_2	1.0000	0.7000	0.6800	0.6500	0.6500
	CP of β_5	1.0000	0.3450	0.4100	0.3850	0.3850
LASSO	ERR	0.4604	1.0676	1.0760	1.1119	1.1465
	FPR	0.0181	0.0920	0.0984	0.1011	0.1025
HIM+SIS	CP of β_1	1.0000	0.9800	0.9200	0.8550	0.7950
	CP of β_2	1.0000	0.8200	0.8100	0.7550	0.6550
	CP of β_5	1.0000	0.6700	0.7400	0.7200	0.6500
HIM+LASSO	ERR	0.4683	1.7865	1.8861	1.8840	2.0630
	FPR	0.0180	0.0233	0.0168	0.0175	0.0201
	POWER	–	0.6133	0.6617	0.6633	0.6650
HD-HIM+SIS	CP of β_1	1.0000	1.0000	0.9800	0.9800	0.9800
	CP of β_2	1.0000	0.8250	0.8600	0.8900	0.8950
	CP of β_5	1.0000	0.7200	0.7550	0.7750	0.7900
HD-HIM+LASSO	ERR	1.0625	1.6806	1.6200	1.6022	1.4976
	FPR	0.0231	0.0212	0.0241	0.0266	0.0314
	POWER	–	0.8200	0.8017	0.7917	0.7867

Table 3.10: Simulation results for case 2 with $K = 4$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9500	0.9400	0.9300	0.9400
	CP of β_2	1.0000	0.7100	0.6850	0.7150	0.7100
	CP of β_5	1.0000	0.3700	0.4250	0.4300	0.4300
LASSO	ERR	0.4604	1.0668	1.0827	1.1387	1.2033
	FPR	0.0181	0.0936	0.0989	0.1015	0.1028
HIM+SIS	CP of β_1	1.0000	0.9450	0.8700	0.7950	0.7950
	CP of β_2	1.0000	0.7050	0.7200	0.6600	0.6600
	CP of β_5	1.0000	0.5850	0.6600	0.6350	0.6100
HIM+LASSO	ERR	0.4683	2.0711	2.2145	2.2286	2.2653
	FPR	0.0180	0.0258	0.0156	0.0173	0.0201
	POWER	–	0.4938	0.7388	0.7475	0.7488
HD-HIM+SIS	CP of β_1	1.0000	0.9900	0.9800	0.9750	0.9750
	CP of β_2	1.0000	0.8000	0.8150	0.8100	0.8000
	CP of β_5	1.0000	0.5800	0.6450	0.6650	0.6650
HD-HIM+LASSO	ERR	1.0625	1.9895	1.9872	1.9680	1.9634
	FPR	0.0231	0.0249	0.0311	0.0360	0.0378
	POWER	–	0.7800	0.7575	0.7475	0.7475

Table 3.11: Simulation results for case 2 with $K = 5$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9550	0.9500	0.9250	0.9300
	CP of β_2	1.0000	0.6700	0.6450	0.6500	0.6550
	CP of β_5	1.0000	0.3800	0.4150	0.4050	0.4100
LASSO	ERR	0.4604	1.0659	1.0903	1.1519	1.2517
	FPR	0.0181	0.0945	0.1002	0.1024	0.1049
HIM+SIS	CP of β_1	1.0000	0.9100	0.8000	0.7450	0.7500
	CP of β_2	1.0000	0.6500	0.6200	0.5700	0.5850
	CP of β_5	1.0000	0.5300	0.5500	0.4950	0.5000
HIM+LASSO	ERR	0.4683	2.3097	2.3728	2.5147	2.5182
	FPR	0.0180	0.0287	0.0178	0.0191	0.0200
	POWER	–	0.7460	0.7910	0.7960	0.7980
HD-HIM+SIS	CP of β_1	1.0000	0.9800	0.9700	0.9650	0.9600
	CP of β_2	1.0000	0.7450	0.7250	0.7600	0.7300
	CP of β_5	1.0000	0.4950	0.5700	0.5950	0.5850
HD-HIM+LASSO	ERR	1.0625	1.9051	1.8481	1.8325	1.8610
	FPR	0.0231	0.0404	0.0466	0.0528	0.0556
	POWER	–	0.7370	0.7230	0.7170	0.7120

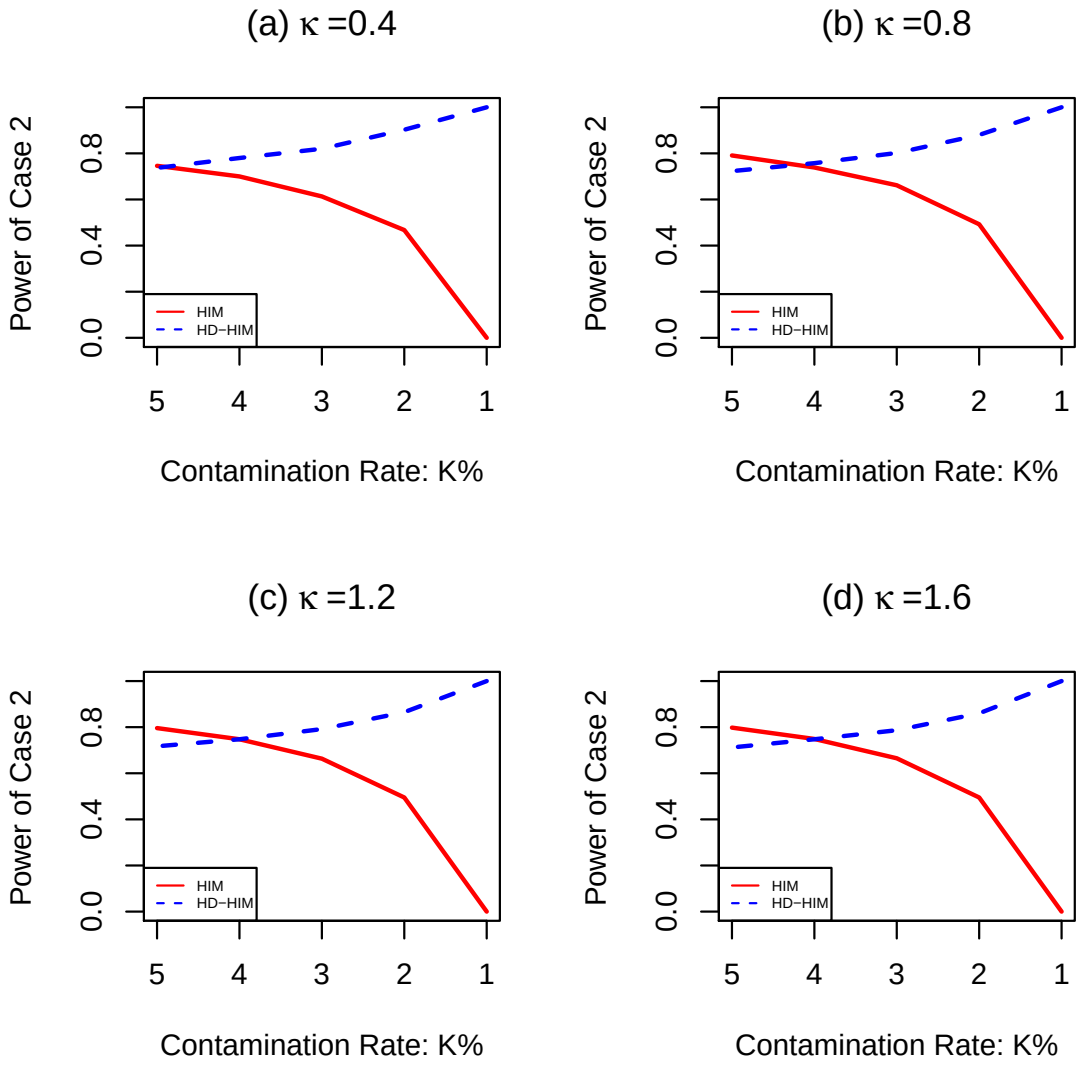

Figure 3.2: Power comparison between HIM and HD-HIM of case 2

Case 3: Both the regression coefficients and covariates are subjected to changes

In this scenario, we set $\hat{\beta} = (3, 1.5, \kappa, \kappa, 2, \kappa, \ldots, \kappa)^T$ and $\tilde{X}_{ij} = X_{ij} + 30\kappa I_{i \in S}$. Similar to case1 and 2, $i = 1, \ldots, p$ and $j = 1, \ldots, p$. We considered $K\%$ mixed leverage points and outliers in this analysis.

