5-29-2019

Sure Independence Screening in the Presence of Data That is Missing at Random

Adriano Zanin Zambom
California State University

Gregory J. Matthews
Loyola University Chicago, gmatthews1@luc.edu

Recommended Citation

Zanin Zambom, Adriano and Matthews, Gregory J.. Sure Independence Screening in the Presence of Data That is Missing at Random. Statistical Papers, , : , 2019. Retrieved from Loyola eCommons, Mathematics and Statistics: Faculty Publications and Other Works, http://dx.doi.org/10.1007/s00362-019-01115-w

Noname manuscript No.
(will be inserted by the editor)

Sure Independence Screening in the Presence of Missing Data

Adriano Zanin Zambom · Gregory J. Matthews

Received: date / Accepted: date

Abstract Variable selection in ultra-high dimensional data sets is an increasingly prevalent issue with the readily available data arising from, for example, genome-wide associations studies or gene expression data. When the dimension of the feature space is exponentially larger than the sample size, it is desirable to screen out unimportant predictors in order to bring the dimension down to a moderate scale. In this paper we consider the case when observations of the predictors are missing at random. We propose performing screening using the marginal linear correlation coefficient between each predictor and the response variable accounting for the missing data using maximum likelihood estimation. This method is shown to have the sure screening property. Moreover, a novel method of screening that uses additional predictors when estimating the correlation coefficient is proposed. Simulations show that simply performing screening using pairwise complete observations is out-performed by both the proposed methods and is not recommended. Finally, the proposed methods are applied to a gene expression study on prostate cancer.

Keywords maximum likelihood estimator · correlation coefficient · EM algorithm · missing at random · ultrahigh dimensionality

1 Introduction

With the advances of biotechnologies in sequencing genomes of a wide variety of organisms, new statistical challenges arise and methodological development is essential to locating genes underlying important traits. The selection of genes

A. Z. Zambom
California State University Northridge
E-mail: adriano.zambom@csun.edu

G. J. Matthews
Loyola University Chicago

whose expression levels significantly contribute to the prediction of the probability of a certain disease (cancer for instance) has been the focus of several (interdisciplinary) genome-wide association studies (GWAS), see for example [19], [73], [30], [5], [45], [63], [18], [78], [68] among several more. Including insignificant variables in the model increases the complexity and decreases its interpretability and predictive power. For this reason, high-dimensional data such as those recorded in genetic studies, have inspired pioneer statistical research in developing parsimonious predictive models.

There is a large number of papers in the statistics literature dedicated to variable selection in high-dimensional models, see [32], [77], [58], [37], [44] and references therein. An interesting overview can be found in [27]. In the last two decades, an effective approach based on the minimization of a constrained penalized likelihood has been the focus of several authors, including the Lasso [70], Smoothly Clipped Absolute Deviation penalty (SCAD) [25], Adaptive Lasso [85], Least Angle Regression [20] and the Dantzig Selector [9]. Nevertheless, these methods may fail to correctly identify the significant covariates for ultra-high dimensional models, that is, when the dimension of the feature space is exponentially larger than the sample size. This drawback is observed due to complicated stability of the algorithms, computational burden and statistical accuracy ([28]). Practical examples of ultra-high dimensional data where there is a much larger number of variables than the sample size are found in a variety of cutting-edge research such as biomedical imaging, genomics, tumor classifications, and finance, just to cite a few. In order to alleviate the computational complexity and difficulties in ultra-high dimensional statistical analysis, [26] established theoretical grounds for screening out unimportant predictors in linear models, thereby reducing the dimensionality to a moderate scale. The idea is to rank the importance of each covariate using its estimated marginal linear correlation with the response variable and select a set of covariates with the highest correlation. They showed that with probability tending to 1 exponentially fast, a well-chosen subset of predictors with highest estimated correlations will contain the true set of predictors that significantly contribute to the underlying predictive model, hence the name Sure Independence Screening (SIS). Since then, a number of authors have explored this idea in different areas, see for example [47], [13], [53], [83], and references therein.

In practice, it is not uncommon to come across missing or incomplete data ([38], [35]) in a wide variety of applied statistical settings including cost effectiveness analysis ([29]), education ([11]), spatial data ([41], [4]), AIDS research ([34]), genome-wide association studies ([55,8,57,15,46]), or gene expression studies ([48],[72],[23],[82]). The simplest method for handling missing data is to remove records with missing values. This, however, can lead to biased statistical results unless the missingness mechanism is completely at random (MCAR). More principled methods for addressing the missing data issue include the well-known maximum likelihood estimation (MLE) ([49]), which can be implemented for more complex likelihoods via the EM algorithm [17]. Another common method is multiple imputation ([65]), which fills in the missing

values in the data by drawing randomly from an appropriate distribution to generate multiple completed data sets. Analysis can then proceed using each of the full data sets and results can be combined across imputations to arrive at a final result. Many of these procedures that handle missing data were developed decades ago when high and ultra-high dimensional data were still squarely in the future. In high-dimensional space, multiple imputation is difficult to perform and research in this area is still rudimentary [84,16].

Some methods for variable selection have been developed in the case of missing covariate or response values for regression models with moderate dimension. [40] use the EM algorithm to develop a novel model criteria for covariates missing at random in regression models or longitudinal response variables and covariates. [31] propose a method for selecting variables in the presence of missing data for Cox proportional hazard models. Garcia et al. (2012) propose an adaptation of the SCAD and Lasso penalized approaches and introduced an algorithm to simultaneously optimize the penalized likelihood and estimate the penalty parameters. [12] proposes MI-LASSO, which employs the group LASSO penalty to perform variable selection after multiple imputation. [51] combines multiple imputation with the random Lasso procedure ([75]) and simultaneously performs variable selection. Other methods include [59,39,81,40,14,67,52,79,80] and references therein. However, none of these methods were designed to deal with ultra-high dimensional data.

Therefore, statistical inference with missing data in ultra-high dimensional spaces is an important applied problem that needs to be explored. However, screening high-dimensional models in the presence of missing data is a mostly unexplored area despite its crucial practical importance.

While in recent years some authors have studied the problem of screening with missing data (e.g. [43], [69], [74]), they are all primarily concerned with missing data only in the response variable whereas the proposed method addresses the case of missing covariate values. To the best of our knowledge, there is no method in the literature of statistics that addresses the challenge of screening covariates with missing values in ultra-high dimensional feature spaces. In this paper we propose a screening procedure based on the maximum likelihood estimator of the linear correlation coefficient under MCAR and MAR missing mechanisms. We show the sure screening property of the proposed method under the assumption of marginal bivariate normality of the predictors and the response variable. When the marginal bivariate distribution is not known, the EM algorithm can be used to estimate the correlations. Moreover, we propose a two-stage screening procedure that first uses screening for imputing the missing values of covariates, and then performs screening with the rankings of the covariates according to their correlations with the response. Simulations suggest that the use of the proposed screening procedure yields higher probability of retaining the true significant predictors compared to screening after simply removing records with missing values, especially for covariate spaces with high correlations.

The remainder of the paper is organized as follows. Section 2 discusses the asymptotic properties of the correlation coefficient estimated through maxi-

mum likelihood in the presence of missing data and introduces the two proposed screening methods. In Section 3 we investigate the finite-sample performance of the proposed methods through Monte Carlo simulations with comparisons to existing methods. Finally, a data set is analyzed in Section 4 with the new and existing screening procedures, and the results of the different analyses are compared.

2 Sure Independence Screening

Let Y denote the response variable, $\mathbf{X} = (X_1, \ldots, X_d)$ the vector of available predictors, and with some abuse of notation, let X_{ij} be the i-th observation of the j-th covariate. Assume that n samples from the response Y are observed, however for the data pair (Y, X_j), only n_j observations are complete, that is, $r_j = n - n_j$ observations of $X_j, j = 1, \ldots, d$, are missing. We consider the ultra-high dimensional setting where the dimension d of the predictor space greatly exceeds the sample size n. The usual assumption in high-dimensional analysis is sparsity of the covariate space, that is, only a small number of predictors belong to the true underlying regression model. In this case, variable selection procedures that identify the significant predictors can improve model interpretability with parsimonious representation and greatly increase model accuracy by eliminating irrelevant covariates.

In [26] Sure Independence Screening, the covariates are ranked according to their marginal correlation coefficient with the response. A simple application of SIS would be to disregard the rows of observations whose X values are missing, however this approach neglects the possible information contained in the excluded data. A better method would be to estimate each marginal correlation using all the intrinsic information that can be extracted from the full dataset, through maximum likelihood or an algorithm such as the Expectation Maximization (EM).

2.1 The correlation coefficient when data is missing

Denote the bivariate distribution of the random vector (X_j, Y) by f_{X_jY} with mean vector (μ_j, μ_y) and finite covariance matrix $\Sigma = [\sigma_j^2, \sigma_j\sigma_y\rho_j; \sigma_j\sigma_y\rho_j, \sigma_y^2]$. It is well known that the maximum likelihood estimator of the linear correlation coefficient ρ_j when n_j complete pairs of observations are available, namely $\tilde{\rho}_j = \sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)(Y_i - \bar{Y})/\sqrt{\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)^2 \sum_{k=1}^{n_j}(Y_k - \bar{Y})^2}$, is consistent and has asymptotic distribution ([3], p.122, Theorem 4.2.4)

$$\sqrt{n_j}(\tilde{\rho}_j - \rho_j) \xrightarrow{d} N\left(0, (1 - \rho_j^2)^2\right). \tag{1}$$

Consider the case when the distribution of (X_j, Y) is bivariate Gaussian and denote the density f_{X_jY} by
$N_2(y, x|\mu_j, \mu_y; \sigma_j^2, \sigma_y^2; \rho_j)$. When one of the variables has observations missing

at random, the likelihood estimator of $\boldsymbol{\theta}_j = (\mu_j, \mu_y, \sigma_j^2, \sigma_y^2, \rho_j)$ can actually be derived algebraically. [2] writes the bivariate probability density as the product of the marginal density of Y and the conditional density of X_j given Y, more specifically

$$N_2(y, x|\mu_j, \mu_y; \sigma_j^2, \sigma_y^2; \rho_j) = N(y|\mu_y, \sigma_y^2)N(x|\mu_{j.y} - \beta_{jy}y, \sigma_{j.y}^2),$$

where $\mu_{j.y} = \mu_j - \beta_{jy}\mu_y, \beta_{jy} = \rho_j\sigma_j/\sigma_y$ and $\sigma_{j.y}^2 = \sigma_j^2(1-\rho_j^2)$. In this way, the parameter vector $\boldsymbol{\theta}_j = (\mu_j, \mu_y, \sigma_j^2, \sigma_y^2, \rho_j)$ can be expressed in terms of the parameter vector $\boldsymbol{\phi}_j = (\mu_y, \sigma_y^2, \mu_{j.y}, \beta_{jy}, \sigma_{j.y}^2)$, where $\mu_j = \mu_{j.y} + \beta_{jy}\mu_y, \sigma_j^2 = \beta_{jy}^2\sigma_y^2 + \sigma_{j.y}^2$, and $\rho_j = \beta_{jy}\sigma_y/\sqrt{\beta_{jy}^2\sigma_y^2 + \sigma_{j.y}^2}$. [2] then shows that the log-likelihood
$\ell(\cdot|\boldsymbol{\phi}_j) := -(2\sigma_{j.y}^2)^{-1}\sum_{i=1}^{n_j}(x_{ij} - \mu_{j.y} - \beta_{jy}y_i)^2 - (1/2)(n_j\log(\sigma_{j.y}^2) + n\log(\sigma_y^2)) - (2\sigma_y^2)^{-1}\sum_{i=1}^{n}(y_i - \mu_y)^2$ is maximized at

$$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \hat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}_y)^2 \qquad \hat{\beta}_{jy} = s_{jy}/s_y^2$$
$$\hat{\mu}_{j.y} = \bar{x}_j - \hat{\beta}_{jy}\bar{y} \qquad \hat{\sigma}_{j.y}^2 = s_j^2 - s_{jy}^2/s_y^2,$$

where $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}, \bar{y} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_i, s_y^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_i - \bar{y})^2, s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$, and $s_{jy} = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_i - \bar{y})(x_{ij} - \bar{x}_j)$. Thus, the estimator $\hat{\boldsymbol{\theta}}_j$ of the parameter vector $\boldsymbol{\theta}_j$ can be written according to the estimator $\hat{\boldsymbol{\phi}}_j$, more specifically for the correlation coefficient we have

$$\hat{\rho}_j^{MLE} = \hat{\rho}_j = \tilde{\rho}_j\left(\frac{\hat{\sigma}_y}{s_y}\right)\left(\frac{s_j^2}{\left(s_j^2 - (1 - \hat{\sigma}_y^2/s_y^2)s_{jy}^2/s_y^2\right)}\right)^{1/2} = \frac{s_{jy}}{s_j s_y}\frac{\hat{\sigma}_y}{s_y}\frac{s_j}{\hat{\sigma}_j}, \quad (2)$$

where $\hat{\sigma}_j^2 = s_j^2 - (1 - \hat{\sigma}_y^2/s_y^2)s_{jy}^2/s_y^2$. Note that the estimator $\hat{\rho}_j$ is a weighted version of $\tilde{\rho}_j$, based on an adjustment computed with the ratio $\hat{\sigma}_y^2/s_y^2$, so that when $\hat{\sigma}_y^2 = s_y^2$, i.e. there is no missing data, we have $\hat{\rho}_j = \tilde{\rho}_j$.

To gain insight about the efficiency of the MLE $\hat{\rho}_j$ compared to that of the estimator $\tilde{\rho}_j$, which is based only on the complete pairs of observations, we now look at the asymptotic distribution of $\hat{\rho}_j$. By computing the large sample covariance matrix of $(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)$ through the inverse of the information matrix, we obtain the following theorem.

Theorem 21 *Assume n_j complete pairs of observations from (X_j, Y) and an additional $n - n_j$ univariate observations of Y are available, where the missing observations of X are missing at random (MAR). Assume $(X_j, Y) \sim N_2(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_j, \mu_y)$ and $\Sigma = [\sigma_j^2, \ \rho_j\sigma_j\sigma_y; \rho_j\sigma_j\sigma_y, \ \sigma_y^2]$. The maximum likelihood estimator of the correlation coefficient has, as $n \to \infty$ and $n_j \to \infty$, asymptotic distribution*

$$\left[(1 - \tilde{\rho}_j^2)^2\left(\frac{\hat{\sigma}_y^2}{s_y^2}\right)\left(\frac{1}{nn_j}\right)\left(\frac{\tilde{\rho}_j^2(n_j - n)/2 + n}{\left(\tilde{\rho}_j^2(\frac{\hat{\sigma}_y^2}{s_y^2} - 1) + 1\right)^3}\right)\right]^{-1/2}(\hat{\rho}_j - \rho_j) \xrightarrow{d} N(0, 1).$$

Note that if there is no missing data, i.e. $n = n_j$ and $\hat{\sigma}_y = s_y$, the result in Theorem 21 is equivalent to the convergence in (1).

An alternative method to the maximum likelihood estimator when an explicit solution of the likelihood equation is not feasible is to estimate the correlation coefficient with the EM algorithm. The algorithm starts by filling in the missing observations of the covariate X_j with some initial values, say the average of the X_j observations. Denote this artificial data by $\{(X_{ij}^{(0)}, Y)\}_{i=1}^n$. The M step consists of maximizing the likelihood to find the parameter vector $\hat{\boldsymbol{\theta}}^{(0)}$. The E step then replaces the missing values of the original data with their expected value conditional on $\hat{\boldsymbol{\theta}}^{(0)}$. In the case of a Gaussian distribution for example, the missing value X_{ij} is replaced with

$$\bar{x}_j^{(0)} + \frac{\hat{\sigma}_j^{(0)}}{\hat{\sigma}_y}(y_i - \hat{\mu}_y),$$

where $\bar{x}_j^{(0)}$ and $\hat{\sigma}_j^{(0)}$ are the mean and variance respectively of the artificial data $\{X_{ij}^{(0)}\}_{i=1}^n$. The iteration of steps E and M yields the k-th artificial data set $\{(X_{ij}^{(k)}, Y)\}_{i=1}^n$. The algorithm stops when $||\hat{\theta}_j^{(k)} - \hat{\theta}_j^{(k-1)}||$ is suitably small.

When the number of available covariates is ultra-high, one would like to reduce the dimension of the predictor space by screening out those covariates that are likely to be uncorrelated with the response. In Section 2.2 we establish the sure independence screening property when ranking the covariates according to their estimated correlation $\hat{\rho}_j$ with Y, computed after the imputation of the missing values via maximum likelihood.

To conclude this section, we provide an insight on the advantages and possible disadvantages of using $\hat{\rho}_j$ instead of $\tilde{\rho}_j$ as a ranking utility for screening. We estimate via bootstrap the variance and bias of the correlation estimators $\hat{\rho}$ and $\tilde{\rho}$ of the correlation ρ between X and Y when (X, Y) are bivariate Normal both with variance 1. Figure 1 shows the bootstrap variance and average absolute value of the bias of $\hat{\rho}$ and $\tilde{\rho}$ computed from 1000 simulated datasets $\{X_i, Y_i\}_{i=1}^n$, for $n = 400$, when the true value of ρ ranges from -1 to 1 for the following missing patterns. Let $r_i = I(X_i \text{ is missing})$, where $I()$ is the indicator function. We consider the missing patterns: a) $P(r_i = 1|Y_i) = \exp(2Y_i)/(1 + \exp(2Y_i))$, b) $P(r_i = 1|Y_i < 0) = 0.3$ and $P(r_i = 1|Y_i \geq 0) = 0.7$, c) $P(r_i = 1||Y_i| < 1) = 0.7$ and $P(r_i = 1||Y_i| \geq 1) = 0.3$, and d) $P(r_i = 1||Y_i| < 1) = 0.7$ and $P(r_i = 1||Y_i| < 1) = 0.3$.

Note that the variance of $\hat{\rho}$ is slightly lower for high (in absolute value) values of ρ but higher for small (in absolute value) values of ρ for missing patterns a) b) and d), however the opposite is observed in case c). As expected, when X and Y are uncorrelated, the variance of $\hat{\rho}$ is larger than that of $\tilde{\rho}$ since the complete pairs of observations may, by chance, contain spurious correlation, inflating (or deflating) the estimator $\hat{\rho}$. Interestingly, when most of the missing data happens within the 2nd and 3rd quartiles of Y, that is case c), the variance of $\hat{\rho}$ is smaller than that of $\tilde{\rho}$, which may be due to the little effect these missing values in the center of the data have on the correlation

estimator. The bias of $\hat{\rho}$ is always smaller than that of $\tilde{\rho}$ for large values (in absolute value) of ρ, however it tends to be larger when X and Y have weak correlation.

For the screening procedure this means that, when the covariate X_j is uncorrelated with Y, the larger bias and variance of $\hat{\rho}_j$ compared to that of $\tilde{\rho}_j$ may cause the ranking assigned to X_j to be less accurate more often. Given the assumption of sparsity of the model, this will likely cause a disarrangement of many of the rankings assigned to the uncorrelated covariates. This effect should also be observed with the use of $\tilde{\rho}_j$ as ranking utility, yet with a smaller proportion of disarrangement given it has a slightly smaller bias when $\rho_j \approx 0$. On the other hand, the MLE $\hat{\rho}_j$ is a more precise estimator (smaller bias and variance) of the correlation coefficient ρ_j when $|\rho_j|$ is large, so that the high rankings of significant covariates will be more accurately maintained.

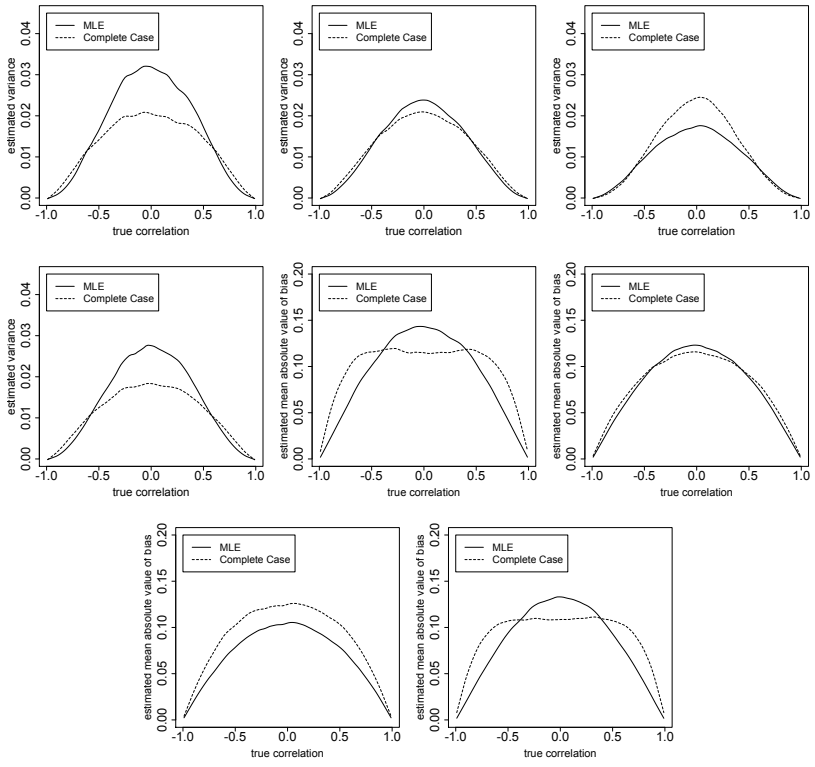

Fig. 1 Bootstrap variance (top row) and absolute value of bias (bottom row) of $\hat{\rho}^{MLE}$ and $\tilde{\rho}$ (complete pairs only) from 1000 simulation runs for the missing patterns: from left to right a) $P(r_i = 1|Y_i) = \exp(2Y_i)/(1 + \exp(2Y_i))$, b) $P(r_i = 1|Y_i < 0) = 0.3$ and $P(r_i = 1|Y_i \geq 0) = 0.7$, c) $P(r_i = 1||Y_i| < 1) = 0.7$ and $P(r_i = 1||Y_i| \geq 1) = 0.3$, and d) $P(r_i = 1||Y_i| < 1) = 0.7$ and $P(r_i = 1||Y_i| < 1) = 0.3$.

2.2 Sure Screening Method

Consider the maximum likelihood estimator $\hat{\rho}_j$ of ρ_j defined in (2). Let $\mathcal{A} = \{1 \leq j \leq p : \beta_j \neq 0\}$ be the active set of predictors that compose the underlying true sparse model. The proposed method aims at identifying a set $\hat{\mathcal{A}} = \{j : |\hat{\rho}_j| \geq cn^{-\kappa}\}$, where c and κ are pre-specified threshold values defined below, such that it contains with high probability the true set $\mathcal{A}$.

For theoretical proofs of the sure screening property, assume the following conditions:

(C1) for all small positive constants m, $\sup_d \max_{j=1,\ldots,d} E[\exp(m(Y_i - Y_k)(X_{ij} - X_{kj}))] < \infty$,

(C2) $\min_{j\in\mathcal{A}} |\rho_j| \geq 2cn^{-\kappa}$, for some constants $c > 0$ and $0 \leq \kappa < 1/2$.

Conditions C1 and C2 are similar to those considered in [26], [46], and [24]. In fact, condition C1 holds when X and Y are uniformly bounded or have a multivariate Normal distribution. Condition C2 assumes that the signal strength, measured by the true correlation coefficient ρ_j, is not too weak, and can be detected by the proposed approach. Theorem 22 shows that the estimated set $\hat{\mathcal{A}}$ contains the true set $\mathcal{A}$ with probability increasing to 1 exponentially fast as the sample size increases.

Theorem 22 *Under condition C1, for any $0 < \gamma < 1/2 - \kappa$, there exists constants $c_1 > 0$ and $c_2 > 0$ such that*

$$P(|\hat{\rho}_j - \rho_j| \geq cn_j^{-\kappa}, \text{ for all } j) \leq \sum_{j=1}^{d} O([\exp(-c_1 n_j^{1-2(\gamma+\kappa)}) + n_j \exp(-c_2 n_j^{\gamma})]), \text{ and}$$

$$P(\max_{j=1,\ldots,d} |\hat{\rho}_j - \rho_j| \geq cn_j^{-\kappa}) \leq O(d\exp(-c_1 \min_j n_j^{1-2(\gamma+\kappa)}) + \max_j\{n_j \exp(-c_2 n_j^{\gamma})\}).$$

Consequently, if C2 is also true, then

$$P(\mathcal{A} \subseteq \hat{\mathcal{A}}) \geq 1 - \sum_{j\in\mathcal{A}} O(\exp(-c_1 \min_j n_j^{1-2(\gamma+\kappa)}) + \max_j\{n_j \exp(-c_2 n_j^{\gamma})\}).$$

The first part of Theorem 22 establishes the exponential convergence rate in probability of the estimators $\hat{\rho}_j$ to the true parameters $\rho_j, j = 1, \ldots, d$. The probability rate of their maximum absolute difference is based on the minimum and maximum number of complete pairs of observations between the response Y and the d covariates. Consequently, the second part of the theorem shows that the estimated reduced set $\hat{\mathcal{A}}$ contains all covariates in the true model, defined by $\mathcal{A}$, with overwhelming probability, that is, the proposed screening method posesses the sure screening property. Note that the total number of covariates d, as well as the number of significant covariates in the model $|\mathcal{A}|$ (size of $\mathcal{A}$), are allowed to increase with the sample size n.

The individual convergence rates of $\hat{\rho}_j$ are proportional to the rates of $\tilde{\rho}_j$, namely, their order is based on n_j, the number of complete of observations of

each (X_j, Y) pair. However, the convergence rate stated in Theorem 22 contains information about the missing data extracted from the joint distribution of X_j and Y, since the missing values were imputed with maximum likelihood estimation. Such information is not taken into consideration when using $\tilde{\rho}_j$. Hence, although asymptotically equivalent, for small samples sizes the proportion of information gained with the use of $\hat{\rho}_j$ may increase the probability that $\hat{\mathcal{A}}$ contains $\mathcal{A}$.

Remark 1 Variable screening procedures may fail to detect significant covariates that are marginally uncorrelated but jointly correlated with the response. Another drawback of screening methods is the identification of spurious correlations, that is, when the procedure selects covariates that are uncorrelated with the response but are correlated with significant covariates of the model. An iterative version of the screening method proposed in this section, similar to the Iterative SIS, Iterative DCSIS and INIS in [26], [46], and [24] respectively, can be used to reduce the impact of these issues.

2.3 Two-Stage Screening

The objective of using the MLE $\hat{\rho}_j$, instead of $\tilde{\rho}_j$, for the computation of the set $\hat{\mathcal{A}}$, is to improve the accuracy of the ranking. However, $\hat{\rho}_j$ is calculated using solely the pairwise data from (X_j, Y), not taking into consideration the information about the missing values of X_j, which is possibly also available in the remainder of the covariates $X_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d)$. This is specially important in the case of highly correlated covariates. In fact, a more precise estimation of the missing values of X_j can be obtained by maximizing the likelihood function (or using the EM algorithm) of the entire set of available variables $(Y, \mathbf{X})$, which in consequence can yield a more precise estimation of $\rho_j, j = 1, \ldots, d$.

Although this approach would produce the most accurate values possible for the missing data based on the available observations, it is an extreme solution to variable screening due to its very high computational complexity. Using all variables for imputation would require the estimation of $\frac{d(d-1)}{2} + d$ parameters in the covariance matrix alone, which for large values of d quickly becomes intractable (e.g. for $d = 1000$, the number of parameters in over 500000, and for $d = 1000000$, the number of parameters is over 500 billion).

To solve this problem, we propose a pre-screening step to select a set of variables $\mathcal{B}_j$ which should be included in the likelihood for the estimation of the correlation between X_j and Y. Let η_j denote the size of the set $\mathcal{B}_j$. When considering how large η_j should be, two separate interests need to be balanced. By choosing η_j too small, one potentially misses out on useful information about the missing values of X_j contained in other variables that may be highly correlated with X_j that will not be included in $\mathcal{B}_j$. Choosing η_j too large, however, can lead to complicated high dimensional likelihoods that may be intractable. We, therefore, seek a value of η_j that is large enough to contain most useful information from other variables, while still being small enough to

be computationally feasible. A simplistic solution is to pick a fixed moderate value of η_j (e.g. 10 or 20) for all j. However, this can lead to biased imputation when few (or even none) of the variables are correlated with the target variable X_j. If all covariates were completely independent, by fixing η_j to be 10, for instance, one would include the ten covariates that happen to be the most highly correlated, but only by chance, with X_j. A better solution is to let the data dictate the value of η_j. We propose choosing η_j based on the [7] false discovery rate (FDR) correction on the hypothesis tests for the significance of the correlation coefficient between X_j and $X_k, k = 1, \ldots, j-1, j+1, \ldots, d$. For computational purposes, let the maximum value of η_j be M_η, which can be useful when many covariates are highly correlated and the FDR method would select computationally infeasible sizes η_j. In simulation studies we use $M_\eta = 10$.

The proposed method, which we call Two-Stage screening, defines the set $\hat{\mathcal{A}} = \{j : |\hat{\rho}_j^{\mathcal{B}}| \geq cn^{-\kappa}\}$, where $\hat{\rho}_j^{\mathcal{B}}$ is computed as follows.

1. Compute the maximum likelihood estimator $\hat{\rho}_{jk}, k = 1, \ldots, j-1, j+1, \ldots, d$, of the correlation coefficient ρ_{jk} between (X_j, X_k) from maximizing the likelihood $L(\cdot|X_j, X_k, Y)$.

2. Let $\pi_{jk} = 1 - F_{t_{n-2}}\left(\left|\frac{\hat{\rho}_{jk}}{\sqrt{(1-\hat{\rho}_{jk}/(n-2))}}\right|\right), k = 1, \ldots, j-1, j+1, \ldots, d$ be the p-value for the test $H_0 : \rho_{jk} = 0$, where $F_{t_{n-2}}(\cdot)$ is the cumulative distribution function of the t-student distribution with $n-2$ degrees of freedom.

3. Let $\gamma_j = \max\left\{\ell : \pi_{j(\ell)} \leq \frac{\ell}{d-1}\frac{\alpha}{\sum_{j=1}^{d-1} j^{-1}}\right\}$, where α is the desired level of the test, and $\pi_{j(1)}, \ldots, \pi_{j(d-1)}$ denote the ordered p-values from step 2.

4. Let $\eta_j = \min\{\gamma_j, M_\eta\}$.

5. Compute $\mathcal{B}_j = \{k : |\hat{\rho}_{jk}| \geq |\hat{\rho}_{jk}|_{(d-\eta_j)}\}$ where $|\hat{\rho}_{jk}|_{(q)}$ is the q-th order statistic of $|\hat{\rho}_{j1}|, \cdots, |\hat{\rho}_{j(j-1)}|, |\hat{\rho}_{j(j+1)}|, \cdots, |\hat{\rho}_{jd}|$.

6. Finally compute $\hat{\rho}_j^{\mathcal{B}}$ as the maximum likelihood estimator of the correlation between Y and X_j from maximizing the likelihood $L(\cdot|X_j, X_{\mathcal{B}_j}, Y)$, where $X_{\mathcal{B}_j} = \{X_\ell : \ell \in \mathcal{B}_j\}$.

The response variable Y is included in the likelihood for the estimation of each ρ_{jk} because we assume Y is fully observed and the missingness mechanism can only depend on Y. Therefore, excluding Y from the likelihood in this setting could lead to great loss of information, bias and, more importantly, a violation of the MAR assumption. Obviously, the Two-Stage screening is more computationally expensive than simply using the m.l.e. $\hat{\rho}_j$ from the pairwise likelihood. However, the number of parameters (mean vectors and variance matrices) to be estimated in the Two-Stage screening method is significantly reduced from computing all parameters in the likelihood of the full set of predictors.

3 Simulation Study

In this section we analyze the finite sample properties of the two proposed screening methods. The first uses the MLE $\hat{\rho}_j$ defined in (2) as ranking utility and will from here on be called Maximum Likelihood Sure Independence Screening (ML-SIS). The second method is the Two-Stage screening described in Section 2.3, which will be denoted as TS-SIS. For comparison purposes, we also include the results of the screening procedure that uses $\tilde{\rho}_j$ as ranking utility, which is based only on the complete pairs of observations and is denoted by CP-SIS (Complete Pairs-SIS), and the ideal screening that uses the full set of n observations before missingness is imposed to the data, which is denoted by FULL-SIS.

In this simulation study we generate the data from the linear model $Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ with $\beta_1 = \beta_2 = \beta_{22} = 2, \beta_{12} = 3$ and $\beta_j = 0$ for $j \notin \{1, 2, 12, 22\}$. The predictor variables $\mathbf{X}$ are generated from a multivariate Normal distribution with zero mean and covariance matrix $\Sigma = (\sigma_{k\ell})_{d\times d}$ for three different covariance structures

1. $\sigma_{k\ell} = 0.5^{|k-\ell|}$ for all k, ℓ (medium exponential decay),
2. $\sigma_{k\ell} = 0.9^{|k-\ell|}$ for all k, ℓ (large exponential decay),
3. $\sigma_{k\ell} = \begin{cases} 1 & \text{if } k = \ell \\ 0.6 & \text{if } k \neq \ell \end{cases}$ (full matrix of medium correlation).

We fix the sample size n to be 100 and the number of variables $d = 1000$. After the generation of the dataset, the following four missing data mechanisms are imposed. Let q_γ be the empirical γ-th percentile of the observations $\{y_i\}_{i=1}^n$ and let $r_{ij} = I(X_{ij}$ is missing) for $i = 1, \ldots, n$ and $j = 1, \ldots, d$, and consider

- MAR0 (MCAR): $P(r_{ij} = 1) = 0.25$, that is, X_{ij} is missing completely at random (MCAR).
- MAR1: $P(r_{ij} = 1|Y_i \leq 0) = 0.65$ and $P(r_{ij} = 1|Y_i > 0) = 0.35$.
- MAR2: $P(r_{ij} = 1) = \text{expit}(0.15Y_i) = \frac{e^{0.15Y_i}}{1+e^{0.15Y_i}}$.
- MAR3: $P(r_{ij} = 1) = \text{expit}(-1 + 0.2|Y_i - q_{0.50}|) = \frac{e^{-1+0.2|Y_i-q_{0.50}|}}{1+e^{-1+0.2|Y_i-q_{0.50}|}}$.

Figure 2 illustrates the four missingness scenarios considered.

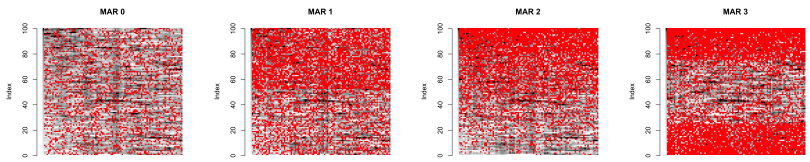

Fig. 2 Visual illustration of the missingness mechanisms considered in the simulation study. Columns represent the $d = 1000$ covariates: red is missing and black is observed.

Note that in this simulation we consider for simplicity the case where the missing pattern of each predictor depends on the fully observed response variable, however the missing pattern could depend on other fully observed variables. We repeat each experiment 500 times, and evaluate the performance of the screening methods by computing the minimal model size $\mathcal{S}$ to include all active predictors.

Tables 1 and 2 show the results for the exponential decay covariance structure and $\sigma_\epsilon = 4$ and $\sigma_\epsilon = 8$ respectively. As expected, the ideal situation where no observations are missing (FULL-SIS) yields the smallest model sizes, while the screening methods applied to datasets with missing observations (TS-SIS, ML-SIS and CP-SIS) require a larger model size to include all active predictors.

MAR	Method	$\sigma_{k\ell} = 0.5^{\lvert k-\ell \rvert}$				$\sigma_{k\ell} = 0.9^{\lvert k-\ell \rvert}$			
		50%	75%	90%	95%	50%	75%	90%	95%
MAR0	FULL-SIS	7.00	22.20	71.30	122.60	20.00	22.00	22.10	23.00
	TS-SIS	16.00	52.20	153.20	262.10	20.00	22.00	23.00	24.00
	ML-SIS	17.50	53.20	168.40	298.00	20.00	23.00	27.00	36.00
	CP-SIS	17.00	52.00	170.00	295.50	20.00	23.00	27.00	36.00
MAR1	FULL-SIS	7.00	22.20	71.30	122.60	20.00	22.00	22.10	23.00
	TS-SIS	75.50	213.50	534.10	723.50	20.00	23.00	27.10	34.00
	ML-SIS	79.00	219.80	526.90	722.20	22.00	34.00	73.00	130.30
	CP-SIS	77.00	235.20	535.20	734.20	22.00	33.20	83.20	128.00
MAR2	FULL-SIS	7.00	22.20	71.30	122.60	20.00	22.00	22.10	23.00
	TS-SIS	107.50	235.20	489.80	641.80	20.00	24.00	35.10	54.10
	ML-SIS	108.00	236.20	491.50	640.30	23.00	38.00	100.20	217.10
	CP-SIS	110.00	248.20	517.10	653.10	23.00	40.20	106.00	226.10
MAR3	FULL-SIS	7.00	22.20	71.30	122.60	20.00	22.00	22.10	23.00
	TS-SIS	190.50	419.20	676.10	828.20	23.00	33.20	75.20	124.40
	ML-SIS	191.50	424.00	666.50	808.90	50.00	112.20	256.70	423.10
	CP-SIS	200.00	425.50	677.00	828.70	50.00	121.20	277.90	465.00

Table 1 Percentiles of the minimum model size $\mathcal{S}$ out of 500 replications for exponential decay covariance structure of predictors and $\sigma_\epsilon = 4$.

It is interesting to note that most percentiles of model sizes in the case $\sigma_{k\ell} = 0.5^{|k-\ell|}$ are larger than those when $\sigma_{k\ell} = 0.9^{|k-\ell|}$. This may be due to the fact that there is a chance any of the d covariates can randomly become correlated with the response variable when observations go missing. On the other hand, when there is high correlation between covariates, the effective dimension of the covariates is less then d, so that less random spurious correlation will appear when missingness is imposed.

When covariates have relatively low correlation ($\sigma_{k\ell} = 0.5^{|k-\ell|}$), all methods have similar performance for all missingness patterns considered. When $\sigma_\epsilon = 4$, FULL-SIS requires an average set size of 71.3 to capture all of the true covariates 90% of the time. For other methods the average 90-th percentile of $\mathcal{S}$ is approximately 163.8, 531.2, 499.4, and 673.2 for MAR0, MAR1, MAR2, and MAR3 respectively. When $\sigma_\epsilon = 8$, much larger model sizes are needed to

MAR	Method	$\sigma_{k\ell} = 0.5^{\lvert k-\ell \rvert}$				$\sigma_{k\ell} = 0.9^{\lvert k-\ell \rvert}$			
		50%	75%	90%	95%	50%	75%	90%	95%
MAR0	FULL-SIS	84.00	230.00	481.80	695.20	21.00	28.00	55.10	91.00
	TS-SIS	155.00	359.20	671.60	848.00	22.00	30.00	64.60	119.10
	ML-SIS	163.00	376.20	639.10	875.00	25.50	47.20	128.30	223.10
	CP-SIS	165.50	378.50	655.40	871.10	25.00	46.00	130.30	211.00
MAR1	FULL-SIS	84.00	230.00	481.80	695.20	21.00	28.00	55.10	91.00
	TS-SIS	331.00	609.00	834.50	930.00	25.00	48.00	112.50	211.40
	ML-SIS	328.50	608.20	844.20	934.00	60.50	150.00	342.30	474.10
	CP-SIS	334.50	600.20	848.00	937.10	54.50	148.50	359.00	484.30
MAR2	FULL-SIS	84.00	230.00	481.80	695.20	21.00	28.00	55.10	91.00
	TS-SIS	485.00	708.00	867.00	935.00	40.00	108.20	278.40	447.20
	ML-SIS	473.00	715.80	869.00	934.00	107.00	262.20	481.30	721.30
	CP-SIS	482.00	712.50	880.00	934.00	102.00	266.20	492.00	722.00
MAR3	FULL-SIS	84.00	228.20	472.90	695.20	21.00	28.00	55.10	91.00
	TS-SIS	669.50	864.50	964.00	983.00	215.00	489.00	742.10	869.10
	ML-SIS	670.50	862.50	963.00	980.10	403.50	719.20	881.00	937.30
	CP-SIS	678.50	870.00	962.20	982.00	405.50	733.00	887.10	944.00

Table 2 Percentiles of the minimum model size $\mathcal{S}$ out of 500 replications for exponential decay covariance structure of predictors and $\sigma_\epsilon = 8$.

capture all the true predictors 90% of the time, from 481.8 for FULL-SIS to as large as 964 for the other methods.

When there are some covariates that are highly correlated to each other ($\sigma_{k\ell} = 0.9^{|k-\ell|}$) and $\sigma_\epsilon = 4$, all methods have similar performance for MAR0, however, the advantage of TS-SIS over the other procedures becomes clear for MAR1, MAR2, and MAR3. Specifically for MAR1 and MAR2, TS-SIS maintains all percentiles very close to that of FULL-SIS, and for MAR3 the 95th percentile is just above the sample size 100. On the other hand, both CP-SIS and ML-SIS require much larger set sizes to include all significant predictors in comparison to TS-SIS, reaching up to 4 times TS-SIS set size for high percentiles.

For the case when the residuals have large variance, that is $\sigma_\epsilon = 8$, the set sizes required by all methods become much larger. When there is low correlation between covariates, similar results to those with $\sigma_\epsilon = 4$ are obtained. For large exponential decay correlation, TS-SIS maintains set sizes much smaller than those of the other methods in all cases, with set sizes about the order of the sample size $n = 100$ up to the 95th, 90th, and 75th percentiles for MAR0, MAR1, and MAR2 respectively. However, for CP-SIS and ML-SIS the quantiles near the sample size correspond to the 90th, 75th, and 50th for MAR0, MAR1, and MAR2 respectively. MAR3 poses a challenge for all methods dealing with the missing data given the large variance of the errors, with TS-SIS having again much lower set sizes especially below the 75th percentile.

Table 3 shows the results for covariance structure 3., full matrix of medium correlation, with $\sigma_\epsilon = 1$ and $\sigma_\epsilon = 4$, which poses a challenge to all methods including FULL-SIS. This constant correlation between all covariates inflates $\mathcal{S}$ because of the numerous spurious correlations introduced in the data, which

causes serious disarrangement of the screening rankings. In all missing data scenarios considered, ML-SIS outperformed CP-SIS. Under MCAR and MAR3, TS-SIS has set sizes only slightly larger than those of ML-SIS, and hence smaller than those of CP-SIS. However, under MAR1 and MAR2, TS-SIS requires larger sets than CP-SIS to include all active predictors. This is probably due to the fact that, when observations go missing more at one end of the spectrum (see Figure 2 for MAR1 and MAR2), several spurious correlations with each gene are created with possible inaccurate relationships. This may cause the first step of the TS-SIS method to select covariates that do not have useful information about the missing values of the gene to be imputed. Nevertheless, the performance of TS-SIS could be improved with an increase in the number of covariates allowed for the first step in the algorithm, namely M_η, however this increases the computational burden and was not investigated.

MAR	Method	$\sigma_\epsilon = 1$				$\sigma_\epsilon = 4$			
		50%	75%	90%	95%	50%	75%	90%	95%
MAR0	FULL-SIS	16.00	48.00	116.30	217.00	26.50	78.20	167.10	237.10
	TS-SIS	50.00	122.50	245.10	356.00	74.50	185.00	305.00	425.00
	ML-SIS	41.00	103.20	216.10	338.00	66.50	161.20	290.00	366.60
	CP-SIS	59.00	133.00	254.10	393.20	89.00	186.50	338.30	455.40
MAR1	FULL-SIS	16.00	48.00	116.30	217.00	26.50	78.20	167.10	237.10
	TS-SIS	222.00	387.20	604.30	719.50	248.50	432.00	655.40	736.10
	ML-SIS	115.00	217.50	348.50	471.10	149.00	260.20	460.50	590.70
	CP-SIS	160.00	323.00	473.00	585.00	192.00	360.20	540.40	665.50
MAR2	FULL-SIS	16.00	48.00	116.30	217.00	26.50	78.20	167.10	237.10
	TS-SIS	218.00	407.00	634.20	770.10	257.00	470.00	692.20	802.00
	ML-SIS	129.00	275.50	441.20	587.80	173.00	348.20	510.00	648.20
	CP-SIS	174.00	313.20	547.30	677.00	223.00	394.50	582.00	715.10
MAR3	FULL-SIS	16.00	48.00	116.30	217.00	26.50	78.20	167.10	237.10
	TS-SIS	227.00	426.00	611.40	715.30	296.00	502.20	696.70	796.10
	ML-SIS	226.50	402.00	613.30	709.10	276.00	461.00	663.10	778.40
	CP-SIS	297.50	477.00	675.10	767.10	334.50	541.00	719.50	830.20

Table 3 Percentiles of the minimum model size $\mathcal{S}$ out of 500 replications for full matrix of medium correlation.

Simulations not reported here show that in a scenario where $P(r_{ij} = 1|Y_i \leq q_{.25}) = P(r_{ij} = 1|Y_i > q_{.75}) = 0$, and $P(r_{ij} = 1|q_{.25} < Y_i \leq q_{.75}) = p \sim beta(50, 50)$, that is, when data is missing in the middle of Y (opposite of MAR3), all screening methods have similar performance. This is likely due to the fact that missing observations in the center of the data do not cause much loss of information.

For additional insight on the properties of the screening methods, we compute $\mathcal{P}$, the probability that all significant predictors are selected for a user-specified model size for the exponential decay correlation cases. Table 4 shows the results for model sizes $d_1 = n - 1$, $d_2 = 2n/log(n)$, and $d_3 = nlog(n)$. When the predictor variables have low correlation, unsurprisingly, there is little difference in the probability of capturing all the relevant variables between

CP-SIS, ML-SIS, and TS-SIS. However, there are some covariates with high correlation, TS-SIS uniformly outperforms both CP-SIS and ML-SIS, with gain in probability reaching up to 29%. Tables not included here for the sake of space show that the proposed ML-SIS outperforms TS-SIS and CP-SIS for the case of a full matrix of constant correlation, as can be expected from the conclusions based on Table 3.

			$\sigma_\epsilon = 4$				σ_ϵ=8			
Size	$\sigma_{k\ell}$	MAR	FULL-SIS	TS-SIS	ML-SIS	CP-SIS	FULL-SIS	TS-SIS	ML-SIS	CP-SIS
d_1	$0.5^{\lvert k-\ell\rvert}$	MAR0	0.93	0.85	0.85	0.84	0.54	0.40	0.39	0.39
		MAR1	0.93	0.56	0.55	0.56	0.54	0.15	0.14	0.15
		MAR2	0.93	0.49	0.49	0.47	0.54	0.08	0.07	0.08
		MAR3	0.93	0.31	0.32	0.31	0.54	0.01	0.01	0.01
	$0.9^{\lvert k-\ell\rvert}$	MAR0	1.00	1.00	0.99	0.99	0.96	0.94	0.85	0.86
		MAR1	1.00	0.99	0.93	0.93	0.96	0.89	0.65	0.67
		MAR2	1.00	0.98	0.90	0.89	0.96	0.73	0.49	0.49
		MAR3	1.00	0.93	0.72	0.70	0.96	0.25	0.09	0.09
d_2	$0.5^{\lvert k-\ell\rvert}$	MAR0	0.74	0.57	0.55	0.56	0.20	0.08	0.05	0.06
		MAR1	0.74	0.20	0.20	0.22	0.20	0.01	0.01	0.01
		MAR2	0.74	0.16	0.15	0.16	0.20	0.00	0.00	0.00
		MAR3	0.74	0.05	0.04	0.04	0.20	0.00	0.00	0.00
	$0.9^{\lvert k-\ell\rvert}$	MAR0	0.70	0.71	0.65	0.65	0.51	0.46	0.36	0.36
		MAR1	0.70	0.64	0.45	0.49	0.51	0.38	0.15	0.16
		MAR2	0.70	0.60	0.41	0.41	0.51	0.24	0.08	0.08
		MAR3	0.70	0.43	0.17	0.19	0.52	0.03	0.01	0.01
d_3	$0.5^{\lvert k-\ell\rvert}$	MAR0	0.85	0.72	0.71	0.71	0.36	0.20	0.19	0.18
		MAR1	0.85	0.34	0.34	0.35	0.36	0.04	0.04	0.05
		MAR2	0.85	0.27	0.27	0.29	0.36	0.02	0.02	0.02
		MAR3	0.85	0.12	0.12	0.13	0.36	0.00	0.00	0.00
	$0.9^{\lvert k-\ell\rvert}$	MAR0	0.99	0.99	0.97	0.96	0.86	0.84	0.73	0.73
		MAR1	0.99	0.97	0.82	0.81	0.86	0.72	0.39	0.41
		MAR2	0.99	0.93	0.78	0.76	0.86	0.53	0.27	0.27
		MAR3	0.99	0.82	0.44	0.46	0.86	0.11	0.02	0.02

Table 4 The proportion $\mathcal{P}$ that all significant predictors are selected for user-specified model sizes $d_1 = n - 1$, $d_2 = n/log(n)$, and $d_3 = 2n/log(n)$.

4 Real data application

In this section we present an application of the proposed method to a real data set. The data in this example comes from the gene expression study on prostate cancer of [71]. The data consists of 104 observations of cDNA microarray data across 20,000 gene locations. Of these 20,000 genes, 18,106 contain at least 1 missing value and 2,570 variables are missing more than half of their observations, making this data set an ideal example with p >> n and a substantial amount of missing data.

As a preliminary step, prior to any application of screening, genes were excluded from the analysis if they were missing 90 or more of the 104 ob-

servations, since imputation of 90% of data missing is unreliable. Hence the number of remaining genes in the dataset is 19,363. We perform four different analysis using BRCA1, BRCA2 ([62,10]), HOXB13 ([22,56,6,61]), and STAT3 ([1,60]) as response variables in each analysis, which have all been previously associated with prostate cancer.

These four genes have different numbers of missing observations. Specifically, all 104 values are observed in STAT3, whereas BRCA1, BRCA2, and HOXB13 have 81, 77, and 102 values observed, respectively. Thus, only rows with an observed target variable are considered, i.e. only 81 observations are used in the BRCA1 analysis for instance. For each target gene as response variable, screening is performed to find the set of genes with highest predictive power among all available genes in the data using TS-SIS, ML-SIS, and CP-SIS. We examine the results for a fixed set size of $\frac{n}{log(n)}$.

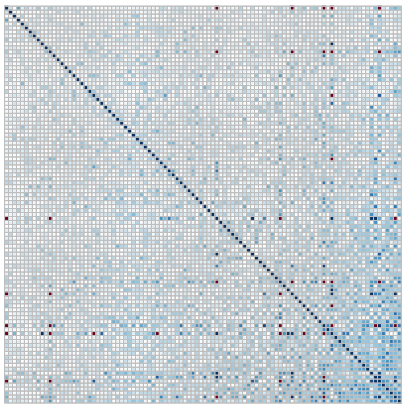

Fig. 3 Correlation plot of 100 randomly chosen genes arranged using their first principal component: white means 0 correlation, dark blur means high correlation in absolute value and red means not enough data to compute correlation.

Figure 3 shows the correlation plot of 100 randomly chosen genes arranged using first principal component order. It can be seen that each gene has high correlation with only a few other genes. This structure suggests that the proposed Two-Stage screening method should be used since, as seen in simulations, it yields the highest probability of retaining the important predictors. It is also prudent to evaluate the top predictors ranked by ML-SIS, as in some cases medium-to-high correlation is observed among several genes. Since the true set of important predictor genes for each case is unknown, we compare the predictor genes top ranked by each method.

The top 22 genes ranked for STAT3 as the response gene are displayed in Table 5. There are 7 genes that are in the top ranked in all three screening methods, namely IMAGE.35807, CCNG1, NFIX.2, SSC.23, ADRA2A, CCSER2.2, and IMAGE.1377071. Five genes are in the top ranked by CP-

SIS only (SSC.13, IMAGE.46647, RPL10.1, CNBP.2, and ANAPC16) and four genes are identified by TS-SIS only (DPYSL2.1, ST8SIA2, PXDN, and PTCH1.1). There are no genes that are uniquely ranked as top 22 by the ML-SIS method. Note that TS-SIS and ML-SIS select genes ST8SIA2 and PTCH1, which are not selected by CP-SIS. ST8SIA2 is potentially related to the expression of polysialic acid (PSA) ([21]), which has been found to be associated with neuroendocrine tumor progression. Additionally, PTCH1.1, which acts as a receptor for sonic hedgehog, has been found to be related to certain types of tumors ([76]).

CP-SIS	ML-SIS	TS-SIS
IMAGE.35807	CNG1	CCNG1
CNTRL.A20	SSC.23	SSC.23
GCC1	IMAGE.35807	CNTRL.CASP14
CCNG1	AP3D1	SSC.83
RAB11B	CNTRL.A20	EGFR.1
NFIX.2	CNTRL.CASP14	ARNTL
GKAP1	SPPL2B	IMAGE.35807
SSC.13	SSC.83	SLC9A9.1
SSC.23	EGFR.1	ADRA2A
IMAGE.46647	GCC1	IMAGE.1377071
SPPL2B	ARNTL	EMX1]
CNTRL.IRAK2	CNTRL.IRAK2	IMAGE.125665
NBL1	GKAP1	RAB11B
ADRA2A	ADRA2A	SSC.2
IMAGE.839579	EMX1	NFIX.2
RPL10.1	IMAGE.1377071	DPYSL2.1
IMAGE.795840	SLC9A9.1	CCSER2.2
CCSER2.2	SSC.2	ST8SIA2
CNBP.2	IMAGE.125665	PXDN
ANAPC16	IMAGE.795840	PTCH1.1
IMAGE.1377071	NFIX.2	IMAGE.839579
AP3D1	CCSER2.2	NBL1

Table 5 Top genes selected by each of the three screening methods for STAT3 as the response variable.

The 18 genes top ranked with BRCA1 as the response variable are shown in Table 6. Of these, 11 of them are found using all three methods. The sets chosen by ML-SIS and TS-SIS are exactly the same and yield seven genes that are not selected by CP-SIS, namely GRIA2, SMYD2, ZNF599, SSR1, LRRC8A, HDGFRP3, and RNF208. SMYD2 is notable as it has previously been found to be related to gastric cancer ([42]), and LRRC8A has been found to be associated with volume-regulated channels for anions (VRAC) which have reduced levels in drug resistant cancer cells ([33]).

With 77 observed values of BRCA2, $\frac{n}{log(n)}$ yields a set size of 17. The result is found in Table 7. The 5 genes that are ranked top 17 by both ML-SIS and TS-SIS but not CP-SIS are FTO, IMAGE.139490, TECRL, RAB2A.1, and SSC.44. CNTRL.RIP is chosen by ML-SIS only and SLC6A15.2 is chosen by TS-SIS only. Two genes that are identified by ML-SIS and TS-SIS that are of

CP-SIS	ML-SIS	TS-SIS
IMAGE.289742	NUDT1	NUDT1
CNTRL.IRAK2	GRIA2	GRIA2
FUT8	IMAGE.289742	IMAGE.289742
COL6A2.1	SLC1A6	SLC1A6
ZNF92	CNTRL.IRAK2	CNTRL.IRAK2
HRSP12	SMYD2	SMYD2
SOX17	ZNF599	ZNF599
RNF113A	SSR1	SSR1
NUDT1	HRSP12	HRSP12
NOL8.1	IMAGE.825583	LRRC8A
IMAGE.78217	LRRC8A	IMAGE.825583
SLC1A6	GJA4.1	GJA4.1
IMAGE.825583	FUT8	FUT8
IMAGE.309119	IMAGE.78217	IMAGE.78217
GJA4.1	NOL8.1	NOL8.1
AP2B1	COL6A2.1	COL6A2.1
IMAGE.139490	HDGFRP3	HDGFRP3
IMAGE.898259	3 RNF208	RNF208

Table 6 Top genes selected by each of the three screening methods examined with BRCA1 as the response variable

interest are FTO, which has been found to be associated with cancer ([36]) and RAB2A.1, which plays a role in breast cancer ([54]).

CP-SIS	ML-SIS	TS-SIS
SSC.215	SSC.215	SSC.215
IMAGE.731426	SEMA3D	SEMA3D
SEMA3D	CCNG1	CCNG1
CCNG1	IMAGE.731426	IMAGE.731426
SYBU	FTO	FTO
ITIH2	ITIH2	ITIH2
GDPD3	IMAGE.139490	IMAGE.139490
IL10RB	GDPD3	TRAF4
C4BPA	RAB2A.1	GDPD3
TRAF4	TRAF4	PLIN3.1
KLRC2	TECRL	TECRL
KLHDC9	PLIN3.1	RAB2A.1
SSC.52	ST8SIA2	ST8SIA2
PLIN3.1	CNTRL.RIP	IMAGE.839829
ST8SIA2	IMAGE.839829	SSC.44
IMAGE.839829	SSC.44	C4BPA
LIMK1	C4BPA	SLC6A15.2

Table 7 Top genes selected by each of the three screening methods examined with BRCA2 as the response variable

The HOXB13 gene has 102 observations in this data yielding a set size of 22. Results are shown in Table 8. There are 9 genes ranked top 22 by all three methods. Genes IMAGE.201264 and SSC.48 are selected by TS-SIS and ML-SIS but not by CP-SIS. The most notable gene top ranked is CDK9 which

has been shown to be related to prostate cancer [64] and is chosen by CP-SIS and ML-SIS, however missed by TS-SIS.

CP-SIS	ML-SIS	TS-SIS
ERGIC1	ERGIC1	ERGIC1
GOLGA7	AADAT	GOLGA7
CR1L	SETD7	CR1L
AADAT	ANK3.1	HINFP
HINFP	CDK9	AADAT
ARF1	IMAGE.201264	ARF1
SETD7	TRIM64	SETD7
SLC39A13	SSC.48	ANK3.1
CDK9	LRRIQ3	SLC39A13
ANK3.1	GMPR	PPP1CA
PPP1CA	PPP1CA.1	IMAGE.201264
FAAH	TXNRD2.1	SSC.48
TRIM64	VIPR1	TRIM64
ACLY.1	APLP2	FAAH
GMPR	NAPA	ACLY.1
CEBPD	CAPNS1	LRRIQ3
RAB11B	LTBP4	CEBPD
PPP1CA.1	ATP5J2	GMPR
LRRIQ3	IRF3	RAB11B
ERBB2.1	CLSTN1	PPP1CA.1
VIPR1	IMAGE.884766	ERBB2.1
APLP2	CSPG4	APLP2

Table 8 Top genes selected by each of the three screening methods examined with HOXB13 as the response variable

5 Discussion

In this paper we study the problem of performing ultra-high dimensional variable screening in the presence of incomplete data that is missing at random. The sure screening property is shown for the screening method that uses as ranking utility the marginal correlation coefficient of each predictor and the response that is computed after imputation of the missing values with maximum likelihood (ML-SIS). In order to use the information about the missing values of each predictor contained in other predictors, we propose a new screening method composed of two steps: for each predictor, first select the other predictor variables that significantly correlate with it and perform the imputation of its missing values using maximum likelihood on their joint distribution together with the response variable; then perform screening using the correlation coefficient computed after the imputation as ranking utility. This Two-Stage method, called TS-SIS, was compared in simulation studies to the ML-SIS method and screening performed after dropping the missing rows of observations. First, it is unsurprising but important to note that there were no scenarios under consideration in the simulation study where CP-SIS

outperformed MLE-SIS. TS-SIS outperforms both complete case analysis and MLE when covariates are highly correlated with a few other covariates (i.e. $\sigma_{k\ell} = 0.9^{|k-\ell|}$). Alternatively, when the covariance structure of the data has moderate correlation throughout (i.e. $\sigma_{k\ell} = 0.6$ for all k,ℓ) among all pair of variables, there are certain patterns of missingness where we observe TS-SIS performing worse than not only MLE, but also complete case analysis.

Simulations not reported in the paper for the sake of brevity suggest that the performance of the screening methods considered follow a similar pattern when the assumption of normality is not valid. Note that the use of the EM algorithm to estimate the parameters allows one to perform the MLE or the two-stage screening for any type of distributions, while empirical results support the suggestion of using these methods instead of simply using the complete pairwise observations.

The manuscript then concludes with a real data application of micro array dataset from a prostate cancer study and compares the results of the three methods. Four different response variables (STAT3, HOXB13, BRCA1, and BRCA2) that are all associated with prostate cancer were considered as the response variable for screening and in all of those cases there was significant overlap in the sets that were chosen between the three methods. In all cases, we observe genes in the top genes ranked by TS-SIS or ML-SIS that are not in the top ranked by CP-SIS. Because most genes are correlated with a few other genes, we suggest the use of TS-SIS to select the top set of active predictors.

In conclusion, we do not recommend the use of only complete pair of observations CP-SIS for screening on a data set with missing data. Instead, we recommend first studying the empirical structure of the covariance matrix. If there is high correlation between many variables, the proposed MLE-SIS is recommended, whereas if high correlation is observed between a few predictors, the proposed Two-Stage screening is recommended.

6 Appendix

Proof of Theorem 21

Proof Recall that the log-likelihood of $\boldsymbol{\phi}_j$ is

$$\ell(\phi_j|\{(X_i,Y_i)\}_{i=1}^{n_j},\{Y_k\}_{k=n_j+1}^{n}) = -\frac{1}{2\sigma_{j.y}^2}\sum_{i=1}^{n_j}(X_{ij}-\mu_{j.y}-\beta_{jy}Y_i)^2 - \frac{n_j\log(\sigma_{j.y}^2)}{2}$$
$$-\frac{1}{2\sigma_y^2}\sum_{i=1}^{n}(Y_i-\mu_y)^2 - \frac{n\log(\sigma_y^2)}{2}.$$

The inverted hessian of the log-likelihood evaluated at the estimated parameters is

$$H_{\boldsymbol{\phi}_j}^{-1}|_{\hat{\boldsymbol{\phi}}_j} = \begin{bmatrix} \hat\sigma_y^2/n & 0 & 0 & 0 & 0 \\ 0 & 2\hat\sigma_y^4/n & 0 & 0 & 0 \\ 0 & 0 & \hat\sigma_{j.y}^2(1+\bar y^2/s_y^2)/n_j & -\bar y\hat\sigma_{j.y}^2/(n_j s_y^2) & 0 \\ 0 & 0 & -\bar y\hat\sigma_{j.y}^2/(n_j s_y^2) & \hat\sigma_{j.y}^2/(n_j s_y^2) & 0 \\ 0 & 0 & 0 & 0 & 2\hat\sigma_{j.y}^4/n_j \end{bmatrix},$$

so that the large sample covariance matrix for $\boldsymbol{\theta}_j$ can be written as $D(\rho_j)H_{\boldsymbol{\phi}_j}^{-1}|_{\hat{\boldsymbol{\phi}}_j}D(\rho_j)^T$, where
$D(\rho_j) = \left(\frac{\partial\rho_j}{\partial\mu_y}, \frac{\partial\rho_j}{\partial\sigma_y^2}, \frac{\partial\rho_j}{\partial\mu_{j.y}}, \frac{\partial\rho_j}{\partial\beta_{j.y}}, \frac{\partial\rho_j}{\partial\sigma_{j.y}^2}\right)$. It can be shown that

$$D(\rho_j) = \left(0, \frac{\sigma_{j.y}^2\beta_{j.y}}{2\sqrt{\sigma_y^2}(\beta_{j.y}^2\sigma_y^2+\sigma_{j.y}^2)^{3/2}}, 0, \frac{\sigma_{j.y}^2\sqrt{\sigma_y^2}}{(\beta_{j.y}^2\sigma_y^2+\sigma_{j.y}^2)^{3/2}}, -\frac{\beta_{j.y}\sqrt{\sigma_y^2}}{2(\beta_{j.y}^2\sigma_y^2+\sigma_{j.y}^2)^{3/2}}\right),$$

and hence one finds

$$\begin{aligned} D(\rho_j)H_{\boldsymbol{\phi}_j}^{-1}|_{\hat{\boldsymbol{\phi}}_j}D(\rho_j)^T &= \frac{\hat\sigma_{j.y}^4\hat\beta_{j.y}^2}{4\hat\sigma_y^2(\hat\beta_{j.y}^2\hat\sigma_y^2+\hat\sigma_{j.y}^2)^3}\frac{2\hat\sigma_y^4}{n} + \frac{\hat\sigma_{j.y}^4\hat\sigma_y^2}{(\hat\beta_{j.y}^2\hat\sigma_y^2+\hat\sigma_{j.y}^2)^3}\frac{\hat\sigma_{j.y}^2}{n_j s_y^2} + \frac{\hat\beta_{j.y}^2\hat\sigma_y^2}{4(\hat\beta_{j.y}^2\hat\sigma_y^2+\hat\sigma_{j.y}^2)^3}\frac{2\hat\sigma_{j.y}^4}{n_j} \\ &= \Big[2\hat\sigma_y^4 s_y^2 n_j(s_j^2-s_{jy}^2/s_y^2)^2 s_{jy}^2/s_y^4 + 4n\hat\sigma_y^4(s_j^2-s_{jy}^2/s_y^2)^3 \\ &\quad +2ns_y^2\hat\sigma_y^4(s_j^2-s_{jy}^2/s_y^2)^2 s_{jy}^2/s_y^4\Big]/\left[4nn_j\hat\sigma_y^2 s_y^2(\hat\sigma_y^2 s_{jy}^2/s_y^4+s_j^2-s_{jy}^2/s_y^2)^3\right] \\ &= \frac{(1-\tilde\rho)^2\hat\sigma_y^4\left[2s_y^2 n_j\tilde\rho^2 s_j^6/s_y^2 + 4ns_j^6(1-\tilde\rho^2) + 2ns_y^2 s_j^6\tilde\rho^2/s_y^2\right]}{4nn_j s_y^2\hat\sigma_y^2 s_j^6(\tilde\rho^2(\hat\sigma_y^2/s_y^2)+1)^3} \\ &= (1-\tilde\rho_j^2)^2\left(\frac{\hat\sigma_y^2}{s_y^2}\right)\left(\frac{1}{nn_j}\right)\left(\frac{\tilde\rho_j^2(n_j-n)/2+n}{\left(\tilde\rho_j^2(\frac{\hat\sigma_y^2}{s_y^2}-1)+1\right)^3}\right). \end{aligned}$$

Since $\hat\rho_j$ is the maximum likelihood estimator computed from a Normal distribution, it follows that $[D(\rho_j)H_{\boldsymbol{\phi}}^{-1}|_{\hat{\boldsymbol{\phi}}_j}D(\rho_j)^T]^{-1/2}(\hat\rho_j-\rho_j)$ converges to a standard Normal distribution.

Proof of Theorem 22

Proof First note that the estimated covariance s_{jy} based on the completely observed pairs is, except for a scale of $(n_j-1)/(n_j)$, a U-statistic (Kowalski and Tu, 2007)

$$s_{jy} = \frac{1}{n_j}\sum_{i=1}^{n_j}(Y_i-\bar{Y})(X_{ij}-\bar{X}_j) = \frac{n_j-1}{n_j}\binom{n_j}{2}^{-1}\sum_{i\neq k}^{n_j}\frac{1}{2}(Y_i-Y_k)(X_{ij}-X_{kj})$$
$$= \frac{n_j-1}{n_j}\frac{1}{(n_j)(n_j-1)}\sum_{i\neq k}^{n_j}h_j(Y_i,Y_k,X_{ij},X_{kj}) := \frac{n_j-1}{n_j}s^*_{jy},$$

where $\bar{X}_j = \sum_{i=1}^{n} X_{ij}$ and $h_j(Y_i,Y_k,X_{ij},X_{kj}) = (Y_i-Y_k)(X_{ij}-X_{kj})$ is the kernel of the U-statistic s^*_{jy}. Note that $E(s^*_{jy}) = \sigma_{jy} := \sigma_j\sigma_y\rho_j$.

We follow steps similar to those in [46]. First write

$$s^*_{jy} = s^*_{jy,1} + s^*_{jy,2} := \frac{1}{n_j(n_j-1)}\sum_{i\neq k}^{n_j}h_j(Y_i,Y_k,X_{ij},X_{kj})I(h_j(Y_i,Y_k,X_{ij},X_{kj})\leq M)$$
$$+\frac{1}{n_j(n_j-1)}\sum_{i\neq k}^{n_j}h_j(Y_i,Y_k,X_{ij},X_{kj})I(h_j(Y_i,Y_k,X_{ij},X_{kj})> M),$$

and define

$$\sigma_{jy,1} := E(s^*_{jy,1}) = E[h_j(Y_i,Y_k,X_{ij},X_{kj})I(h_j(Y_i,Y_k,X_{ij},X_{kj})\leq M)],$$
$$\sigma_{jy,2} := E(s^*_{jy,2}) = E[h_j(Y_i,Y_k,X_{ij},X_{kj})I(h_j(Y_i,Y_k,X_{ij},X_{kj})> M)].$$

Because $s^*_{jy,1}$ can be written as an average of averages of i.i.d. random variables ([66] - sec. 5.1.6), for any $t>0$ and $\epsilon>0$ we have

$$P(s^*_{jy,1}-\sigma_{yj,1}\geq\epsilon) \leq \exp(-t\epsilon)\exp(-t\sigma_{jy,1})E(\exp(ts^*_{jy,1}))$$
$$= \exp(-t\epsilon)\exp(-t\sigma_{jy,1})E\left(\exp\left(t\frac{1}{n_j!}\sum_{n_j!}\frac{1}{m}\sum_{m}h_j^{(m)}I(h_j^{(m)}\leq M)\right)\right)$$
$$\leq \exp(-t\epsilon)\exp(-t\sigma_{jy,1})E^m\left(\exp\left(\frac{1}{m}th_j^{(m)}I(h_j^{(m)}\leq M)\right)\right)$$
$$= \exp(-t\epsilon)E^m\left(\exp\left(\frac{1}{m}t\left(h_j^{(m)}I(h_j^{(m)}\leq M)-\sigma_{yj,1}\right)\right)\right),$$

where $m=[n_j/2]$ and the last inequality follows from Theorem 5.6.1A in [66]. Choose $t = 4\epsilon m/M^2$ so that $P(s^*_{jy,1}-\sigma_{jy,1}\geq\epsilon)\leq\exp(-2\epsilon^2m/M^2)$ and by symmetry of the U-statistics

$$P(|s^*_{jy,1}-\sigma_{jy,1}|\geq\epsilon)\leq 2\exp(-2\epsilon^2m/M^2). \tag{3}$$

Now we deal with $s^*_{jy,2}$. Note that using Cauchy-Schwarz and Markov inequalities we have

$$\begin{aligned}\sigma^2_{jy,2} &\le E[(Y_i - Y_k)^2(X_{ij} - X_{kj})^2]P[(Y_i - Y_k)(X_{ij} - X_{kj}) \ge M] \\ &\le E[(Y_i - Y_k)^2(X_{ij} - X_{kj})^2]E[\exp(s(Y_i - Y_k)(X_{ij} - X_{kj}))]\exp(-sM)\end{aligned}$$

for any $s > 0$. Using assumptions C1, if we choose $M = cn_j^\gamma$ for $0 < \gamma < 1/2-k$, then $\sigma_{jy,2} \le \epsilon/2$ when n_j is sufficiently large. Consequently,

$$\begin{aligned}P(|s^*_{jy,2} - \sigma_{jy,2}| > \epsilon) &\le P(|s^*_{jy,2}| > \epsilon/2) \le P(\cup\{(Y_i - Y_k)(X_{ij} - X_{kj}) > M\} \\ &\le n_j P((Y_i - Y_k)(X_{ij} - X_{kj}) > M) \\ &= n_j P[\exp(s(Y_i - Y_k)(X_{ij} - X_{kj})) > \exp(sM)] \\ &\le n_j \exp(-sM)E(\exp\{s(Y_i - Y_k)(X_{ij} - X_{kj})\}) = n_j C\exp(-sM),\end{aligned}$$

for any $s > 0$. Hence

$$\begin{aligned}P(|s^*_{jy} - \sigma_{jy}| > 2\epsilon) &= P(|s^*_{jy,1} + s^*_{jy,2} - \sigma_{jy,1} - \sigma_{jy,2}| \ge 2\epsilon) \\ &\le P(|s^*_{jy,1} - \sigma_{jy,1}| > \epsilon) + P(|s^*_{jy,2} - \sigma_{jy,2}| > \epsilon) \\ &\le O(\exp(-c_1\epsilon^2 n_j^{1-2\gamma}) + n_j\exp(-c_2 n_j^\gamma)). \qquad (4)\end{aligned}$$

Recall $\hat\rho_j = \frac{s_{jy}}{s_j s_y}\frac{\hat\sigma_y}{s_y}\frac{s_j}{\hat\sigma_j}$. Using similar arguments, one can show that the convergence rate of $s_y, s_j, \hat\sigma_y$ and $\hat\sigma_j$ have the same form of (4) and hence by Lemma S4 in [50] so does $\hat\rho_j$, so that we have

$$\begin{aligned}P(|\hat\rho_j - \rho_j| \ge cn_j^{-\kappa}) &\le P(|\hat\rho_j - \rho_j| \ge cn_j^{-\kappa}) \\ &= O([\exp(-c_1 n_j^{1-2(\gamma+\kappa)}) + n_j\exp(-c_2 n_j^\gamma)]).\end{aligned}$$

$$\begin{aligned}P(|\hat\rho_j - \rho_j| \ge cn_j^{-\kappa}, \text{ for all } j) &\le \sum_{j=1}^d P(|\hat\rho_j - \rho_j| \ge cn_j^{-\kappa}) \\ &= \sum_{j=1}^d O([\exp(-c_1 n_j^{1-2(\gamma+\kappa)}) + n_j\exp(-c_2 n_j^\gamma)]).\end{aligned}$$

Letting $\epsilon = cn_j^{-\kappa}$ we have

$$\begin{aligned}P(\max_{j=1,\dots,d}|\hat\rho_j - \rho_j| \ge cn_j^{-\kappa}) &\le d\max_{j=1,\dots,d}P(|\hat\rho_j - \rho_j| \ge cn_j^{-\kappa}) \\ &= d\max_{j=1,\dots,d}O(\exp(-c_1 n_j^{-2\kappa}n_j^{1-2\gamma}) + n_j\exp(-c_2 n_j^\gamma)) \\ &\le O(d\exp(-c_1\min_j n_j^{1-2(\gamma+\kappa)}) + \max_j\{n_j\exp(-c_2 n_j^\gamma)\}).\end{aligned}$$

If $\mathcal{A} \not\subseteq \hat{\mathcal{A}}$, then there exists a $j \in \mathcal{A}$ such that $\hat{\rho}_j < cn_j^{-\kappa}$. From condition C2 it follows that $|\hat{\rho}_j - \rho_j| > cn_j^{-\kappa}$ for some $j \in \mathcal{A}$. This implies that $\{\mathcal{A} \not\subseteq \hat{\mathcal{A}}\} \subseteq \{|\hat{\rho}_j - \rho_j| > cn_j^{-\kappa} \text{ for some } j \in \mathcal{A}\}$. Then

$$\begin{aligned} P(\mathcal{A} \subseteq \hat{\mathcal{A}}) &\geq P(|\hat{\rho}_j - \rho_j| \leq cn_j^{-\kappa}, \text{ for all } j \in \mathcal{A}) = 1 - P(|\hat{\rho}_j - \rho_j| > cn_j^{-\kappa}, \text{ for some } j \in \mathcal{A}) \\ &\geq 1 - \sum_{j \in \mathcal{A}} P(|\hat{\rho}_j - \rho_j| > cn_j^{-\kappa}) \\ &= 1 - \sum_{j \in \mathcal{A}} O(\exp(-c_1 \min_j n_j^{1-2(\gamma+\kappa)}) + \max_j \{n_j \exp(-c_2 n_j^{\gamma})\}). \end{aligned}$$

References

1. Abdulghani, J., Gu, L., Dagvadorj, A., Lutz, J., Leiby, B., Bonuccelli, G., et al. (2008). Stat3 promotes metastatic progression of prostate cancer. The American Journal of Pathology, 172(6):1717-1728.
2. Anderson, T. (1957). Maximum-likelihood estimation for the multivariate normal distribution when some observations are missing. Journal of the American Statistical Association, 52:200-203.
3. Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis. Wiley.
4. Attouch, M., Laksaci, A., and Messabihi, N. (2017). Nonparametric relative error regression for spatial random variables. Statistical Papers, 58(4):987-1008.
5. Barnett, G. C., Thompson, D., Fachal, L., Kerns, S., Talbot, C., Elliott, R. M., et al. (2014). A genome wide association study (gwas) providing evidence of an association between common genetic variants and late radiotherapy toxicity. Radiotherapy and oncology: Journal of the European Society for Therapeutic Radiology and Oncology, 111(2):178-185.
6. Beebe-Dimmer, J., Hathcock, M., Yee, C., Okoth, L., Isaacs, W., Cooney, K., et al. (2015). The hoxb13 g84e mutation is associated with an increased risk for prostate cancer and other malignancies. Cancer Epidemiology, Biomarkers, and Prevention, 24(9):1366-1372.
7. Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4):1165-1188.
8. Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. Human genetics, 124(5):439-450.
9. Candes, E. and Tao, T. (2007). The dantzig selector statistical estimation when p is much larger than n. The Annals of Statistics, 35(6):2313-2351.
10. Castro, E. and Eeles, R. (2012). The role of brca1 and brca2 in prostate cancer. Asian Journal of Andrology, 14(3):409-414.
11. Cheema, J. (2014). A review of missing data handling methods in education research. Review of Educational Research, 84(4):487-508.
12. Chen, Q. and Wang, S. (2013). Variable selection for multiply imputed data with application to dioxin exposure study. Statistics in Medicine, 32(21):3646-3659.
13. Chen, X., Chen, X., and Liu, Y. (2017). A note on quantile feature screening via distance correlation. Statistical Papers.
14. Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. Biometrics, 64:1062-1069.
15. Dai, J., Ruczinski, I., LeBlanc, M., and Kooperberg, C. (2006). Imputation methods to improve inference in snp association studies. Genetic Epidemi- ology, 30(8):690-702.
16. Dang, Y., Chang, C., Ido, M., and Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1-38.
17. Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B. (Methodological), 39(1):1-38.

18. Deters, K. D., Nho, K., Risacher, S. L., Kim, S., Ramanan, V. K., Crane, P. K., et al. (2017). Genome-wide association study of language performance in alzheimer's disease. Brain and language.
19. Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature, 447(7148):10871093.
20. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. The Annals of Statistics, 32(2):407-499.
21. Elkashef, A., Allison, S., Sadiq, M., Basheer, H., Morais, G., Loadman, P., et al. (2016). Polysialic acid sustains cancer cell survival and migratory capacity in a hypoxic environment. Scientific Reports, 6(33026).
22. Ewing, C. M., Ray, A. M., Lange, E. M., Zuhlke, K. A., Robbins, C. M., Tembe, W. D., et al. (2012). Germline mutations in hoxb13 and prostate-cancer risk. New England Journal of Medicine, 366(2):141-149. PMID: 22236224.
23. Faisal, S. and Tutz, G. (2017). Missing value imputation for gene expression data by tailored nearest neighbors. Statistical Applications in Genetics and Molecular Biology, 16(2):95-106.
24. 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(494):544-557.
25. 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.
26. 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.
27. Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica, 20:101-148.
28. Fan, J., Samworth, R., and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. Machine Learning Research, 10:1829-1853.
29. Faria, R., Gomes, M., Epstein, D., and White, I. (2014). A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials. Pharmacoeconomics, 32(12):1157-1170.
30. Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., et al. (2011). Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. Journal of the National Cancer Institute, 103(5):425-435.
31. Garcia, R. I., Ibrahim, J. G., and Zhu, H. (2010). Variable selection in the cox regression model with covariates missing at random. Biometrics, 66:97-104.
32. Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. Bernoulli, 10:971-988.
33. Haffmann, E., Sorenson, B., Sauter, D., and Lambert, I. (2015). Role of volume-regulated and calcium-activated anion channels in cell volume homeostasis, cancer and drug resistance. Channels (Austin), 9(6):380-396.
34. Harel, O., Pellowski, J., and Kalichman, S. (2012). Are we missing the importance of missing values in hiv prevention randomized clinical trials? reviews and recommendations. AIDS and Behavior, 16(6):1382-1393.
35. Harel, O. and Zhou, X. (2007). Multiple imputation: review of theory, implementation, and software. Statistics in Medicine., 26(16):3057-3077.
36. Hernandez-Caballero, M. and Sierra-Ramirez, J. (2015). Single nucleotide polymorphisms of the fto gene and cancer risk: an overview. Molecular Biology Reports, 42(3):699-704.
37. Horowitz, J. L. (2015). Variable selection and estimation in high-dimensional models. Canadian Journal of Economics/Revue canadienne deconomique, 48(2):389-407.
38. Horton, N. and Kleinman, K. (2007). Much ado about nothing: A comparison of missing data methods and software to
fit incomplete data regression models. The American Statistician, 61(1):79-90.
39. Ibrahim, J. G., Lipsitz, S. R., and Chen, M. H. (2001). Missing responses in generalized linear mixed models when the missing data mechanism is nonignorable. Biometrika, 88:551-564.

40. Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. Journal of the American Statistical Association, 103:1648-1658.
41. Karimi, O. and Mohammadzadeh, M. (2012). Bayesian spatial regression models with closed skew normal correlated errors and missing observations. Statistical Papers, 53(1):205-218.
42. Komatsu, J., Ichikawa, D., Hirajima, S., Nagata, H., Nishimura, Y., Kawaguchi, T., et al. (2015). Overexpression of smyd2 contributes to malignant outcome in gastric cancer. British Journal of Cancer, 112:357-364.
43. Lai, P., Liu, Y., Liu, Z., and Wan, Y. (2017). Model free feature screening for ultrahigh dimensional data with responses missing at random. Comput. Stat. Data Anal., 105(C):201-216.
44. Lansangan, J. R. G. and Barrios, E. B. (2017). Simultaneous dimension reduction and variable selection in modeling high dimensional data. Computational Statistics & Data Analysis, 112:242 -256.
45. