Published in final edited form as:

Commun Stat Case Stud Data Anal Appl. 2018 ; 4(2): 82–95. doi:10.1080/23737484.2018.1521315.

A Comparison of High Dimensional Variable Selection Methods with Missing Covariates in a Prostate Cancer Study

Chi Chen, Jiwei Zhao*, Jeffrey Miecznikowski, Marianthi Markatou

Abstract

Prostate cancer is the most common cancer in American men. Dozens of specific genes have been shown to be correlated to prostate cancer, to benign and non-benign cancer cases, from a biology perspective. In this paper, we apply a penalized logistic regression model with different penalty functions to select genes that contribute to benign and non-benign cases, based on the data from a prostate cancer study. The tuning parameter is determined by cross validation. In order to take into account some specific genes that have been classified as prostate cancer genes through biology research but with missing values, multiple imputation is adopted to create complete data sets. We analyze the prostate cancer data by comparing the selection results with completely observed data only, and the results with imputed data. We also conduct a simulation study to validate our proposed method.

Keywords

Prostate cancer; variable selection; missing covariate; high dimensional; multiple imputation; cross validation

1 Introduction

Prostate cancer is the most common cancer in American men, ranking as the 2nd leading cause of cancer death, only behind lung cancer (Siegel et al., 2016). Life of a prostate cancer patient is seriously threatened as the cancer cells may spread from the prostate to other parts of the body, particularly the bones and lymph nodes. The prostate-specific antigen (PSA) blood test is used mainly to try to find prostate cancer early in men without symptoms. PSA is a type of protein released by prostate tissue that is found in higher levels in a man's blood when there is abnormal activity in the prostate (Oesterling, 1991). According to the American Cancer Society's estimates for prostate cancer in the United States for 2016, there are about 182,890 new cases of prostate cancer, and 26,120 deaths from prostate cancer (Siegel et al., 2016). That means about 1 man in 7 will be diagnosed with prostate cancer during his lifetime.

Several factors can affect an individual's risk of getting prostate cancer. The primary risk factor is age. Prostate cancer is very rare in men younger than 40, but the chance of having prostate cancer rises remarkably after age 50. The morbidity of prostate cancer differs

*Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY 14214 (zhaoj@buffalo.edu).

between different races as well. In addition, prostate cancer seems to run in some families, which suggests that there may be an inherited or genetic factor. Fundamentally, prostate cancer is caused by changes in the DNA of a prostate cell. Certain changes (mutation) in DNA will contribute to the abnormal growth of prostate cells and cancers will be formed consequently. DNA changes can either be inherited from a parent or can be acquired during a person's lifetime.

As genetic mutations are associated with cancer, it is beneficial to research prostate cancer by means of analyzing gene expression profiles of patients so that the critical ones can be recognized. Tomlins et al. (2007) analyzed prostate cancer progression using the Molecular Concept Map (MCM), which computes pairwise associations among all genes in the database. The benefit from MCM is the "enrichment networks" of linked concepts that are able to be identified and visualized. This is especially useful for complex gene expression in understanding cancer biology. In this paper, we will analyze the data from the study of Tomlins et al. (2007) with 104 samples denoting tissues of prostate tumors and 20,000 predictors denoting genes. However, only 10% of the predictors are fully observed in the data set, which is a critical issue to be considered and solved.

This paper aims to statistically identify the genes that are associated with prostate cancer, primarily by means of variable selection methods. As the data set is high dimensional (Tomlins et al., 2007), penalized logistic regression will be adopted for variable selection on the fully observed covariates at first. With respect to the missing values, we are interested in the 6 additional covariates marked as prostate cancer genes in terms of biological analysis in the prostate cancer study. We propose two different tracks to conduct multiple imputation, in order to obtain complete covariates of interest. Then penalized logistic regression is conducted on the imputed data set including fully observed covariates as well as the 6 additional covariates. We obtained the results of this setting as a proportion of whether the given covariate is selected or not.

The structure of the paper is as follows. In Section 2, a prostate cancer study in which the data set is analyzed in this paper is introduced. The structure and missing values of the data set is discussed afterwards. Penalized logistic regression with different penalty functions for variable selection is discussed theoretically in Section 3. We perform penalized logistic regression on fully observed covariates at first in Section 4. Multiple imputation is adopted to generate complete data sets in Section 5. The results of variable selection on both fully observed covariates and covariates with imputed values are presented and compared in this section as well. To validate the proposed, we conduct a simulation study in Section 6. Section 7 contains the conclusions and some discussions of this paper.

2 A Prostate Cancer Case Study

In Tomlins et al. (2007), the tissues of the prostate tumors were from the radical prostatectomy series at the University of Michigan and from the Rapid Autopsy Program, which are both part of University of Michigan Prostate Cancer Specialized Program of Research Excellence Tissue Core. Tissues were also obtained from a radical prostatectomy series at the University Hospital Ulm. All samples were collected with informed consent of

the patients and prior institutional review board approval at each institution. After the cells were collected from the tumor tissues, the cell population were isolated. Approximately 10,000 cells were captured for each sample by means of laser-capture microdissection. Isolated total RNA was amplified before expression profiling on 20,000-element cDNA arrays, where each pixel has a corresponding value. As a result, a raw data set of this study with 104 samples and 20,000 genes is generated and available on the NCBI website (<https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3289>). All the data were log transformed and median centered per array, and standard deviation was normalized to one per array. Two data files GDS3289 full soft table.csv (containing data for all genes across the samples) and GDS3289 full soft columns.csv (containing information of disease status for each sample) are extracted from GDS3289 full.soft.gz. Specifically, the disease status consists of 5 levels, benign (44), atrophic lesion (5), metastatic prostate cancer (20), localized prostate cancer (32) and prostatic intraepithelial neoplasia (13).

A $104 \times 20,000$ numeric matrix is generated based on the data from GDS3289 full soft table.csv. After being transposed, the matrix is regarded as a predictor matrix X with 104 rows denoting samples and 20,000 columns denoting covariates (genes). For each sample, there is a response observation corresponding to its disease status. Thus a response vector Y can be formed, where Y is a vector with 104 binary observations. The response is 0 if the disease status is benign and 1 if the disease status is non-benign. Our purpose is to find statistical association between genes and benign/non-benign cancer status, which can be effectively obtained by the variable selection procedure.

2.1 Missing Values

A matrix M with the same dimension as X is created from X , where $M_{ij} = 1$ if X_{ij} is not missing and $M_{ij} = 0$ if X_{ij} is missing. From the matrix M , we are able to summarize the information of missing values in matrix X .

According to the Figure 1, it is obvious that throughout 104 samples, each one contains missing genes. When it comes to the covariates, there are no missing samples in 1,894 genes, as illustrated in Figure 2.

Another visualization method is through the so-called heatmap (Figure 3), which is a figure for the predictor matrix X where the red pixels denote the ones with missing values. It is clear that, only around 10% of the genes are fully observed, out of 20,000 in total.

We denote the predictor matrix with fully observed variables only as X_{obs} . In our following statistical analysis, we will only use fully observed covariates for preliminary analysis.

3 Penalized Logistic Regression

Logistic regression enables us to find the odds that a sample has non-benign prostate cancer versus benign prostate cancer, given the expressions of each gene of that sample. As a member of the family of generalized linear models, logistic regression is frequently applied when the response Y is categorical, which can take only two values (0 or 1). Consider the response vector $Y = (Y_1, \dots, Y_n)^T$ where n is the sample size, and the covariate vector for the

i^{th} sample $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ for $i = 1, \dots, n$ where X_{ij} is the observed value of the j^{th} covariate for the i^{th} sample and p is the number of covariates. The regression model can be formed as

$$\log \frac{P(Y_i = 1 | \mathbf{X}_i)}{P(Y_i = 0 | \mathbf{X}_i)} = \alpha + \mathbf{X}_i^T \boldsymbol{\beta}, \quad (1)$$

where α is the intercept, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is a vector of coefficients and $\beta \in \mathbb{R}^p$.

Given independent and identically distributed samples, the likelihood function is

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^n \left\{ \frac{\exp(\alpha + \mathbf{X}_i^T \boldsymbol{\beta})}{1 + \exp(\alpha + \mathbf{X}_i^T \boldsymbol{\beta})} \right\}^{Y_i} \left\{ \frac{1}{1 + \exp(\alpha + \mathbf{X}_i^T \boldsymbol{\beta})} \right\}^{1-Y_i}, \quad (2)$$

and the corresponding log-likelihood function is

$$l(\alpha, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n \left[Y_i (\alpha + \mathbf{X}_i^T \boldsymbol{\beta}) - \log \{ 1 + \exp(\alpha + \mathbf{X}_i^T \boldsymbol{\beta}) \} \right]. \quad (3)$$

In high dimensional data settings where the number of covariates is greater than the sample size ($p > n$), the result of parameter estimation tends to be imprecise or infeasible by conventional estimating methods. Penalized logistic regression (PLR) is the method frequently employed for variable selection within high dimensional data. In penalized logistic regression, the estimates of parameters are computed based on minimizing the objective function:

$$l^{(\lambda)}(\alpha, \boldsymbol{\beta}) = -\frac{1}{n} \sum_{i=1}^n \left[Y_i (\alpha + \mathbf{X}_i^T \boldsymbol{\beta}) - \log \{ 1 + \exp(\alpha + \mathbf{X}_i^T \boldsymbol{\beta}) \} \right] + P_\lambda(\boldsymbol{\beta}), \quad (4)$$

where $P_\lambda(\boldsymbol{\beta})$ is the penalty function. Different choices of penalty function $P_\lambda(\boldsymbol{\beta})$ can be employed for variable selection.

3.1 LASSO

The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model (Tibshirani, 1996). For the i^{th} sample, we have the response Y_i and a vector of p covariates $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$, LASSO is applied to solve

$$\arg \min_{\alpha, \boldsymbol{\beta}} \left(-\frac{1}{n} \sum_{i=1}^n \left[Y_i (\alpha + \mathbf{X}_i^T \boldsymbol{\beta}) - \log \{ 1 + \exp(\alpha + \mathbf{X}_i^T \boldsymbol{\beta}) \} \right] + \lambda \|\boldsymbol{\beta}\|_1 \right), \quad (5)$$

where $\lambda > 0$ is the tuning parameter, and $\|\cdot\|_1$ is the standard L_1 norm. Due to its convexity, the coordinate descent algorithm can be efficiently applied for function minimization (Friedman et al., 2010).

3.2 Nonconvex Penalties

Although LASSO is the most popular penalized regression method, it has been shown that the shrinkage introduced by the LASSO results in significant bias towards zero for large regression coefficients. The estimate of parameter vector β , denoted as $\hat{\beta}$, solved by derivation of objective function is approximately the same as the maximum likelihood estimate so that it is an unbiased estimator, given $\lambda\|\beta\|_1$ is small enough. Otherwise, the estimate is no longer the approximate maximum likelihood estimate of β and has bias as a result. Obviously, a large number of variables or a large λ will result in a considerable bias of estimates.

The penalty function $P_\lambda(\beta)$ in LASSO is given as $\lambda\|\beta\|_1$, which is a convex function of β . In fact, the likelihood function of LASSO is also a convex function as it is a sum of two convex functions. It is not complex to solve $\hat{\beta}$. However, the possible bias of coefficient estimates due to large $\lambda\|\beta\|_1$ specified cannot be ignored.

A large family of nonconvex penalties are proposed to diminish the bias in the statistical literature. Two representatives are the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010). Both methods satisfy the so-called oracle property in the sense that, the estimators are asymptotically consistent to the truth if the analyst had known in advance which coefficients were zero and which were not.

When it comes to the logistic regression with nonconvex penalties (SCAD and MCP), the first derivative of the penalty function $P_\lambda(\beta)$ in (4) is given by

$$P'_\lambda(t) = \lambda I\{t \leq \lambda\} + \frac{(a\lambda - t)_+}{a-1} I\{t > \lambda\} \quad (6)$$

for SCAD where $a > 2$ and $t > 0$, and

$$P'_\lambda(t) = \frac{(a\lambda - t)_+}{a} \quad (7)$$

for MCP where $a > 1$ and $t > 0$.

3.3 Tuning Parameter Selection

In penalized logistic regression, the solution of nonzero coefficients is indexed by the tuning parameter λ . As a result, we need to determine the optimal value of λ in advance to solving for the optimal numbers of variables and corresponding estimates of coefficient. There are various methods proposed in the literature to determine the optimal tuning parameter. In this paper, we adopt the K -fold cross-validation approach which can be obtained by the following steps:

1. Partition the original data set T into K separate sets of equal size, where T is define as $T = (T_1, \dots, T_K)$.
2. For $k = 1, \dots, K$, fit the model $l_{-k}^{(\lambda)}(\alpha, \beta)$ to the training set T_{-k} , where $T_{-k} = (T_1, \dots, T_{k-1}, T_{k+1}, \dots, T_K)$ and $l_{-k}^{(\lambda)}(\alpha, \beta)$ is the objective function defined in (4). Derive the estimates $\hat{\alpha}_{-k}$ and $\hat{\beta}_{-k}$ in the form of λ .
3. Compute the cross-validation error (CVE) for each k , which is given by $2l_k(\hat{\alpha}_{-k}, \hat{\beta}_{-k})$, on test data set T_k . The overall cross-validation error for the model is defined as

$$\text{CVE}(\lambda) = \frac{1}{K} \sum_{k=1}^K \{2l_k(\hat{\alpha}_{-k}, \hat{\beta}_{-k})\}. \quad (8)$$

4. Find a λ such that $\text{CVE}(\lambda)$ attains its minimum value, denote it as λ^* . Then λ^* is the optimal tuning parameter.

4 Variable Selection with Fully Observed Data

When it comes to the dataset from Tomlins et al. (2007), we consider the subset of $\mathbf{X}$ with fully observed covariates only. We can partition $\mathbf{X}$ as $\mathbf{X} = (\mathbf{X}_{obs}, \mathbf{X}_{mis})$, where $\mathbf{X}_{obs}$ contains all the fully observed covariates and $\mathbf{X}_{mis}$ contains the covariates with missing values. Specifically, $\mathbf{X}_{obs}$ has 104 rows and 1,894 columns. Meanwhile, we have a trimmed parameter vector β_{obs} .

Based on the data in this study, we will estimate the unknown parameters using 1,894 fully observed variables at first. The predictor matrix is denoted as $\mathbf{X}_{obs}$ and the response matrix is denoted as $\mathbf{Y}$. Penalized logistic regression will be performed on $\mathbf{Y}$ and $\mathbf{X}_{obs}$. We are interested in the estimates of the corresponding 1,894 coefficients. Consider equation (4) again, we are going to solve

$$\arg \min_{\alpha, \beta} \left[-\frac{1}{n} \sum_{i=1}^n [Y_i(\alpha + (\mathbf{X}_{obs,i})^T \beta_{obs}) - \log\{1 + \exp(\alpha + (\mathbf{X}_{obs,i})^T \beta_{obs})\}] + P_{\lambda}(\beta) \right]. \quad (9)$$

Packages glmnet (Friedman et al., 2010) and ncvgreg (Breheny and Huang, 2011) can be applied directly in R to estimate λ^* and $\hat{\beta}_{obs}$ in LASSO and nonconvex penalties respectively.

4.1 LASSO

A plot of solution path generated by glmnet established the graphical relationship between the coefficients of each variable and some statistics from data (mainly includes L_1 -norm, log λ , and percent deviance explained). Figure 4 shows the change of the estimates against λ . Graphically, as λ decreases, there will be more genes being selected. The process of LASSO

involves minimizing the objective function (4), while controlling $\sum_{j=1}^p \lambda \left| \hat{\beta}_j \right|$ plays an important role in objective function minimization. As a result, $\left| \hat{\beta}_j \right|$ will be able to deviate from 0 with less restriction, given λ is close to 0. Otherwise, $\left| \hat{\beta}_j \right|$ will shrink to 0.

Although the plot of solution path has highlighted the relationship between estimates of coefficients against tuning parameter, we can not solve the optimal tuning parameter λ^* based on the graph of solution path only, while estimates of coefficients are not available unless the value of λ^* is given. K -fold cross-validation is performed in glmnet in order to solve for λ^* . Specifically, the number of cross-validation folds is set as $K = 10$ by default in glmnet.

4.2 Nonconvex Penalties

In practice, we can invoke package ncvgreg to perform both SCAD and MCP. Similarly, the plots can be generated to illustrate the result of the solution path. The default number of cross-validation folds is set as $K = 10$ in ncvgreg. Figure 5 shows the change of total number of nonzero coefficients and values for each of them, against the λ sequence through SCAD. The shaded region is the region where the objective function is not locally convex. Similarly, Figure 6 shows the result for MCP.

5 Variable Selection with Missing Data via Imputation

Consider the predictor matrix X (X is a $104 \times 20,000$ matrix) again, where only 1,894 genes are fully observed. The rest occupies around 90% of the total number of genes, among which there probably exists some interesting genes. In the prostate cancer genes (listed in Figure 7 of Tomlins et al. (2007)) specifically, there are 6 genes with missing values that are of interest for further analysis. Table 1 summarizes the 6 genes with missing values in terms of number of missing cells contained. We would like to take into account these genes in our model and figure out if they are associated with the benign/non-benign prostate cancer status. However, some specific methods are essential prior to variable selection by means of penalized logistic regression.

Several methods have been developed to analyze data sets that contain missing values; the two most frequently employed are multiple imputation (MI) and inverse probability weighting (IPW). In the IPW approach, a model for the probability that an individual is a complete case is specified, which is conventionally called missingness model. In the MI approach, a model for the distribution of the missing data given observed data is specified instead. There also exists some literature focusing on nonignorable missing data, or under a generally applied missing data mechanism. For example, based upon the identification conditions, pseudo likelihood type methods have been developed for missing responses (Zhao and Shao, 2015; Zhao, 2017), covariates (Fang et al., 2018) and repeated measurements (Shao and Zhao, 2013), under a general missing data mechanism including both ignorable and nonignorable ones. In general, the MI approach is preferred in that the missingness model can only use fully observed variables unless the data is monotone missing. Let $V = (V^1, \dots, V^K)$ denote a vector of variables measured on each individual from sample. Then V is said to be monotone missing when V^k is observed on an individual only if

V^{k-1} is observed. In addition, MI is generally more efficient in computation (Seaman and White, 2013; Zhao et al., 2015). As a result, MI will be discussed and employed to deal with missing values in the predictor matrix X in the next step of this paper.

5.1 Introduction to Multiple Imputation

In the process of imputation, missing data will be replaced with predicted values. Single imputation is a basic type of imputation technique where each missing value is substituted with a predicted value so that the data set can be treated as a complete one. Consequently, standard statistical analysis approaches can be conducted on the completed data set. However, the bias and variance of parameter estimates will be reduced if the distribution of the missing value can be obtained rather than a single predicted value. Rubin proposed a multiple imputation procedure to replace each missing value with a set of possible values that represent the uncertainty about the right value to impute (Rubin, 2004). In terms of estimating each missing value, the fundamental difference between multiple imputation and single imputation is that we can obtain a random sample of the missing value through multiple imputation, while only a single simulated value is provided in single imputation.

Consider a data set W and let W_j indicate one of p incomplete variables, where $j = 1, \dots, p$. The subset for observed data for W_j is denoted as W_j^{obs} and the subset for missing data is denoted as W_j^{mis} . We can obtain $W^{obs} = (W_1^{obs}, \dots, W_p^{obs})$ and $W^{mis} = (W_1^{mis}, \dots, W_p^{mis})$.

Assume M multiple imputations are performed, the m^{th} imputed data set is denoted as $W^{(m)}$ where $m = 1, \dots, M$. We define $W_{-j} = (W_1, \dots, W_{j-1}, W_{j+1}, \dots, W_p)$, which is the collection of the $p - 1$ variables in W except W_j . Let Q denote the quantity of interest such as the set of parameter to be estimated, which is often a multivariate vector.

Without information or assumption on the missing data, we are not able to make inference on Q from W^{obs} . Multiple imputation is required to perform in order to replace the missing values by plausible data values generated from M imputed versions.

Multiple imputation inference involves three primary phases:

1. The missing values are filled in M multiple imputations. In total M complete data sets $W^{(1)}, \dots, W^{(M)}$ are generated after imputation.
2. Estimate Q on each imputed data set $W^{(m)}$ for $m = 1, \dots, M$, denote as $\hat{Q}^{(m)}$. Although the model assigned to $W^{(m)}$ is identical for each m , the estimates $\hat{Q}^{(1)}, \dots, \hat{Q}^{(M)}$ cannot be guaranteed to be the same to each other due to uncertainty of values imputed.
3. Pool the m estimates $\hat{Q}^{(1)}, \dots, \hat{Q}^{(M)}$ into one estimate $\bar{Q}$ and estimate its variance.

The package mice in R can be adopted to conduct multivariate imputation by chained equations (MICE), which contains two general approaches for imputing multivariate data: joint modeling and fully conditional specification (Buuren and Groothuis-Oudshoorn, 2011).

Consider $\mathbf{W}$ with missing values from a p -variate multivariate distribution $P(\mathbf{W}|\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$ is a vector of p unknown parameters. The posterior distribution of $\boldsymbol{\theta}$ can be obtained by sampling iteratively from the conditional distribution

$$P(\mathbf{W}_j | \mathbf{W}_{-j}, \theta_j) \quad (10)$$

for $j = 1, \dots, p$. We can start drawing from observed marginal distribution and the t^{th} iteration of chained equations is a Gibbs sampler that successively draws

$$\theta_j^{*(t)} \sim P(\theta_j | \mathbf{W}_j^{\text{obs}}, \mathbf{W}_1^{(t)}, \dots, \mathbf{W}_{j-1}^{(t)}, \mathbf{W}_{j+1}^{(t-1)}, \dots, \mathbf{W}_p^{(t-1)}) \quad (11)$$

$$\mathbf{W}_j^{*(t)} \sim P(\mathbf{W}_j | \mathbf{W}_j^{\text{obs}}, \mathbf{W}_1^{(t)}, \dots, \mathbf{W}_{j-1}^{(t)}, \mathbf{W}_{j+1}^{(t-1)}, \dots, \mathbf{W}_p^{(t-1)}, \theta_j^{*(t)}) \quad (12)$$

where $\mathbf{W}_j^{(t)} = (\mathbf{W}_j^{\text{obs}}, \mathbf{W}_j^{*(t)})$ is the j^{th} imputed variable at iteration t (Buuren and Groothuis-Oudshoorn, 2011). The mice() function executes M streams of imputation in parallel, each of which generates one imputed data set.

The ideal track of multiple imputation is to make imputation on the whole data set. However, it is not feasible to be executed in this project since there are 20,000 variables and 90% of them contain at least one missing value. Consequently, it will consume too much memory to finish all the iterations on the missing covariates. We will make imputation on the data subset extracted from the whole data. In the next two different compromising methods of multiple imputation on two different targeted subset data of predictors will be introduced.

5.2 Imputation Method A

We define the first method for multiple imputation as Imputation Method A. Let $\mathbf{X}_{\text{obs}}$ denote the matrix of observed values of all the 1,894 fully observed variables, and $\mathbf{X}_{A_j}$ denote the j^{th} column of variables indicating the 6 genes (MYC, EZH2, PTEN, GSTP1, NKX3-1, AZGP1) with missing values, for $j = 1, \dots, 6$. We make multiple imputation on $\mathbf{X}_A$, which can be partitioned as

$$\mathbf{X}_A = (\mathbf{X}_{A_1}, \dots, \mathbf{X}_{A_6}). \quad (13)$$

After M multiple imputations are performed, M completed data sets are generated for the 6 variables, each of which is denoted as $\mathbf{X}_A^{(m)}$, for $m = 1, \dots, M$. It is notable that 100 multiple imputations make it almost indifferent in terms of the result of variable selection so that the convergence is satisfied. As a result, a maximum of 100 imputations are enough for further inference in this scenario. Assume we have made M multiple imputations, we output M different matrices denoted as $\mathbf{X}_A^{(m)}$ completed by the m^{th} action of M multiple imputations, where $m = 1, \dots, M$. A new matrix $\tilde{\mathbf{X}}^{(m)}$ can be defined as a merge of $\mathbf{X}_{\text{obs}}$ and $\mathbf{X}_A^{(m)}$ with columns, $\tilde{\mathbf{X}}^{(m)} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_A^{(m)})$, where $\tilde{\mathbf{X}}^{(m)}$ is a $104 \times 1,900$ matrix without missing values.

For each m within a given M , we can estimate the coefficients by LASSO, SCAD and MCP by treating $\tilde{X}^{(m)}$ as the predictor matrix, the output of which can be denoted as $\hat{\beta}_{\text{LASSO}}^{(m)}$, $\hat{\beta}_{\text{SCAD}}^{(m)}$ and $\hat{\beta}_{\text{MCP}}^{(m)}$ respectively.

Generally, we have 1,900 coefficients for covariates to estimate, which can be partitioned as $\beta = (\beta_{\text{obs}}^T, \beta_{\text{mis}}^T)^T$. The corresponding estimates of coefficients β denoted as $\hat{\beta}^{(m)}$, can be solved via

$$\arg \min_{\alpha, \beta} \left[-\frac{1}{n} \sum_{i=1}^n \left[Y_i \left(\alpha + (\tilde{X}_i^{(m)})^T \beta \right) - \log \left\{ 1 + \exp \left(\alpha + (\tilde{X}_i^{(m)})^T \beta \right) \right\} \right] + P_{\lambda}(\beta) \right]. \quad (14)$$

The proportion that the j^{th} coefficient is selected as significant based on LASSO given M multiple imputations is denoted as

$$P_{j, \text{LASSO}}^{(M)} = \frac{1}{M} \sum_{m=1}^M I\{\hat{\beta}_{j, \text{LASSO}}^{(m)} \neq 0\}, \quad (15)$$

for $j = 1, \dots, 1900$, where $\hat{\beta}_{j, \text{LASSO}}^{(m)}$ denotes the estimate of β_j via LASSO, given M multiple imputations are performed and the completed data set is generated by the m^{th} action. We can create a vector $\mathbf{P}_{\text{LASSO}}^{(M)} = (P_{1, \text{LASSO}}^{(M)}, \dots, P_{1900, \text{LASSO}}^{(M)})^T$. Similarly, we can output $\mathbf{P}_{\text{SCAD}}^{(M)}$ and $\mathbf{P}_{\text{MCP}}^{(M)}$ respectively.

5.3 Imputation Method B

The other approach is defined as Imputation Method B. The initial step is to generate 3 individual matrices with fully observed variables that have been selected by LASSO, SCAD or MCP respectively, denoted as $\mathbf{X}_{\text{LASSO}}$, $\mathbf{X}_{\text{SCAD}}$ and $\mathbf{X}_{\text{MCP}}$. Consider the case of LASSO, we merge two matrices $\mathbf{X}_{\text{LASSO}}$ and $\mathbf{X}_A$ (defined in (13)) together by column as a new predictor matrix with missing values, denoted as $\tilde{\mathbf{X}}_{\text{LASSO}}$. Thus $\tilde{\mathbf{X}}_{\text{LASSO}}$ is a $104 \times (N_{\text{LASSO}} + 6)$ matrix, where N_{LASSO} is the number of variables selected from 1,894 fully observed covariates through LASSO. Assume we have made M multiple imputations and output completed data set $\tilde{\mathbf{X}}_{\text{LASSO}}^{(m)}$ using the m^{th} action, where $m = 1, \dots, M$. Similarly, we can obtain $\tilde{\mathbf{X}}_{\text{SCAD}}$, $\tilde{\mathbf{X}}_{\text{MCP}}$ and output $\tilde{\mathbf{X}}_{\text{SCAD}}^{(m)}$, $\tilde{\mathbf{X}}_{\text{MCP}}^{(m)}$ in the same way.

Based on $\tilde{\mathbf{X}}_{\text{LASSO}}^{(m)}$, $\tilde{\mathbf{X}}_{\text{SCAD}}^{(m)}$ and $\tilde{\mathbf{X}}_{\text{MCP}}^{(m)}$, we can obtain the estimates of coefficients $\hat{\beta}_{\text{LASSO}}^{(m)}$, $\hat{\beta}_{\text{SCAD}}^{(m)}$ and $\hat{\beta}_{\text{MCP}}^{(m)}$ for given m . In this case, we still have 1,900 coefficients for covariates to estimate and the corresponding estimates $\hat{\beta}^{(m)}$ can be solved via

$$\arg \min_{\alpha, \beta} \left[-\frac{1}{n} \sum_{i=1}^n \left[Y_i \left(\alpha + (\tilde{\mathbf{X}}_i^{(m)})^T \beta \right) - \log \left\{ 1 + \exp \left(\alpha + (\tilde{\mathbf{X}}_i^{(m)})^T \beta \right) \right\} \right] + P_{\lambda}(\beta) \right]. \quad (16)$$

The proportion that the j^{th} coefficient is selected under LASSO given M multiple imputations can be obtained by (15) similarly. Analogous to the procedure in Imputation Method A, we can output $P_{LASSO}^{(M)}$, $P_{SCAD}^{(M)}$ and $P_{MCP}^{(M)}$.

It is notable that in Imputation Method A, we make multiple imputation on the same data set while the difference of $P_{LASSO}^{(M)}$, $P_{SCAD}^{(M)}$ and $P_{MCP}^{(M)}$ results from difference in completed predictor matrix after imputation and penalized logistic regression. However, when it comes to Imputation Method B, the difference of the 3 final proportions results from difference in predictor matrix for imputation (dimensions), as well as completed predictor matrix after imputation (dimensions and values) and penalized logistic regression.

5.4 Results

Table 2 shows the result of variable selection by LASSO. Before imputation, the coefficients are estimated by taking LASSO on all the 1,894 fully observed variables. After imputation, the proportion that the variable is selected under M imputations both in Imputation Method A and Imputation Method B are summarized. Table 3 and Table 4 show the results generated by SCAD and MCP respectively. We make multiple imputations 20, 50, 100 times respectively and the convergence of $P^{(m)}$ is achieved when m is large.

6 Simulation Study

We conduct a simulation to validate our proposed multiple imputation method used in this prostate cancer study. We design our simulation to have similar parameters and dimensions to the prostate cancer data set. In each of the simulation replicate, the sample size n is 100 and dimension of covariate p is 1,900. Let X_i be the covariate matrix for the i^{th} sample ($i = 1, \dots, 100$), where $X_i \sim N_{1,900}(\mathbf{0}, \Sigma)$. The cell in the i^{th} row and j^{th} column of Σ is $\sigma_{ij} = 0.5^{|i-j|}$. Let Y_i denote the response of the i^{th} sample, which can take value of either 1 or 0. Specifically,

$$P(Y_i = 1 | X_i) = \frac{\exp(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_7 X_{i7} + \beta_8 X_{i8} + \beta_9 X_{i9})}{1 + \exp(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_7 X_{i7} + \beta_8 X_{i8} + \beta_9 X_{i9})},$$

where $\alpha = 1$, $(\beta_1, \beta_2, \beta_3) = (1, 1, 1)$ and $(\beta_7, \beta_8, \beta_9) = (1, 1, 1)$. In addition, we assume missing values exist in $X_1, \dots, X_6$ and $P(R_{ij} = 1 | X_{ij}) = 0.9$, where $R_{ij} = 1$ indicates X_{ij} is observed. We let $X_{obs} = (X_7, \dots, X_{1900})$ denote the predictor matrix with fully observed covariates. It can be seen that, we have in total three covariates which have nonzero coefficients and have missing values, three covariates which have nonzero coefficients and are fully observed, three covariates which have zero coefficients and have missing values, and 1,891 covariates which have zero coefficients and are fully observed.

For each simulation, we summarize the following results. Before imputation, we count the number of the selected covariates, either fully observed or has missing values. After imputation, we compute the proportion that each covariate is selected, under Imputation Method A and Imputation Method B, with 20, 50 and 100 imputations. Based on 100 simulation replicates, we take the average of the proportion that each covariate is selected.

We then report the average of these averaged proportions in four different categories of covariates: truth is nonzero and has missing values, truth is nonzero and fully observed, truth is zero and has missing values, truth is zero and fully observed, in Table 5. The simulation is conducted using the R language. The R package mice (Buuren and Groothuis-Oudshoorn, 2011) is used to perform the multiple imputation technique. The coordinate descent algorithm is used for the regularization by using the packages glmnet (Friedman et al., 2010) for the LASSO and ncvmreg (Breheny and Huang, 2011) for the nonconvex penalties.

From Table 5, we can reach the following conclusions. First, the before imputation method cannot select the covariates which have missing values since these covariates have to be discarded if no imputation method is used. This issue could become severe if there are missing values contained in many clinically proven relevant variables. Second, after imputation, we have a much higher probability, although not perfect, to select those potentially important variables which have missing values (the nonzero, missing rows in Table 5). This is the biggest benefit of using the multiple imputation technique in variable selection. Third, we observe that, although it has a relatively higher probability to select the relevant variables (nonzero coefficient) than SCAD and MCP, the LASSO method also has a higher probability to select the irrelevant variables, i.e., a higher false positive rate. Due to this tradeoff, we cannot reach a definite conclusion that LASSO performs better than nonconvex penalties. This observation is also consistent to the phenomenon in our real data analysis that the LASSO selects more variables than nonconvex penalties. Finally, we observe that for nonzero-truth variables the Imputation Method A outperforms the Method B, whereas for zero-truth ones the Method B performs slightly better in a few scenarios. Since the Imputation Method B is computationally more involved and there is no conclusive evidence to be the superior, we generally recommend to apply the Imputation Method A in real applications. The Imputation Method B can certainly be further explored if the computational capacity is reached.

7 Conclusion and Discussion

In this paper, our overarching goal is to compare different high dimensional variable selection methods when the covariates have missing values. We analyze a data set from a prostate cancer study and we adopt the multiple imputation technique to handle the six genes that are of interest but have missing values. We contrast the variable selection results by comparing situations before imputation and after imputation. Since we have multiple complete data sets after multiple imputation, multiple selection results are obtained. Therefore, the selection results after imputation is a probability instead of a definite zero/one.

In general, the method with LASSO selects the most variables, while both SCAD and MCP produce a more parsimonious model. In our case study, the MCP obtains the least number of “important” variables. Out of the four fully observed genes that are of our major interest (blue highlighted in Table 2, Table 3, Table 4), both LASSO and SCAD identify them either before imputation or after imputation, but MCP selects only two of them. Interestingly, for the six genes that are of our major interest but with missing values, the methods before

imputation cannot select them since they are not in the model, and the methods after imputation also donot select them at all.

There indeed exist some disagreements before and after imputation. For example, in the results with the LASSO, the gene types CCDC93 and MAPRE1 are not chosen before imputation but with a relatively high probability to be chosen after imputation. Similar situations happen for CPS1 and ZNF410 in the results with the SCAD, and ANTXR2 in the results with the MCP (red highlighted in Table 2, Table 3, Table 4). Interestingly, all these five genes are shown to be associated with prostate cancer in previous literature (Liu et al., 2014; Sharad et al., 2011; Eick et al., 2006; Ladd et al., 2012; Panagiotou et al., 2015).

Our study is not without limitations. For example, as we discussed in the Introduction, age, race and family history are also important covariates for prostate cancer. However, since these variables are not available in this study, we cannot take them into account. In our prostate cancer data analysis, we cannot take all 20,000 variables into account since around 90% of them have missing values. Instead, we only focus on imputing the clinically important ones, the prostate cancer genes. One can also use some screening method, e.g., sure independence screening (Fan and Song, 2010), to first obtain a different set of genes of their own interest and then apply the imputation methods proposed and detailed in this paper. We can also see that three different penalty functions give rise to different selection results. In all, we cannot reach a certain conclusion on which genes are definitely associated with the prostate cancer benign/non-benign status and which are not. Hopefully, our attempt with the multiple imputation approach for missing covariate data in high dimensional variable selection could help to reveal more insightful clinical findings in the area of prostate cancer investigation.

Finally, the general statistical methodology for addressing missing data in high dimensional variable selection is still lacking. Based upon the results on model identifiability (Zhao and Shao, 2017), Zhao et al. (2018) proposed a regularization framework for variable selection under high dimensional setting, without imputation. The imputation technique proposed in this paper shares some similar spirit to Chen and Xie (2014), where their penalization is applied to multiple data sets, and then the majority voting is used to account for the different results across these data sets. In general, the use of multiple imputation for variable selection under high dimensionality warrants further more rigorous investigations and it is one of our future research topics.

Acknowledgements

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR001412. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. We thank the editor, the associate editor and the anonymous referee for their constructive comments and insightful suggestions, which have led to a significantly improved paper.

References

- Breheny P and Huang J (2011), "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The Annals of Applied Statistics*, 5, 232–253. [PubMed: 22081779]

- Buuren S and Groothuis-Oudshoorn K (2011), “mice: Multivariate imputation by chained equations in R,” *Journal of Statistical Software*, 45.
- Chen X and Xie M.-g. (2014), “A split-and-conquer approach for analysis of extraordinarily large data,” *Statistica Sinica*, 1655–1684.
- Eick G, Kidd M, Mane S, Nadler B, Champaneria M, Pfragner R, and Modlin I (2006), “geNorm assessment of the efficacy of novel reference genes ALG9, TFCEP2 and ZNF410 compared to GAPDH for real time normalization in GI carcinoids,” *Journal of Clinical Oncology*, 24, 20052–20052.
- 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 Song R (2010), “Sure independence screening in generalized linear models with NP-dimensionality,” *The Annals of Statistics*, 38, 3567–3604.
- Fang F, Zhao J, and Shao J (2018), “Imputation-based adjusted score equations in generalized linear models with nonignorable missing covariate values,” *Statistica Sinica*, in press.
- Friedman J, Hastie T, and Tibshirani R (2010), “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software*, 33, 1. [PubMed: 20808728]
- Ladd JJ, Busald T, Johnson MM, Zhang Q, Pitteri SJ, Wang H, Brenner DE, Lampe PD, Kucherlapati R, Feng Z, et al. (2012), “Increased plasma levels of the APC-interacting protein MAPRE1, LRG1, and IGFBP2 preceding a diagnosis of colorectal cancer in women,” *Cancer Prevention Research*, 5, 655–664. [PubMed: 22277732]
- Liu F, Feng Y, Li Z, Pan C, Su Y, Yang R, Song L, Duan H, and Deng N (2014), “Clinic-genomic association mining for colorectal cancer using publicly available datasets,” *BioMed Research International*, 2014.
- Oesterling JE (1991), “Prostate specific antigen: a critical assessment of the most useful tumor marker for adenocarcinoma of the prostate,” *The Journal of Urology*, 145, 907–923. [PubMed: 1707989]
- Panagiotou OA, Travis RC, Campa D, Berndt SI, Lindstrom S, Kraft P, Schumacher FR, Siddiq A, Papatheodorou SI, Stanford JL, et al. (2015), “A genome-wide pleiotropy scan for prostate cancer risk,” *European Urology*, 67, 649–657. [PubMed: 25277271]
- Rubin DB (2004), *Multiple Imputation for Nonresponse in Surveys*, vol. 81, John Wiley & Sons.
- Seaman SR and White IR (2013), “Review of inverse probability weighting for dealing with missing data,” *Statistical Methods in Medical Research*, 22, 278–295. [PubMed: 21220355]
- Shao J and Zhao J (2013), “Estimation in longitudinal studies with nonignorable dropout,” *Statistics and Its Interface*, 6, 303–313.
- Sharad S, Srivastava A, Ravulapalli S, Parker P, Chen Y, Li H, Petrovics G, and Dobi A (2011), “Prostate cancer gene expression signature of patients with high body mass index,” *Prostate Cancer and Prostatic Diseases*, 14, 22–29. [PubMed: 21060327]
- Siegel RL, Miller KD, and Jemal A (2016), “Cancer statistics, 2016,” *CA: a cancer journal for clinicians*, 66, 7–30. [PubMed: 26742998]
- Tibshirani R (1996), “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.
- Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, Dhanasekaran SM, Kalyana-Sundaram S, Wei JT, Rubin MA, Pienta KJ, et al. (2007), “Integrative molecular concept modeling of prostate cancer progression,” *Nature Genetics*, 39, 41–51. [PubMed: 17173048]
- Zhang C-H (2010), “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, 38, 894–942.
- Zhao J (2017), “Reducing bias for maximum approximate conditional likelihood estimator with general missing data mechanism,” *Journal of Nonparametric Statistics*, 577–593. [PubMed: 31551650]
- Zhao J, Cook RJ, and Wu C (2015), “Multiple imputation for the analysis of incomplete compound variables,” *Canadian Journal of Statistics*, 43, 240–264.
- Zhao J and Shao J (2015), “Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data,” *Journal of the American Statistical Association*, 110, 1577–1590.

- Zhao J and Shao J (2017), “Approximate conditional likelihood for generalized linear models with general missing data mechanism,” *Journal of Systems Science and Complexity*, 30, 139–153.
- Zhao J, Yang Y, and Ning Y (2018), “Penalized pairwise pseudo likelihood for variable selection with nonignorable missing data,” *Statistica Sinica*, in press.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1:

Proportion (out of 104 samples) of samples based on the number of missing genes in each sample. Note that the total number of samples with fully observed genes is zero.

Figure 2:
Proportion (out of 20,000 gene types) of genes based on the number of missing samples in each gene. Note that the total number of genes with fully observed samples is 1,894.

Figure 3:
Heatmap of predictor matrix X , where the yellow pixel denotes the corresponding cell is observed, and the red pixel denotes missing.

Figure 4:
Solution pathes using 1,894 fully observed genes with LASSO

Figure 5:
Solution pathes using 1,894 fully observed genes with SCAD

Figure 6:
Solution pathes using 1,894 fully observed genes with MCP

Table 1:

Summary of the number of missing samples (out of 104 in total) for prostate cancer genes.

Gene	Number of Missing Samples (out of 104)
MYC	1
EZH2	53
PTEN	2
GSTP1	1
NKX3-1	48
AZGP1	1
AMACR	0
SLC25A6	0
MYO6	0
MME	0

Table 2:

Comparison (before imputation and after imputation) of the variable selection results using LASSO.

Index	Gene	Before Imputation Parameter Estimation	Imputation Method A Selection Proportion			Imputation Method B Selection Proportion		
			$\hat{\beta}$	$P_{\text{LASSO}}^{(20)}$	$P_{\text{LASSO}}^{(50)}$	$P_{\text{LASSO}}^{(100)}$	$P_{\text{LASSO}}^{(20)}$	$P_{\text{LASSO}}^{(50)}$
91	IMAGE: 202168	0.1892	1	1	1	1	1	1
288	ATP8A1	0	0	0	0	0.05	0.02	0.02
319	IMAGE: 40728	−0.0874	0.25	0.24	0.32	0.35	0.26	0.26
435	IMAGE: 470914	0.3968	1	1	1	1	1	1
665	CKS1B	0.0958	0.65	0.64	0.70	0.60	0.70	0.72
677	EPC1	0	0.05	0.06	0.10	0.20	0.12	0.08
876	ODF2	0	0.30	0.34	0.29	0.20	0.18	0.19
1065	CPS1	0.8814	1	1	1	1	1	1
1079	IMAGE: 416614	0.4348	1	1	1	1	1	1
1248	SPRY1	0.2701	0.95	0.98	0.99	1	0.98	0.97
2525	NAB2	0.6846	0.60	0.60	0.67	0.55	0.68	0.67
2643	MYO1B	1.2384	0.90	0.96	0.96	0.95	0.96	0.94
2851	WIBG	0.4769	1	1	1	1	0.98	0.99
3137	MME	−0.9807	1	1	1	1	1	1
4535	ACOX2	−1.3396	1	1	1	1	1	1
5071	ST14	1.3229	1	1	1	1	1	1
5176	TGOLN2	0	0.05	0.02	0.04	0.15	0.08	0.05
5326	SLC39A1	0.4074	0.90	0.90	0.87	0.95	0.90	0.88
5381	ZNF337	0	0.30	0.34	0.29	0.20	0.18	0.19
5581	SLC25A6	1.2666	1	1	1	1	1	1
6301	NFKB1	0.1823	1	1	1	1	0.98	0.99
7124	IMAGE: 490971	0.2061	0.90	0.96	0.96	1.00	0.98	0.95
7547	POLR2A	0.2620	0.65	0.64	0.70	0.60	0.70	0.72
7605	IMAGE: 111714	2.1862	1	1	1	1	0.98	0.99
7739	SOX4	1.2760	1	1	1	1	1	1
8210	RPS3A	0.1460	0.25	0.24	0.32	0.35	0.26	0.26
8446	LGALS1	−2.0863	1	1	1	1	1	1
9843	AMACR	1.2443	1	1	1	1	1	1
10276	MYO6	0.8293	1	1	1	1	1	1
10281	MEIS2	0	0	0	0	0	0.02	0.01
10295	GAA	0.0123	0.30	0.28	0.32	0.20	0.26	0.28
10365	ATXN7	0.6761	1	1	1	1	1	1

Index	Gene	Before Imputation Parameter Estimation	Imputation Method A Selection Proportion			Imputation Method B Selection Proportion		
			$\hat{\beta}$	$P_{\text{LASSO}}^{(20)}$	$P_{\text{LASSO}}^{(50)}$	$P_{\text{LASSO}}^{(100)}$	$P_{\text{LASSO}}^{(20)}$	$P_{\text{LASSO}}^{(50)}$
10532	CCDC93	0	0.70	0.70	0.63	0.65	0.70	0.69
11322	PIK3R1	−0.3203	0.80	0.78	0.78	0.90	0.88	0.85
11482	MAPRE1	0	0.65	0.52	0.46	0.45	0.46	0.44
12470	HADHA	0	0.15	0.14	0.16	0.10	0.12	0.13
12585	RBM22	0.4493	0.60	0.60	0.67	0.55	0.68	0.67
14003	RPSA	0.3257	1	1	1	1	1	1
14114	RNF141	−0.5457	0.50	0.56	0.62	0.55	0.64	0.63
14398	ZNF410	−1.0376	1	1	1	1	1	1
14695	MAP3K7	0	0.50	0.40	0.32	0.35	0.38	0.32
15278	CLDN4	0.0809	0.95	0.96	0.94	0.85	0.92	0.95
15845	GREM1	0	0.05	0.02	0.04	0.15	0.08	0.05
16565	MAPK10	−1.3125	1	1	1	1	1	1
17469	ITM2A	−2.3793	1	1	1	1	1	1
17474	THUMPD1	−0.3287	0.95	0.98	0.99	1	0.98	0.96
17831	ACSS3	−1.7764	1	1	1	1	1	1
17852	IMAGE: 287365	0.1873	0.15	0.20	0.27	0.30	0.24	0.23
18544	PAIP2	−3.1127	1	1	1	1	1	1
18798	CCK	−0.0314	0.15	0.20	0.27	0.30	0.24	0.23
19471	ZNF585B	−0.4962	0.35	0.36	0.44	0.55	0.50	0.52

Table 3:

Comparison (before imputation and after imputation) of the variable selection results using SCAD.

Index	Gene	Before Imputation Parameter Estimation	Imputation Method A Selection Proportion			Imputation Method B Selection Proportion		
			$\hat{\beta}$	$P_{SCAD}^{(20)}$	$P_{SCAD}^{(50)}$	$P_{SCAD}^{(100)}$	$P_{SCAD}^{(20)}$	$P_{SCAD}^{(50)}$
91	IMAGE: 202168	0.1087	1	1	1	1	1	1
435	IMAGE: 470914	0.1969	1	1	1	1	1	0.99
876	ODF2	0.4764	1	1	1	1	1	1
1065	CPS1	0	0.75	0.78	0.76	0.75	0.74	0.72
1079	IMAGE: 416614	0.0167	0.85	0.88	0.83	0.90	0.84	0.82
3137	MME	−0.3245	1	1	1	1	1	1
5071	ST14	0.4184	1	1	1	1	1	1
5381	ZNF337	0	0.25	0.20	0.18	0.05	0.14	0.20
5581	SLC25A6	1.1927	1	1	1	1	1	1
7739	SOX4	0.4495	1	1	1	1	1	1
8446	LGALS1	−0.1551	1	1	1	1	1	0.99
9843	AMACR	0.8870	1	1	1	1	1	1
10137	IMAGE: 199229	0	0.15	0.08	0.11	0.05	0.10	0.11
10276	MYO6	0.0910	1	1	1	1	1	1
10281	MEIS2	−0.2023	0.75	0.80	0.82	0.95	0.86	0.80
10365	ATXN7	0	0.55	0.46	0.35	0.30	0.36	0.35
10532	CCDC93	0.0834	0.95	0.98	0.99	1	1	0.99
11482	MAPRE1	−0.7445	1	1	1	1	1	1
12470	HADHA	−0.0811	0.95	0.98	0.99	1	1	0.99
14003	RPSA	0	0.15	0.08	0.11	0.05	0.10	0.13
14398	ZNF410	0	0.75	0.70	0.64	0.60	0.62	0.63
15278	CLDN4	0.3472	1	1	1	1	1	1
16565	MAPK10	−0.3930	1	1	1	1	1	1
17469	ITM2A	−0.5140	1	1	1	1	1	1
17831	ACSS3	−0.8144	1	1	1	1	1	1
18544	PAIP2	−0.4286	1	1	1	1	1	0.99
18798	CCK	0	0	0	0	0	0	0.02
19673	ANTXR2	−0.3280	1	1	1	1	1	0.98

Table 4:

Comparison (before imputation and after imputation) of the variable selection results using MCP.

Index	Gene	Before Imputation Parameter Estimation	Imputation Method A Selection Proportion			Imputation Method B Selection Proportion		
		$\hat{\beta}$	$P_{\text{MCP}}^{(20)}$	$P_{\text{MCP}}^{(50)}$	$P_{\text{MCP}}^{(100)}$	$P_{\text{MCP}}^{(20)}$	$P_{\text{MCP}}^{(50)}$	$P_{\text{MCP}}^{(100)}$
665	CKS1B	0.9409	0.10	0.12	0.12	0.05	0.10	0.24
1065	CPS1	0.1058	0.05	0.04	0.04	0	0.04	0.14
5010	RHOB	-0.0486	0.65	0.64	0.70	0.65	0.68	0.68
5581	SLC25A6	3.7771	1	1	1	1	1	1
9436	TMED10	-0.8100	0.05	0.04	0.04	0	0.04	0.14
9843	AMACR	1.6849	1	1	1	1	1	1
10281	MEIS2	0	0.20	0.20	0.17	0.30	0.26	0.18
15278	CLDN4	1.6255	1	0.98	0.99	0.95	0.92	0.96
17469	ITM2A	-0.3465	0.30	0.40	0.40	0.25	0.34	0.49
17831	ACSS3	-0.8503	0.80	0.80	0.83	0.70	0.74	0.82
19673	ANTXR2	0	0.95	0.94	0.95	1	0.96	0.86

Table 5:

Simulation: Comparison (ideal, before imputation and after imputation) of the selection probability results for different covariate categories (based on whether the true coefficient is zero and whether the covariate has missing values) using LASSO, SCAD and MCP. (BI=Before Imputation, IMA=Imputation Method A, IMB=Imputation Method B)

Penalty	Covariate	Ideal	BI	IMA			IMB		
				<i>M</i> = 20	<i>M</i> = 50	<i>M</i> = 100	<i>M</i> = 20	<i>M</i> = 50	<i>M</i> = 100
LASSO	nonzero, missing	1	0	0.8945	0.8989	0.8950	0.8925	0.8903	0.8925
	nonzero, observed	1	0.9367	0.9780	0.9785	0.9777	0.9775	0.9759	0.9776
	zero, missing	0	0	0.1033	0.1027	0.1048	0.1160	0.1223	0.1191
	zero, observed	0	0.0080	0.0170	0.0173	0.0171	0.0168	0.0169	0.0168
SCAD	nonzero, missing	1	0	0.7828	0.7890	0.7832	0.7777	0.7766	0.7757
	nonzero, observed	1	0.9133	0.9020	0.9023	0.9033	0.8962	0.8956	0.8954
	zero, missing	0	0	0.0712	0.0712	0.0724	0.0867	0.0894	0.0900
	zero, observed	0	0.0062	0.0114	0.0115	0.0115	0.0111	0.0111	0.0110
MCP	nonzero, missing	1	0	0.6175	0.6223	0.6203	0.6222	0.6201	0.6176
	nonzero, observed	1	0.7300	0.7640	0.7640	0.7653	0.7512	0.7518	0.7531
	zero, missing	0	0	0.0272	0.0282	0.0279	0.0305	0.0290	0.0318
	zero, observed	0	0.0026	0.0033	0.0034	0.0034	0.0032	0.0032	0.0032