Table 3.12: Simulation results for case 3 with $K = 1$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9850	0.8350	0.6500	0.5500
	CP of β_2	1.0000	0.9500	0.7650	0.5750	0.4600
	CP of β_5	1.0000	0.8800	0.5900	0.4200	0.3200
LASSO	ERR	0.4592	1.0117	1.8117	2.4665	2.8606
	FPR	0.0176	0.0174	0.0243	0.0345	0.0413
HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	0.9950	0.9750
	CP of β_2	1.0000	1.0000	1.0000	0.9900	0.9550
	CP of β_5	1.0000	0.9950	1.0000	0.9900	0.9600
HIM+LASSO	ERR	0.4670	0.5786	0.5462	0.5671	0.7822
	FPR	0.0175	0.0174	0.0197	0.0189	0.0204
	POWER	–	0.0000	0.0000	0.0000	0.0000
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	0.9950	1.0000
	CP of β_2	1.0000	0.9950	0.9950	0.9750	0.9900
	CP of β_5	1.0000	0.9950	1.0000	0.9900	0.9900
HD-HIM+LASSO	ERR	1.0620	0.9801	0.8155	0.7124	0.6111
	FPR	0.0232	0.0234	0.0219	0.0202	0.0188
	POWER	–	0.8200	0.9200	0.9400	0.9700

Table 3.13: Simulation results for case 3 with $K = 2$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9650	0.7000	0.4900	0.3750
	CP of β_2	1.0000	0.9200	0.6250	0.3950	0.2900
	CP of β_5	1.0000	0.7450	0.3600	0.2200	0.1650
LASSO	ERR	0.4592	1.3334	2.5328	3.2868	3.7328
	FPR	0.0176	0.0206	0.0397	0.0446	0.0422
HIM+SIS	CP of β_1	1.0000	1.0000	0.9950	0.9800	0.9550
	CP of β_2	1.0000	1.0000	0.9950	0.9800	0.9400
	CP of β_5	1.0000	0.9950	0.9800	0.9650	0.9300
HIM+LASSO	ERR	0.4670	0.6518	0.6182	0.7440	1.0126
	FPR	0.0175	0.0174	0.0205	0.0207	0.0216
	POWER	–	0.3800	0.4725	0.4950	0.5000
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	0.9900	0.9950	0.9900	0.9850
	CP of β_5	1.0000	0.9850	1.0000	0.9850	0.9900
HD-HIM+LASSO	ERR	1.0620	0.9578	0.7393	0.6291	0.5729
	FPR	0.0232	0.0211	0.0208	0.0207	0.0211
	POWER	–	0.8000	0.8900	0.9200	0.9375

Table 3.14: Simulation results for case 3 with $K = 3$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9550	0.6200	0.3550	0.2000
	CP of β_2	1.0000	0.8900	0.4700	0.2400	0.1400
	CP of β_5	1.0000	0.6700	0.2300	0.0600	0.0350
LASSO	ERR	0.4592	1.6528	3.0060	3.7954	4.3189
	FPR	0.0176	0.0196	0.0288	0.0306	0.0288
HIM+SIS	CP of β_1	1.0000	1.0000	0.9900	0.9400	0.8850
	CP of β_2	1.0000	1.0000	0.9900	0.9150	0.8500
	CP of β_5	1.0000	0.9950	0.9750	0.9100	0.8500
HIM+LASSO	ERR	0.4670	0.7353	0.6853	0.8702	1.1877
	FPR	0.0175	0.0179	0.0225	0.0198	0.0202
	POWER	–	0.4783	0.6367	0.6600	0.6667
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	1.0000
	CP of β_2	1.0000	0.9950	1.0000	1.0000	0.9950
	CP of β_5	1.0000	0.9950	0.9950	0.9850	0.9750
HD-HIM+LASSO	ERR	1.0620	0.9306	0.6721	0.5953	0.5939
	FPR	0.0232	0.0210	0.0191	0.0222	0.0234
	POWER	–	0.7867	0.8800	0.8983	0.9067

Table 3.15: Simulation results for case 3 with $K = 4$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9350	0.5050	0.2250	0.1250
	CP of β_2	1.0000	0.8200	0.3750	0.1450	0.0800
	CP of β_5	1.0000	0.5850	0.1750	0.0500	0.0350
LASSO	ERR	0.4592	2.0268	3.4600	4.2060	4.6360
	FPR	0.0176	0.0223	0.0278	0.0238	0.0201
HIM+SIS	CP of β_1	1.0000	1.0000	0.9900	0.9350	0.8500
	CP of β_2	1.0000	0.9950	0.9900	0.9250	0.8250
	CP of β_5	1.0000	0.9850	0.9700	0.9000	0.8050
HIM+LASSO	ERR	0.4670	0.7920	0.7395	0.9784	1.4173
	FPR	0.0175	0.0188	0.0249	0.0222	0.0218
	POWER	–	0.5325	0.7138	0.7388	0.7475
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	0.9950	0.9900
	CP of β_2	1.0000	0.9900	1.0000	1.0000	0.9800
	CP of β_5	1.0000	0.9900	0.9850	0.9600	0.9300
HD-HIM+LASSO	ERR	1.0620	0.8959	0.6921	0.6489	0.6872
	FPR	0.0232	0.0228	0.0218	0.0240	0.0266
	POWER	–	0.7712	0.8462	0.8738	0.8812

Table 3.16: Simulation results for case 3 with $K = 5$

Method	Criterion	κ				
		0	0.4	0.8	1.2	1.6
SIS	CP of β_1	1.0000	0.9050	0.4300	0.1900	0.1150
	CP of β_2	1.0000	0.7900	0.2950	0.1100	0.0750
	CP of β_5	1.0000	0.4850	0.1250	0.0350	0.0200
LASSO	ERR	0.4592	2.2324	3.5894	4.2872	4.6531
	FPR	0.0176	0.0214	0.0199	0.0168	0.0140
HIM+SIS	CP of β_1	1.0000	1.0000	0.9950	0.9300	0.8750
	CP of β_2	1.0000	0.9900	0.9800	0.8950	0.8300
	CP of β_5	1.0000	0.9850	0.9650	0.8700	0.8000
HIM+LASSO	ERR	0.4670	0.8788	0.7817	1.0918	1.4363
	FPR	0.0175	0.0196	0.0239	0.0240	0.0235
	POWER	–	0.5420	0.7460	0.7860	0.7940
HD-HIM+SIS	CP of β_1	1.0000	1.0000	1.0000	1.0000	0.9900
	CP of β_2	1.0000	1.0000	0.9950	0.9850	0.9500
	CP of β_5	1.0000	0.9950	0.9900	0.9450	0.8600
HD-HIM+LASSO	ERR	1.0620	0.8953	0.7243	0.7335	0.8011
	FPR	0.0232	0.0194	0.0233	0.0271	0.0302
	POWER	–	0.7550	0.8250	0.8480	0.8560

Table 3.12-3.16 show the simulation results of case 3. Figure 3.3 shows that the powers of HDHI perform better than those of HIM for contamination rate within 5%.

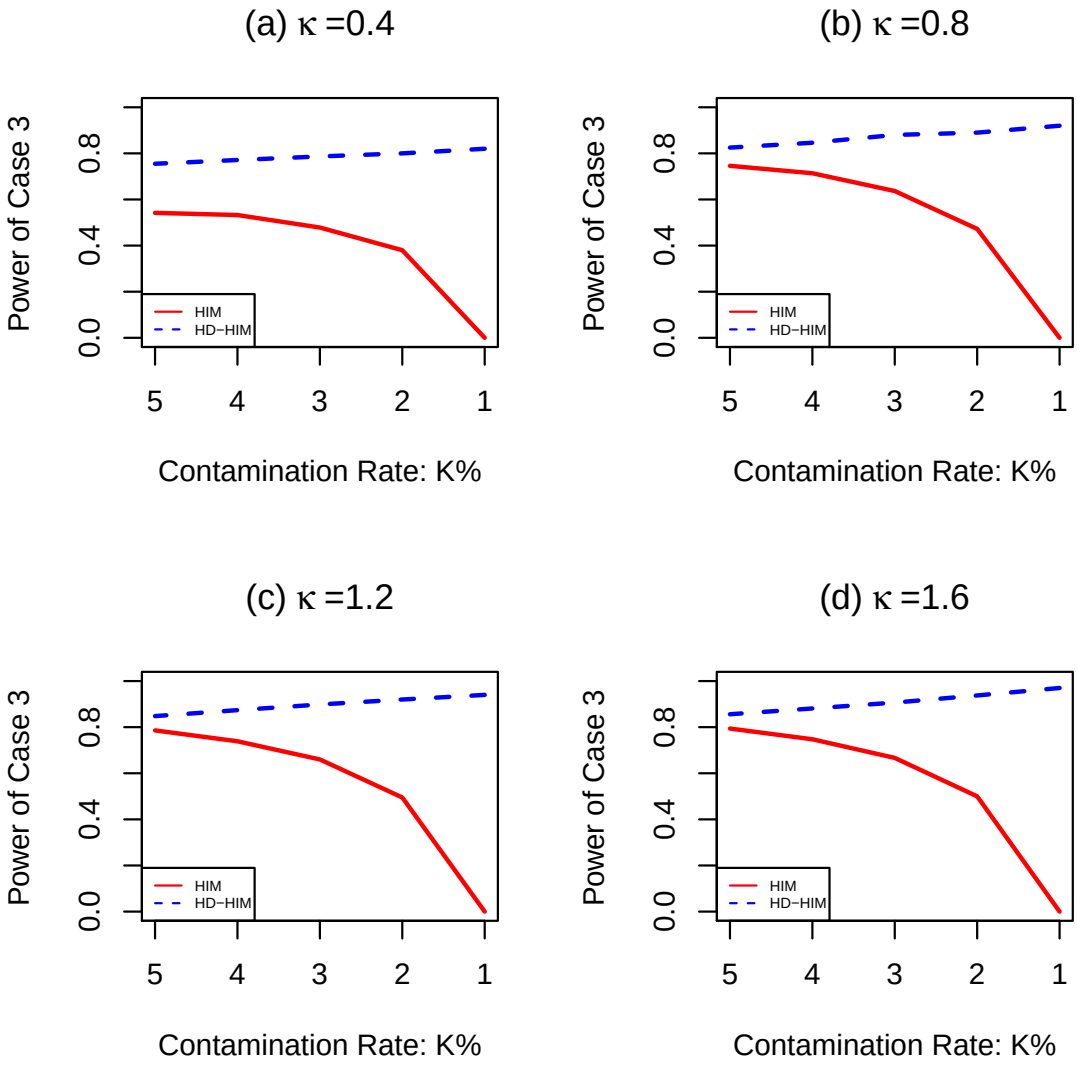

Figure 3.3: Power comparison between HIM and HD-HIM of case 3

The comparison of our method (HD-HIM) to HIM shows that the HD-HIM performed better than the latter in detecting the influential observation for contamination rate from 1% to 5%. Based on the above result, we can conclude that the proposed method, HD-HIM, is applicable in solving the problem of high-dimensional influence measure. After conducting the simulation thoroughly, we can conclude that

1. When the regression coefficients are subjected to perturbation holding the covoriates constant, we noticed that there was not much difference in the perfor-

mance of HIM and HD-HIM.

2. The performances of the HD-HIM and HIM are close when the covariates are subjected to disturbances holding the regression coefficients constant.

3. The performance of HD-HIM is comparable with that of HIM when the perturbation is caused from both the regression coefficient and the covariate predictors. Also, HD-HIM does yield much better performance than HIM with smaller ERR values.

4. From Figure 3.1-3.3, we can conclude that HD-HIM has larger power of detecting the influential observation for smaller contamination rate.

3.5 Conclusion and Discussion

In this chapter, we proposed two methods for detecting influential observation(s) in high-dimensional statistics: one is from the perspective of extreme value distribution; one is from the perspective of the robustness of design.

For the first method, we propose the EVD-type statistics instead of the sum-of-squares type statistics $\|\mathbf{M}^{1/2}(T_n - T_{n-1})\|_2^2$ in high-dimensional statistics, and term it *Extreme Value Distribution for High-dimensional Influence Measure* (EVD-HIM). The EVD type statistics is based on a linear transformation of $(T_n - T_{n-1})$ by the precision matrix Ω of T. Suppose for the moment that the precision matrix $\Omega = \Sigma^{-1}(T)$ is known. This new statistics is theoretically powerful against sparse alternatives in the high dimensional setting under dependence. We use a short simulation to show this promising new method by using a quick precision solving. However, in most cases Ω is unknown and computational expensive to estimate. The possible extension is to work with a better and quicker prediction of the high-dimensional precision matrix. Also, comparing α-level test based on different selection of $\mathbf{M}$ is another potential directions for future research.

For the second method, we construct a data driven method, *Hellinger Distance High-dimensional Influence Measure* (HD-HIM). The HD-HIM is the test statistic

which is expressed as the inner product of the transformed marginal correlations from the whole and deleted dataset. To carry out inference on the proposed method, we established the asymptotic properties of the HD-HIM. The derivation of these properties were based on the fact that $p \cdot BWHD(\frac{1}{2})$ behaves in the same manner as the usual chi-squared test statistics when the dimension of the explanatory variable approaches infinity. Hence, the hypothesis test based on the HD-HIM are compared with the chi-square statistics. Possible extensions of the HD-HIM is to apply the Hellinger-type inner product kernel to the other correlation estimator and the FDR control in the hypothesis testing.

Chapter 4

Cosine Distribution in the Post-Selection Inference of Least Angle Regression

Statistical inference associated with model selection has been discussed for decades. Taking traditional linear regression as an example, we first fit a linear model with all variables included, then preserve the significant ones after drawing the hypothesis testings to all the predictors, and eventually refit the linear model with these significant variables. As the increasing of the predictor's dimension, we usually use a multi-stage procedure and obtain candidate models by a data-driven method. Most data-driven methods have their roots in two ideas: penalized optimization and correlation learning both of which build the path towards a parsimonious model. Post-selection inference about the penalized regression has been discussed recently, for examples, see details in Lockhart et al. (2014) and Lee et al. (2016). In this chapter, we discuss post-selection inference on the correlation learning by using a geometric argument in the LARS solution path.

The rest of this chapter is organized as the follows. The next section introduces a recent development in post-selection inference. Our proposed methodology is discussed in Section 4.3. Numerical studies are reported in Section 4.4. This chapter concludes with a short discussion in Section 4.5. Note that the design matrix is

assumed to be deterministic throughout this chapter.

4.1 Post-selection Inference

Data-driven methods are widely used in high-dimensional statistical problems, but methods from classical statistical inference theory maybe invalid due to the stochastic components in the high-dimensional structure. Under the deterministic design matrix, the response or the current residual bring the stochastic aspects. In this section, we review some recent development on making inference after variable selection by LASSO and forward-type regression. The idea of *post-selection inference* has appeared in literature for decades. Relatively recent ideas on this topic can be found in Berk et al. (2013), Lockhart et al. (2014) and Lee et al. (2016). Berk et al. (2013) produced a valid *post-selection inference* (PoSI) problem by forming statistical tests and confidence intervals of linear models after selecting a subset of the predictors in a data-driven way. Lockhart et al. (2014) and Lee et al. (2016) illustrated post-selection inference of LASSO by forming exact hypothesis testing and confidence intervals respectively.

Recall the linear model (1.2) and the LASSO solution (1.11). The LASSO solution $\hat{\boldsymbol{\beta}}(\lambda)$ is a continuous and piecewise linear function of a sequence of decreasing knots λ's, i.e., $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ where λ_n is the nth knot (tuning parameter) on the LASSO solution path. To test the significance of a predictor that enters the LASSO solution path at a corresponding knot, Lockhart et al. (2014) proposed the covariance test statistic. Let $\mathcal{M}_k = \{j_1, \ldots, j_k\}$ be the LASSO solution set with increasing complexity. The corresponding knots (tuning parameters) after each step are λ_i, $i = 1, \ldots, k$. Note that $\lambda_0 = \infty$ corresponds $\mathcal{M}_0 = \emptyset$. Before the j_kth ($j_k \geq 2$) predictor is added into the model, we have solution set $\mathcal{M}_{k-1}$. Let the estimates at the end of the j_kth step be $\hat{\boldsymbol{\beta}}(\lambda_k)$. If we refit the LASSO by using just the variables in $\mathcal{M}_{k-1}$ with the knot λ_k, the estimates at the end of this step is $\hat{\boldsymbol{\beta}}_{\mathcal{M}_{k-1}}(\lambda_k)$. Then the *covariance test statistic* of the j_kth predictor is defined by

$$\mathbf{T}_{j_k} = \frac{1}{\sigma^2} \cdot \left(\langle \mathbf{Y}, \mathbf{X}\hat{\boldsymbol{\beta}}(\lambda_k) \rangle - \langle \mathbf{Y}, \mathbf{X}_{\mathcal{M}_{k-1}} \hat{\boldsymbol{\beta}}_{\mathcal{M}_{k-1}}(\lambda_k) \rangle \right). \tag{4.1}$$

The statistic $\mathbf{T}_{j_k}$ measures how much contribution $\mathbf{X}_{jk}$ made to improve the fitted model over the interval $(\lambda_{k-1}, \lambda_k)$. At a high probability, large value of $\mathbf{T}_{j_k}$ determines big contribution of variable $\mathbf{X}_{jk}$ in the model $\mathcal{M}_{k-1} \bigcup \{j_k\}$. Under the null hypothesis that all truly active variables are contained in the model $\mathcal{M}_{k-1} \bigcup \{j_k\}$, $\mathbf{T}_{j_k} \xrightarrow{d} \exp(1)$, as $n, p \to \infty$. Post-selection inference does not assume any of the models under consideration to be correct, but evaluate whether a model with a certain predictor surpass the previous model without this predictor.

Lee et al. (2016) discussed a general scheme for post-selection inference which yields exact p-values and confidence intervals in the Gaussian case. Recall the linear model (1.2) with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \Sigma)$. Under the deterministic design matrix setting, $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$. For some matrix M and vector b, a set of linear inequalities in $\mathbf{y}$, i.e., $\{M\mathbf{y} \leq b\}$ can be used as the selection events. Let $\mathcal{M}$ be the current solution set and $\eta = \mathbf{X}_{\mathcal{M}}(\mathbf{X}_{\mathcal{M}}^T \mathbf{X}_{\mathcal{M}})^{-1} e_j$ where e_j is a vector having 1 for the jth element and 0's elsewhere. Inferences about $\eta^T \boldsymbol{\mu}$ conditional on the event $\{M\mathbf{y} \leq b\}$ can been made from a truncated normal distribution. This property gives the possibility of constructing a $1 - \alpha$ level selection interval for $\eta^T \boldsymbol{\mu}$. The confidence bounds of this interval can be solved by inverting the inequalities $\eta^T \boldsymbol{\mu}$ such that $P(\eta^T \boldsymbol{\mu}) \geq 1 - \alpha/2$ and $P(\eta^T \boldsymbol{\mu}) \leq \alpha/2$ respectively.

The path-based regression algorithms are widely used in high-dimensional statistics (Fan and Lv 2010), such as forward-type regression, LASSO, LARS, SIS and DTCCS. Forming a final model under these methodologies, variables are added either one by one such as in LARS and LASSO or one group after another such as in DTCCS. For the LASSO method, the number of non-zero variables in the 'best' final model only depend on a single tuning parameter which means that a sequence of 'knots' of tuning parameters determine different final models. For the DTCCS, the candidate model size is predetermined and a group of monotone value of tuning parameters have been used to form a final model. The nature of high-dimensional statistics may lead to high false discovery rate (FDR) which is the expected fraction of false discoveries among all discoveries (Li and Barber 2017). Let $\mathcal{M}_k = \{j_1, \ldots, j_k\}$ be the active solution set with increasing complexity. To test the model adequacy

and control the FDR, Li and Barber (2017) develop a family of 'accumulation tests' to choose a cutoff $\hat{k}$ to control FDR at level α. In our proposed post-selection inference method, we also include a new stopping criteria which is in the family of 'accumulation tests', see details in Section 4.2.

4.2 Methodology

Recall linear model (1.2) with $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Under the deterministic design matrix setting, $\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$. Assume σ^2 is known in this chapter. For a fixed matrix $\mathbf{X}_{n\times p}$ of predictor variables, we assume that all the covariates have been standardized to have mean 0 and unit length, and the response is also centered. We consider the forward procedure in the LARS context. Note that we only consider the procedure of adding variables, ignore the possibility of deleting variables.

Let $\mathcal{M}_k$ be the active (equicorrelation) set along the LARS solution path, X_{jk} be the j_kth entering predictor, s_{jk} be the sign of the current correlation. Following Efron et al. (2004), we define the matrix

$$X_{\mathcal{M}_k} = (\dots s_{jk}X_{jk} \dots)_{jk\in\mathcal{M}_k}. \tag{4.2}$$

Let $S_{\mathcal{M}_k}$ be the vector containing the signs in the active set with the entering order and $X_S = X_{\mathcal{M}_k}S_{\mathcal{M}_k}$ be the corresponding submatrix formed by extracting the columns of $\mathbf{X}$ in the entering order.

Let

$$G_{\mathcal{M}_k} = X_{\mathcal{M}_k}^T X_{\mathcal{M}_k} \quad \text{and} \quad A_{\mathcal{M}_k} = (1_{\mathcal{M}_k}^T G_{\mathcal{M}_k}^{-1} 1_{\mathcal{M}_k})^{-1/2}, \tag{4.3}$$

where $1_{\mathcal{M}_k}$ being a vector of 1's of length equaling $|\mathcal{M}_k|$, the cardinality of $\mathcal{M}_k$.

The direction of LARS solution path is

$$v_{\mathcal{M}_k} = X_{\mathcal{M}_k}(X_{\mathcal{M}_k}^T X_{\mathcal{M}_k})^{-1}1_{\mathcal{M}_k}, \tag{4.4}$$

then the unit equiangular vector

$$u_{\mathcal{M}_k} = \frac{v_{\mathcal{M}_k}}{\|v_{\mathcal{M}_k}\|}, \tag{4.5}$$

where $\|v_{\mathcal{M}_k}\| = 1/A_{\mathcal{M}_k}$. Hence, $X_{\mathcal{M}_k}^T u_{\mathcal{M}_k} = A_{\mathcal{M}_k} 1_{\mathcal{M}_k}$. The correlation vector between the equiangular direction and all predictors can be calculated by

$$\mathbf{a} = X^T u_{\mathcal{M}_k}, \tag{4.6}$$

then $S_{\mathcal{M}_k}^T X_{\mathcal{M}_k}^T u_{\mathcal{M}_k}$ is a subvector of $\mathbf{a}$ for $|\mathcal{M}_k| < n$. Recall (1.16), at the kth stage, $\hat{C}_k$ is the biggest absolute value of the correlation between the entering variable and the current residual Z_k. LARS finds the variable that has the smallest angle with the current residual and then proceeds in the direction of $u_{\mathcal{M}_k}$ which has the same angle with all X_{jk}'s, $j_k \in \mathcal{M}_k$ for a theoretical step size $\hat{\gamma}$ until the next variable earns its 'most correlated' position. By the end of each stage, LARS updated the mean function, i.e.,

$$\hat{\mu}_{\mathcal{M}_{k+1}} = \hat{\mu}_{\mathcal{M}_k} + \hat{\gamma} u_{\mathcal{M}_k}, \tag{4.7}$$

where

$$\hat{\gamma} = \min_{l \notin \mathcal{M}_k}{}^{+} \left\{ \frac{\hat{C}_k - \hat{c}_l}{A_{\mathcal{M}_k} - a_l}, \frac{\hat{C}_k + \hat{c}_l}{A_{\mathcal{M}_k} + a_l} \right\}, \tag{4.8}$$

where $\hat{c}_l$ is the current correlation of the lth remaining predictor variable and $\min^+$ indicates the smallest positive value such that a new index joins the active set. The mean function $\hat{\boldsymbol{\mu}}$ can be written as

$$\hat{\boldsymbol{\mu}}_{\mathcal{M}_k} = U_{\mathcal{M}_k} \Gamma_{\mathcal{M}_k}, \tag{4.9}$$

where $U_{\mathcal{M}_k} = \left(\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_k \right)$ and $\Gamma_{\mathcal{M}} = (\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_k)^T$. Denote $\hat{\beta}(\hat{C}_k)$ as the regression coefficients of active predictors at stage k, $\hat{\beta}(\hat{C}_k) = (X_S^T X_S)^{-1} X_S^T U_{\mathcal{M}_k} \Gamma_{\mathcal{M}_k}$. The current correlation can also be expressed as the score vector of the least squares criterion with entering predictor:

$$\hat{C}_k = -\frac{s_{jk}}{2} \frac{\partial}{\partial \beta_{jk}} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \Big|_{\beta = \hat{\beta}(\hat{C}_k)}. \tag{4.10}$$

Define the angle $\theta(X_{jk}, Z_k)$ as the angle between the vector X_{jk} and Z_k. Since X_{jk} is standardized, we have

$$\cos\{\theta(X_{jk}, Z_k)\} = \frac{\|X^T Z_k\|_\infty}{\|Z_k\|_2} = \frac{\hat{C}_k}{\|Z_k\|_2}. \tag{4.11}$$

In general, $|\cos\{\theta(X_{jk}, Z_k)\}|$, $k = 1, 2, 3, \ldots$, diminish stochastically. LARS solution path ends at a predetermined step or when the angle $\theta(X_{jk}, Z_k)$ is very close to $\frac{\pi}{2}$, i.e., the remaining variable is almost orthogonal to the current residual.

Lemma 4.2.1. *For $A_{\mathcal{M}_k} \geq 1$, $|\cos\{\theta(X_{jk}, Z_k)\}|$, $k = 1, 2, \ldots, n-1$, is nonincreasing by the LARS solution path. For notational simplicity, we will use θ_k instead of $\theta(X_{jk}, Z_k)$.*

Proof. We know that $\hat{C}_k$ declines with k from Efron et al. (2004), and want to show $1 \geq \frac{\hat{C}_1}{\|Z_1\|_2} \geq \frac{\hat{C}_2}{\|Z_2\|_2} \geq \ldots$ which is equivalent to show $\frac{\hat{C}_k}{\hat{C}_{k+1}} \geq \frac{\|Z_k\|_2}{\|Z_{k+1}\|_2} \geq 1$, for $k = 1, 2, \ldots$.

By Eq. (4.7), $Z_k - Z_{k+1} = \hat{\gamma}_k u_{\mathcal{M}_k}$. Hence, $\hat{\gamma}_k^2 = (Z_k - Z_{k+1})^T (Z_k - Z_{k+1})$, for $k = 1, 2, \ldots$.

From Eq. (1.16), (4.3), (4.6) and (4.7), we obtain

$$\hat{C}_k - \hat{C}_{k+1} = \hat{\gamma}_k A_k \geq \hat{\gamma}_k = \|Z_k - Z_{k+1}\|_2 \geq \|Z_k\|_2 - \|Z_{k+1}\|_2.$$

The last inequality is from the Triangle inequality, then we obtain $\frac{\hat{C}_k}{\hat{C}_{k+1}} \geq \frac{\|Z_k\|_2}{\|Z_{k+1}\|_2}$, that is, $|\cos(\theta_k)| \geq |\cos(\theta_{k+1})|$, for $k = 1, 2, \ldots, n-1$.

Note that in the traditional linear regression model with intercept, $(1/A_{\mathcal{M}_k})^2$ is the first element of the diagonal of hat matrix which contains column one, and it is always bounded by $\frac{1}{n}$ and 1. □

Lemma 4.2.2. *For $Z(\neq 0) \in \mathbb{R}^n$, the following events are equivalent:*

$$\{\|Z_{k+1}\|_2 \cos\theta_{k+1} \leq \|Z_k\|_2 \cos\theta_k \leq \|Z_{k-1}\|_2 \cos\theta_{k-1}\} = \{\theta_{k-1} \leq \theta_k \leq \theta_{k+1}\}$$

Proof. The event in the left hand is equivalent to $\{\hat{C}_{k+1} \leq \hat{C}_k \leq \hat{C}_{k-1}\}$, for $k = 2, 3, \ldots$, which has the monotone property as shown in Efron et al. (2004). The monotonicity of θ's and the one-to-one correspondence of $\hat{C}_k$ and θ_k, $k = 2, 3, \ldots$ have been verified in Lemma 4.2.1. Hence, the above events are equivalent.

□

Recall in the linear regression model (1.2), negligible or zero value of $e_i (= y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}})$ shows a good prediction. In the LARS context, the absolute value of the

corresponding angle at each knot is bounded by $\frac{\pi}{2}$, no more predictor will enter the model once the angle is 'big' enough. We consider the angle close to $\frac{\pi}{2}$ to be 'big' enough.

In this section, we make inference about the angle based on the assumption that the angles follow a (truncated) cosine distribution. The distribution of the cosine value of the angles is shown in Figure 4.1 through a large sample simulation where θ_k are obtained LARS context with $n = 1000$, $p = 2000$ and $\boldsymbol{\beta} = \big(\underbrace{5, \cdots, 5,}_{n \text{ elements}} \quad \underbrace{0, \cdots, 0}_{(p-n) \text{ elements}} \quad \big)^T$. We connect the angle θ_k of each LARS solution path to the incremental null hypothesis which measures whether $\mathcal{M}_k$ statistically surpass $\mathcal{M}_{k-1}$ or not. The limiting distribution of the maximum angle can be used to do an efficient and robust significance test for each predictor variable.

Under the domain of $[-\frac{\pi}{2}, \frac{\pi}{2}]$, Burrows (1986) defined the following cosine distribution with the density function:

$$f(\theta) = \begin{cases} \frac{1}{2}\cos\theta & \text{if } |\theta| \leq \ \pi/2, \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

Its cumulative density function (CDF) is give by:

$$F(\theta) = \begin{cases} 0 & \text{if } \theta < -\pi/2, \\ \sin^2\left(\frac{\theta}{2} + \frac{\pi}{4}\right) & \text{if } |\theta| \leq \pi/2, \\ 1 & \text{if } \theta > \pi/2. \end{cases} \tag{4.13}$$

This CDF, $F(\theta)$, of cosine distribution can be used to do hypotheses testing of whether '$\mathcal{M}_k$ improves over $\mathcal{M}_{k-1}$' by the following theorem.

Theorem 4.2.1. *Assume that the covariate vectors $\mathbf{X}_j$'s, $j = 1, \ldots, p$, are linearly independent in the LARS solution path. Let θ_j, $j = 1, \ldots, n$, be the corresponding angle at each knot $\hat{C}_j$ in the first n steps. If Lemma 4.2.1 and 4.2.2 hold,*

$$\frac{n}{2}\left(\frac{\pi}{2} - \theta_n\right)^2 \xrightarrow{d} \chi^2_2 \text{ as } n \to \infty, \tag{4.14}$$

where χ^2_2 denotes a chi-square random variable with df $= 2$.

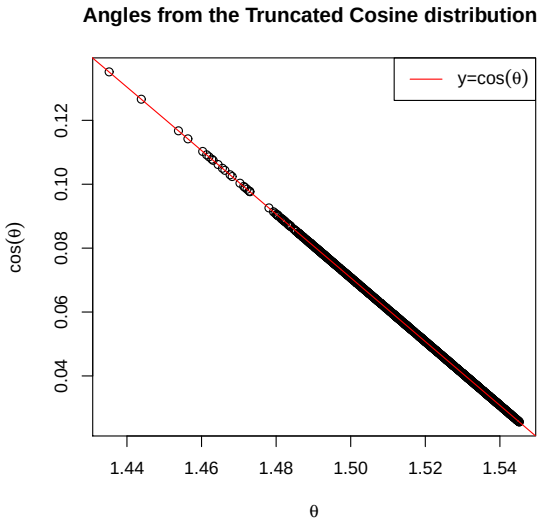

Figure 4.1: Comparing truncated cosine curve (solid) with the simulation results.

Proof. From Lemma 4.2.1, we know that θ_j's, $j = 1, \ldots, n$, are monotone increasing. Hence, θ_1 and θ_n can be considered as the minimum and the maximum order statistics of θ's. As the dimension increases, $\frac{\pi}{2} - \theta_n$ will diminish stochastically.

Let $\tilde{\theta}_n = \frac{n}{2}(\frac{\pi}{2} - \theta_n)^2$. From the CDF of the cosine distribution (Eq. (4.13)) and the basic trigonometric formula, the distribution of $\tilde{\theta}_n$ can be derived as follows:

$$\begin{aligned} P(\tilde{\theta}_n \leq g) &= P\left\{\frac{n}{2}(\frac{\pi}{2} - \theta_n)^2 \leq g\right\} \\ &= P\left\{\theta_n \geq \frac{\pi}{2} - (\frac{2g}{n})^{1/2}\right\} \\ &= 1 - \sin^{2n}\left[\frac{\pi}{2} - (\frac{g}{2n})^{1/2}\right] \\ &= 1 - \cos^{2n}\left[(\frac{g}{2p})^{1/2}\right], \text{ over } 0 \leq g \leq 2n(\pi/2)^2. \end{aligned}$$

Therefore, the limiting distribution of $\tilde{\theta}_n$ is obtained as

$$\begin{aligned} \lim_{n\to\infty} P(\tilde{\theta}_n \leq g) &= 1 - \lim_{n\to\infty} \cos^{2n}[(\frac{g}{2n})^{1/2}] \\ &= 1 - e^{-g/2}, \ g \geq 0, \end{aligned}$$

since $\cos^{2n}[(\frac{g}{2n})^{1/2}] \approx (1 - \frac{g}{4n})^{2n} = e^{-g/2}$ as $n \to \infty$.

Hence, $\tilde{\theta}_n \xrightarrow{d} \chi_2^2$ converge in distribution, where χ_2^2 denotes a chi-square random variable with $df = 2$. □

The limiting distribution of $\tilde{\theta}_n$ determines if the corresponding angle at knot $\hat{C}_k$ is 'big' enough. A sequence of p-values can be obtained by using the above property $P(\chi_2^2 > \tilde{\theta}_j)$, $j = 1, \ldots, n$.

Selection Criteria

Definition 4.2.1. *(Family of 'Accumulation Tests', Li and Barber 2017)*

Let $\mathcal{M}_m$ be the model which includes the first m entries. For an integer $k \in \{1, \ldots, m\}$, a sequence of null hypotheses, H_j, $j = 1, 2, \ldots, k$, measures whether model $\mathcal{M}_j$ statistically surpasses $\mathcal{M}_{j-1}$ or not. Suppose there is a sequence of uniformly distributed $p-$value, $p_1, p_2, \ldots, p_k \in [0, 1]$ corresponding to the hypotheses H_j. Choosing any function $\phi : [0, 1] \mapsto [0, \infty)$ satisfying $\int_{t=0}^{1} \phi(t)dt = 1$, ϕ is termed 'accumulation function'. The 'accumulation tests' determines the stoping point $\hat{k}$ to control FDR at level α and are expressed as

$$\hat{k}_\phi = \max \left\{ k \in \{1, \ldots, m\} : \frac{1}{k} \sum_{j=1}^{k} \phi(p_j) \leq \alpha \right\}. \tag{4.15}$$

We suggest a new $\phi(x) = \frac{x}{\sqrt{1-x^2}}$ to choose a stoping point $\hat{k}_\phi$. We test the hypothesis with H_0 : the jth angle is the maximum one. This null hypothesis is equivalent to test wether the current model is adequate along the LARS solution path. By doing this, we reject all hypotheses up to $\hat{k}_\phi$ and none thereafter.

4.3 Numerical Studies

A few related R packages have been added to the current R community since 2015. The most important packages are **PoSI** (Berk et al. 2013), **covTest** (Lockhart et al. 2014) and **selectiveInference** (Lee et al. 2016). Among them, package **PoSI** and **covTest** cannot support the case of 'small n and large p'. The functions recorded in **selectiveInference** are from Lockhart et al. (2014), Lee et al. (2016), G'Sell et al. (2016)

and etc. We call it *LARS-sI* for the methods form **selectiveInference** in the LARS context. We are going to assess the performance of the proposed cosine post-selection inference (*cosine PoSI*) method by extensive simulation studies and compare the results with that from *LARS-sI*. The proposed ϕ function is used to determine the stop point along the LARS solution path and the testing level is set to be 0.01. Package **selectiveInference** uses the *ForwardStop* (G'Sell et al. 2016) to determine the stop point. *ForwardStop* uses $\phi(x) = \log(\frac{1}{1-x})$ and it is a special case of *accumulation test*. In this simulation, the same testing level has been set for the stopping criteria after the *cosine PoSI* and *LARS-sI*. The selected model size and the selection accuracy are calculated by the expected value after some replications and are defined as $E(|\mathcal{M}_s|)$ and the frequency $P(\mathcal{T} \subset \mathcal{M}_s)$ where $\mathcal{M}_s$ respectively.

In Theorem 4.2.1, the covariate vectors are assumed to be independent, but we still want to see wether the proposed method is robust against the correlated predictors. In general, a strong correlation among the predictors creates difficulty in high dimensional variable screening/selection.

4.3.1 Simulation Studies

To show good performance of the proposed method, we examine two scenarios. In the first scenario, compound symmetry structure of Σ's are used to see wether the proposed method can overcome issues associated with strong correlation among predictors. In the second scenario, auto-regressive correlation structure of Σ's are used to show that the proposed method is good in parsimonious interpretation. 100 replications of simulation are run for each scenario. The results of proposed *cosine PoSI* and *LARS-sI* are reported in Table 4.1-4.2.

Scenario I: Compound Symmetry Structure of Σ

For the first scenario, we use model (1.2) with true $\boldsymbol{\beta} = (5, 5, 5, 0, \ldots, 0)^T$. In this model, $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are p predictors and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ is the noise that is independent of the predictors. In this simulation, a sample of $(X_1, \ldots, X_p)$ with size n was drawn

from a multivariate normal distribution $N(\mathbf{0}, \Sigma)$ with covariance matrix $\Sigma = (1-\rho)I_p + \rho\mathbf{1}\mathbf{1}^T$, where $\mathbf{1} = (1, \ldots, 1)^T$. 16 models are generated by using $n = 100$, or 200, $p = 100$ or 1000, $\rho = 0$, 0.1, 0.5 or 0.9, respectively. This scenario modifies Example I of Fan and Lv (2008) with a fixed $\sigma^2 = 1$.

Table 4.1 shows that the proposed *cosine PoSI* method works perfectly for the case $n = 200$, $p = 100, 1000$ and $\rho = 0$ (independent predictor variables) and works very good for the case $n = 100$, $p = 100, 1000$ and $\rho = 0$. The selected model size increases as the value of ρ increases, but it is still on the level of $O(n)$. The selection accuracy are all 1 for all the cases which means the selected final model always contains the entire set of truly nonzero coefficients. We also found that *LARS-sI* works also very good for the low dimensional case ($n = 200, p = 100$) and it can achieve above 90% selection accuracy for this case. But for the high-dimensional cases, *LARS-sI* works conservatively and only keep the 'strongest' (the first one) variable in the model.

Table 4.1: Selected Model Size and Selection Accuracy for Scenario I

			$p = 100$				$p = 1000$			
n	Method	Result	ρ=0	ρ=0.1	ρ=0.5	ρ=0.9	ρ=0	ρ=0.1	ρ=0.5	ρ=0.9
100	cosine PoSI	$E(\lvert\mathcal{M}_s\rvert)$	3.05	3.17	7.77	11.35	3.23	4.83	16.24	23.90
		$P(\mathcal{T} \subset \mathcal{M}_s)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LARS-sI	$E(\lvert\mathcal{M}_s\rvert)$	1.00	1.04	1.02	1.00	1.02	1.00	1.00	1.00
		$P(\mathcal{T} \subset \mathcal{M}_s)$	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00
200	cosine PoSI	$E(\lvert\mathcal{M}_s\rvert)$	3.00	3.00	5.68	10.02	3.00	3.05	10.98	19.94
		$P(\mathcal{T} \subset \mathcal{M}_s)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LARS-sI	$E(\lvert\mathcal{M}_s\rvert)$	2.94	3.00	3.00	2.96	1.04	1.04	1.00	1.00
		$P(\mathcal{T} \subset \mathcal{M}_s)$	0.93	0.99	0.98	0.98	0.01	0.00	0.00	0.00

Scenario II: Auto-Regressive Correlation

In this scenario, we use model (1.2) with true $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^T$. The predictors $\mathbf{X}_1, \ldots, \mathbf{X}_p$ and the noise $\boldsymbol{\epsilon}$ are again generated the same as in the first scenario, but having different covariance matrix for the predictors. The covariance matrix Σ has entries $\sigma_{ii} = 1$, $i = 1, \ldots, p$ and $\sigma_{ij} = \rho^{|i-j|}$, $i \neq j$. This example is modified from Example 1 of Tibshirani (1996) with ρ set at 0, 0.5, 0.7 or 0.9.

The results are reported in Table 4.2. The proposed post-selection method is always able to select a parsimonious model with accuracy rate of 100% even when the data are highly correlated. From Table 4.1, the proposed *cosine PoSI* method works perfectly for the independent predictor variables and works very good for the case of correlated predictors. The selected model size increases as the value of ρ increases, but it is still on the level of $O(\log(n))$. The selection accuracy of *cosine PoSI* are all 1 for all the cases. We conclude that, for the model with auto-regressive correlation, the *cosine PoSI* method accords the parsimony philosophy in statistics and contains the entire set of truly nonzero coefficients. We also found that *LARS-sI* works conservatively than the proposed method. *LARS-sI* works fine for the low dimensional case ($n = 200, p = 100$) with modest selection accuracy. But for the high-dimensional cases, *LARS-sI* works conservatively and only keeps about one variable in the model.

Table 4.2: Selected Model Size and Selection Accuracy for Scenario II

			$p = 100$				$p = 1000$			
n	Method	Result	ρ=0	ρ=0.5	ρ=0.7	ρ=0.9	ρ=0	ρ=0.5	ρ=0.7	ρ=0.9
100	cosine PoSI	$E(\lvert\mathcal{M}_s\rvert)$	3.03	3.15	3.51	4.24	3.23	3.33	3.79	4.71
		$P(\mathcal{T} \subset \mathcal{M}_s)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LARS-sI	$E(\lvert\mathcal{M}_s\rvert)$	1.12	1.04	1.00	1.00	1.02	1.00	1.00	1.00
		$P(\mathcal{T} \subset \mathcal{M}_s)$	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
200	cosine PoSI	$E(\lvert\mathcal{M}_s\rvert)$	3.00	3.04	3.39	4.10	3.00	3.04	3.39	4.01
		$P(\mathcal{T} \subset \mathcal{M}_s)$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	LARS-sI	$E(\lvert\mathcal{M}_s\rvert)$	2.74	2.73	2.77	2.37	1.00	1.00	1.00	1.00
		$P(\mathcal{T} \subset \mathcal{M}_s)$	0.76	0.80	0.82	0.46	0.00	0.00	0.00	0.00

4.3.2 A Real Data Application

We are going to use the data reported in Scheetz et al. (2006) to show the usefulness of our proposed post-selection inference method. In this data set, $F1$ rates were intercrossed and eye tissues from 120 twelve-week-old male $F2$ offspring were used for microarray analysis. The microarray data used to analyze the RNA from the eye tissues contain over 31042 different genes. Among the genes, one gene with the label '$TRIM32$' was recently found to cause Bardet-Biedl syndrome and it is believed to be linked with a small number of other genes. A subset of this microarray data can be found in the R package **flare**, it contains following two parts:

(1). $\mathbf{X}$ - an 120×200 matrix, which is the data of 120 rats with 200 gene probes.

(2). $\mathbf{Y}$ - a vector with length 120, which is the expression level of gene '$TRIM32$'.

To compare the results from *cosine PoSI* and *LARS-sI* at the same FDR significant level, Leave one out (LOO) technique has been considered such that each observation in the sample is used once as the validation data. We obtain the variables after post-selection inference procedure on the training set and then obtain the OLS estimator of those variables via a linear regression. To evaluate the prediction accuracy, square error $(Y_i - \hat{Y}_i)^2$, $i = 1, \ldots, n$, is recorded for each validation observation. In Table 4.3, we report the means and the standard deviation (SD) of the square errors for prediction and the mean and median of model sizes from n training sets. It can be seen from Table 4.3 that models selected by the proposed *cosine PoSI* has smaller cross-validation error than that from *LARS-sI*, which justifies that the proposed *cosine PoSI* method keeps the useful variables in the post-selection inference procedure, while *LARS sI* is too conservative and most likely screens out many relevant variables. We found that *LARS sI* only contains the very first entered variable for this real data analysis.

We continue to apply the *cosine PoSI* method, in contrast to the *LARS-sI* approaches, to obtain a final model from the full data by first applying the post-selection inference methods to select relevant variables and then obtaining the final model using a linear fit. Table 4.4 reports the selected final model size, The mean square

Table 4.3: Data Analysis of Eye Microarray Data (LOOCV)

Method	Mean square errors	SD square errors	Model size (mean)	Model size (median)
cosine PoSI	0.3548	0.3523	27.6750	28.0000
LARS-sI	15.5772	1.5019	1.0000	1.0000

error (MSE) and adjusted R^2 using different approaches. The MSE and adjusted R^2 obtained after applying an OLS estimator to the final selected variables. We see that the proposed *cosine PoSI* method contains a larger model than the *LARS-sI* procedure. The final model of *cosine PoSI* method keeps 29 variables (ID in package **flare**): $\{153, 55, 99, 87, 42, 85, 180, 177, 109, 90, 199, 112, 36, 185, 62, 136, 200, 155, 187, 146, 188, 134, 141, 172, 127, 11, 54, 181, 164\}$. Comparing to *cosine PoSI*, the *LARS-sI* procedure is too conservative and only includes the first variable into the final model and some relevant variables may lost in this procedure. In this example, we showed the usefulness of the proposed cosine post-selection method which is able to select a final model with size at level $O(n)$ and we also verified our LARS code generates the same solution path as that of the function 'lar' from package **selectiveInference**.

Table 4.4: Final Models for Eye Microarray Full Data using Different Methods

Method	Model Size	MSE	$adjusted\, R^2$
cosine PoSI	29	0.0041	0.8009
LARS-sI	1	0.0087	0.5776

4.4 Discussion

When using a traditional linear regression model, a fixed hypothesis test is conducted to observe which variables are significant at significance level α and report a $(1-\alpha)$ confidence intervals for the significant variables. The randomness aspect in the high-dimensional context brought confliction between model selection and the inference. In

high-dimensional statistics, the data-driven selection procedure is critical important and the model should be selected to be adaptive to the data instead of devising a model before collecting data. Hence, a sequence of random hypothesis tests is required today to do post-selection inference (also termed selective inference). In this chapter, we proposed *cosine PoSI* which is a novel post-selection inference method based on the cosine distribution. We discuss the geometric aspect in LARS and apply the *cosine PoSI* in the LARS solution path to do inference. Comparing with the method in R package **selectiveInference**, the proposed *cosine PoSI* did a better job for the combination of 'small n and large p as measured through a comparison of the methods from **selectiveInference**. The proposed *cosine PoSI* method is strong in providing a parsimony model for independent predictor variables and is robust when the data has 'multicollinearity'.

Lee et al. (2016)'s 'Polyhedral selection'draw inferences about $\eta^T\boldsymbol{\mu}$ conditional on the event $\{M\mathbf{y} \leq b\}$ from a truncated normal distribution. $\eta^T\mathbf{y}$ denotes the parameter estimator constrained to a variable in $\mathcal{M}$ and $\eta^T\mathbf{y} \sim N(\eta^T\boldsymbol{\mu}, \eta^T\Sigma\eta)$. Let $\gamma = \Sigma\eta(\eta^T\Sigma\eta)^{-1}$, $\mathbf{d} = \mathbf{y} - \gamma\eta^T\mathbf{y}$, $\mathcal{V}^-(\mathbf{d}) = \max_{j:(M\gamma)_j<0} \frac{b_j-(M\mathbf{d})_j}{(M\gamma)_j}$, $\mathcal{V}^+(\mathbf{d}) = \min_{j:(M\gamma)_j>0} \frac{b_j-(M\mathbf{d})_j}{(M\gamma)_j}$, $\mathcal{V}^0(\mathbf{d}) = \min_{j:(M\gamma)_j=0} \{b_j - (M\mathbf{d})_j\}$ and $\mathcal{V}^-$, $\mathcal{V}^+$, $\mathcal{V}^0$ are independent of $\eta^T\mathbf{y}$, $\{M\mathbf{y} \leq b\}$ can be rewritten in term of $\eta^T\mathbf{y}$ and $\mathbf{d}$ as follows: $e_j^T(\mathbf{X}_\mathcal{M}^T\mathbf{X}_\mathcal{M})^{-1}\mathbf{X}_\mathcal{M}^T\mu = \eta^T\mu$ for some η.

$$\{M\mathbf{y} \leq b\} = \{\mathcal{V}^-(\mathbf{d}) \leq \eta^T\mathbf{Y} \leq \mathcal{V}^+(\mathbf{d}), \mathcal{V}^0 \geq 0\}. \tag{4.16}$$

Hence, $\eta^T\mathbf{y}|\{M\mathbf{y} \leq b, \mathbf{d} = \mathbf{d}_0\}$ is a truncated normal between $\mathcal{V}^-(\mathbf{d}_0)$ and $\mathcal{V}^-(\mathbf{d}_0)$ where $\mathbf{d}_0$ is a fix value of $\mathbf{d}$ and its CDF follows about a standard uniform distribution.

Inspired by the 'Polyhedral selection', another cosine distribution can also be constructed to approximate normal distribution. The density function and cumulative density function are given by:

$$f(\theta) = \begin{cases} \frac{1}{2\pi}(1 + \cos\theta) & \text{if } |\theta| \leq \ \pi, \\ 0 & \text{otherwise.} \end{cases} \tag{4.17}$$

$$F(\theta) = \begin{cases} 0 & \text{if } \theta < -\pi, \\ \dfrac{1}{2\pi}(\pi + \theta + \sin\theta) & \text{if } |\theta| \leq \pi, \\ 1 & \text{if } \theta > \pi. \end{cases} \tag{4.18}$$

We conjecture that some statistics based on this cosine distribution are able to measure how much improvement the kth entering predictor variable X_{jk} made over the interval $(\hat{C}_{k-1}, \hat{C}_{k+1})$. Then the predictor variables having negligible contribution on this interval can be screened out. We may also combine the results with other post-selection inference method to refine the candidate predictors and make LARS as a powerful and reliable high-dimensional screener.

Chapter 5

Concluding Remarks and Future Work

Analysing high-dimensional data is one of the most challenging problems in the era of big data and artificial intelligence. In this thesis, three important problems are explored for high-dimensional data analysis: variable screening, influence measure and post-selection inference. This chapter summarizes the main contributions made in this thesis and discusses some potential directions for future research.

5.1 Conclusions and Discussions

In Chapter 2, the high-dimensional variable screening problem in linear regression was considered under the assumption of a sparse structure. We proposed a new estimator of measuring the correlation between the predictor variables and the current residual dynamically. The new estimator adaptively reduce the spurious correlation among the predictor variables. Based on this estimator, a new variable screening method termed *Dynamic Tilted Current Correlation Screening* (DTCCS) has been proposed to ensure the screening accuracy especially when the data encounter low signal-to-noise ratio and/or multicollinearity. We theoretically and numerically showed that the DTCCS procedure effectively preserves the relevant variables and reduces the entering chance of the irrelevant variables under certain conditions. A parsimonious

final model can be obtained by combining the DTCCS and the recent development of the model selection criteria for high-dimensional data. In Section 2.5, we extended the DTCCS procedure to the case of the deterministic design matrix. Different from the random design matrix, the value of the 'tilting parameter' can be theoretically determined.

In Chapter 3, we proposed two frameworks to deal with high-dimensional influence measure problem, one is from the extreme value distribution (EVD), another is derived from the robustness of design. The sum-of-squares type statistics have been widely used in traditional statistics for decades, but EVD-type statistics have been proven to be more powerful than the sum-of-squares type statistics in the high dimensional sparse setting (Cai et al. 2011). To measure high-dimensional influence, we first proposed an EVD-type statistic which is based on a linear transformation of $(T_n - T_{n-1})$ by the precision matrix Ω of T. This new statistic is theoretically powerful against sparse alternatives in the high dimensional setting under dependence. However, in most cases the precision matrix is unknown and numerically difficult to estimate. It is a theoretically feasible but time-consuming method. The EVD-type statistic is involved in the future work of obtaining alternative or efficient approach of the precision matrix. From the perspective of robustness of design, we proposed another numerically efficient method termed *Hellinger distance for high-dimensional influence measure* (HD-HIM). We first construct two discrete probability mass functions (PMF) from the marginal correlations between the predictor variables or quantities of interest. Similar to the kernel idea in machine learning, an inner product of two transformed influence function is used to measure the Hellinger distance of those two PMFs. This construction gives detecting power to flag the observations that have unusual effect on high-dimensional models. The HD-HIM method has been thoroughly illustrated theoretically and numerically.

In Chapter 4, we proposed a new numerically feasible post-selection inference method termed *Cosine PoSI* in high-dimensional framework. This method is motivated by the seminal theory of *Least Angle Regression* (LARS, Efron et al. 2004) and it focus on the geometric aspect of LARS solutions. LARS efficiently provides a

solution path along which the entered predictors always have the same absolute correlation with the current residual. At each step of the LARS algorithm, the proposed *Cosine PoSI* method employs an angle from the correlation between the entering variable and current residual and considers this angle as a random variable from the cosine distribution. The post-selection inference is then conducted based on the order statistics of this cosine distribution. Given the collection of the possible angles, we propose a new ϕ function to perform multiple hypothesis tests on the limiting distribution of the maximum angle. Base on our knowledge, there is only one R package **selectiveInference** for post-selection inference in the high-dimensional context. By comparing with **selectiveInference**, we illustrated that the proposed *Cosine PoSI* method can do efficient and robust significant tests for the first n predictor variables on the LARS solution path. The usefulness and the effectiveness of the proposed *Cosine PoSI* method is also established via real-life data analysis.

5.2 Future Work

This thesis is centering on the correlation learning in high-dimensional statistics. Topics on the correlation learning merit further statistical and machine learning research. Part of the theories and the methodologies in this thesis can be generalized to broader family of models, such as the generalized linear model and generalized additive model. There are a bunch of direct applications in the areas such as unusual credit card transaction, abnormal medical screening image and feature engineering. As the rapid development of computational power, the advantage of correlation learning will be more and more prominent. In this section, we will briefly discuss some future work which is expanded from our methodologies on the correlation learning.

Implicit DTCCS

A nature extension of the DTCCS procedure is to use the residual vector from the explicit ridge to other implicit methods when regressing $\mathbf{X}_j$ against all other variables $\tilde{\mathbf{X}}_{-j}$, such as LASSO, SCAD, etc., for identifying accurate relationships among them.

For instance, let $J_\lambda(|\beta_k|) = \lambda|\beta_k| \cdot \mathbb{I}(|\beta_k| \leq \lambda) + \frac{a\lambda|\beta_k| - (\beta_k^2+\lambda^2)/2}{(a-1)\lambda}\mathbb{I}(\lambda < |\beta_k| \leq a\lambda) + \frac{(a+1)\lambda^2}{2}\mathbb{I}(|\beta_k| > a\lambda)$ for $\lambda > 0$ and some $a > 2$, the SCAD-generated residual vector is

$$z_j = X_j - \tilde{\mathbf{X}}_{-j}\hat{\gamma}_j, \text{ and } \hat{\gamma}_j = \arg\min_{\beta} \left\{ \frac{X_j - \tilde{\mathbf{X}}_{-j}\beta}{2n} + \sum_{k=1}^{p} J_\lambda(|\beta_k|) \right\}, \tag{5.1}$$

where $\hat{\gamma}_j$ be the vector of coefficients from the SCAD regression of X_j on $\tilde{\mathbf{X}}_{-j}$. In the case of the random design matrix, the value of 'tilting parameter' λ_j can be preassigned by a descending sequence of positive integers. For the deterministic design matrix, the selection of the 'tuning parameter' may be determined as the deterministic case which was discussed in Chapter 2.

Extension of the Influence Measure

The accuracy of the EVD-type statistics, $\|\Omega(T)(T_n - T_{n-1})\|_\infty$ is highly associated with the estimates of the precision matrix $\Omega(T)$. When we choose T as the HOLP estimator, the preliminary numerical example shows the usefulness of this new diagnostic idea. As the computing power increasing rapidly, we expect to apply an efficient way to calculate the precision matrix which is one of the most important step in the EVD-type statistics, for instance, Wu et al. (2018) suggest a 'low rank + diagonal' decomposition to obtain the high-dimensional inverse. Besides, another potential extension is to figure out an efficient way to obtain the precision matrix of the *high-dimensional correlation estimator* (HDCE). HDCE is one of the most powerful methods to tilt the spurious correlation. We expect its EVD-type statistics is still powerful to spot the influence observation in high-dimensional statistics.

The proposed HD-HIM is efficient and robust to flag influential observations. It can be considered as a new inner-product kernel and applied to many machine learning problems, such as classification. We will continue to explore the asymptotical relationship between the original distance and the its counterpart for a transformed higher or lower dimension. This exploration may relate to the Johnson-Lindenstrauss Lemma (Johnson and Lindenstrauss 1984). This lemma is stated as follows.

Lemma 5.2.1. *(Johnson-Lindenstrauss Lemma, Johnson and Lindenstrauss 1984)*

Suppose we have n points $u_1, \ldots, u_n \in \mathbb{R}^d$. *Given* $\epsilon \in (0,1)$, *we are going to map* $f : \mathbb{R}^d \to \mathbb{R}^k$, *where* $k < d$ *and* $k = O(\epsilon^{-2}\log(n))$, *such that for each* $i, j, 1 \leq i \leq j \leq n$, *we have*

$$(1-\epsilon)\|u_i - u_j\|_2^2 \leq \|f(u_i) - f(u_j)\|_2^2 \leq (1+\epsilon)\|u_i - u_j\|_2^2. \tag{5.2}$$

The above formula can be rearranged to

$$(1+\epsilon)^{-1}\|f(u_i) - f(u_j)\|_2^2 \leq \|u_i - u_j\|_2^2 \leq (1-\epsilon)^{-1}\|f(u_i) - f(u_j)\|_2^2. \tag{5.3}$$

PoSI of DTCCS

The *post-selection inference* (PoSI) allows us to test hypothesis suggested by the data. The PoSI of LASSO and LARS has been discussed in the past few years. In the high or ultra-high dimensional context, some criteria, such as emphextended BIC (Chen and Chen 2008) emphquadratically supported risks (QSR, Kim and Jeon 2016), are used to select a final sparse model one by one after the screening procedure. Our proposed DTCCS admits candidate variables group by group. Post-selection inference in DTCCS with groups of variables is an interesting future work. Under the deterministic design matrix, confidence intervals for the DTCCS can also be examined since $\frac{\hat{\beta}_j - \beta_j}{\tau_j \sigma} \overset{(d)}{\approx} N(0,1)$ where τ_j was defined in Section 2.5.

Bibliography

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, New York.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *Annals of Statistics*, 5:445–463.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41:802–837.

Biba, M. and Xhafa, F. (2011). *Learning Structure and Schemas from Documents*. Springer, Berlin.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*, 37:1705–1732.

Burrows, P. M. (1986). Extreme statistics from the sine distribution. *The American Statistician*, 40:216–217.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l-1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106:594–607.

Cai, T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:349–372.

Cai, T. T., Guo, Z., et al. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of statistics*, 45:615–646.

Cai, T. T., Ren, Z., Zhou, H. H., et al. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10:1–59.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351.

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

Cho, H. and Fryzlewicz, P. (2012). High dimensional variable selection via tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74:593–622.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19:15–18.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74:169–174.

Cook, R. D. and Sanford, W. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics*, 22:495–508.

De Haan, L. and Ferreira, A. (2007). *Extreme Value Theory: An Introduction*. Springer Science & Business Media, New York.

Donoho, D. L. and Liu, R. C. (1988). The 'automatic' robustness of minimum distance functionals. *Annals of Statistics*, 16:552–586.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Fan, J., Feng, Y., Saldana, D. F., Samworth, R., and Wu, Y. (2018). R package 'sis'.

Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106:544–557.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148.

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with np-dimensionality. *Annals of Statistics*, 38:3567–3604.

Fang, J. and Grzymala-Busse, J. (2006). Leukemia prediction from gene expression data–a rough set approach. *Artificial Intelligence and Soft Computing–ICAISC 2006*, pages 899–908.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537.

Graybill, F. A. (1983). *Matrices with Applications in Statistics, Second Edition.* The Wadsworth Statlstics/Probabllitv Series, Belmont.

G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:423–444.

Hampel, F. (1968). *Contributions to the theory of robust estimation, unpublished Ph.D. thesis, University of California, Berkeley.*

Hampel, F. (1973). Robust estimation: A condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 27:87–104.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer, New York.

Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313:504–507.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Huber, P. (1972). The 1972 wald lecture robust statistics: A review. *Annals of Mathematical Statistics*, 43:1041–1067.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *Annals of Statistics*, 1:799–821.

Jin, Z. and He, W. (2016). Local linear regression on correlated survival data. *Journal of Multivariate Analysis*, 147:285 – 294.

Johnson, W. (1985). Influence measures for logistic regression: Another point of view. *Biometrika*, 72:59–65.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26:189–206.

Karunamuni, R. J., Tang, Q., and Zhao, B. (2015). Robust and efficient estimation of effective dose. *Computational Statistics & Data Analysis*, 90:47 – 60.

Kim, Y. and Jeon, J. J. (2016). Consistent model selection criteria for quadratically supported risks. *Annals of Statistics*, 44:2467–2496.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44:907–927.

Li, A. and Barber, R. F. (2017). Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112:837–849.

Li, G., Peng, H., Zhang, J., and Zhu, L. (2012a). Robust rank correlation based screening. *Annals of Statistics*, 40:1846–1877.

Li, G., Peng, H., and Zhu, L. (2011). Nonconcave penalized m-estimation with a diverging number of parameters. *Statistica Sinica*, 21:391–419.

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

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Annals of Statistics*, 22:1081–1114.

Lindsay, B. G. (2004). Statistical distances as loss functions in assessing model adequacy. In Taper, M. and Lele, S., editors, *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*, pages 439–464. The University of Chicago Press, Chicago.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, 42:413–468.

Lv, J. (2013). Impacts of high dimensionality in finite samples. *Annals of Statistics*, 41:2236–2262.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37:246–270.

Pardo, L. (2005). *Statistical Inference Based on Divergence Measures*. CRC press, New York.

Queiró, J. F. (1987). On the interlacing property for singular values and eigenvalues. *Linear Algebra and Its Applications*, 97:23–28.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer, New York.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing, Volume 1*, pages 319–362. MIT Press, Cambridge, Massachusetts.

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, pages 3859–3869. Curran Associates, Inc., Red Hook.

Sakhinia, E., Faranghpour, M., Liu Yin, J. A., Brady, G., Hoyland, J. A., and Byers, R. J. (2005). Routine expression profiling of microarray gene signatures in acute leukaemia by real-time pcr of human bone marrow. *British journal of haematology*, 130:233–248.

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103:14429–14434.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics*, 40:812–831.

Sherman, J. and Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, 21:124–127.

Shorack, G. (2017). *Probability for statisticians, Second Edition*. Springer, New York.

Tang, L. J., Jiang, J. H., Wu, H. L., Shen, G. L., and Yu, R. Q. (2009). Variable

selection using probability density function similarity for support vector machine classification of high-dimensional microarray data. *Talanta*, 79:260–267.

Tang, Q. and Karunamuni, R. (2013). Minimum distance estimation in a finite mixture regression model. *Journal of Multivariate Analysis*, 120:185–204.

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

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *Annals of Statistics*, 39:1335–1371.

Tropp, J. A. (2012). A comparison principle for functions of a uniformly random subspace. *Probability Theory and Related Fields*, 153:759–769.

Tukey, J. (1970). *Exploratory Data Analysis*. Addison-Wesley, Boston.

Van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3. Cambridge university press.

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

Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. *Annals of Applied Statistics*, 5:468–485.

Wang, X. and Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:589–611.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of statistics*, 37:2178–2201.

Wu, Y., Qin, Y., and Zhu, M. (2018). High-dimensional covariance matrix estimation using a low-rank and diagonal decomposition. *arXiv preprint, arXiv:1802.06048*.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594.

Zhang, C. H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:217–242.

Zhao, J., Leng, C., Li, L., and Wang, H. (2013). High-dimensional influence measure. *Annals of Statistics*, 41:2639–2667.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7:2541–2563.

Zheng, Z., Fan, Y., and Lv, J. (2014). High dimensional thresholded regression and shrinkage effect. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:627–649.

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106:1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101:1418–1429.

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

Curriculum Vitae

Name: Bangxin Zhao

Research Interests

High-dimensional statistics, variable screening, influence measure, post-selection inference, machine learning

Education & Thesis

Ph.D. in Statistics, University of Western Ontario, *expected:* Spring 2018

- Thesis: *Analysis Challenges for High Dimensional Data*
- Supervisor: Prof. Wenqing He

M.S. in Statistics, University of Alberta, Summer 2013

- Thesis: *On Minimum Distance Estimation in Dose Response Studies*
- Supervisor: Prof. Rohana J. Karunamuni

Peer-reviewed Publications

Zhao, B., and He, W. (2018), "Dynamic Tilted Current Correlation for High Dimensional Variable Screening," *submitted*.

Karunamuni, R.J., Tang, Q. and Zhao, B. (2015), "Robust and efficient estimation of effective dose," *Computational Statistics & Data Analysis*, 90:47- 60.

Awards

The Queen Elizabeth II Graduate Scholarships in Science and Technology (QEII-GSST), 2015-2017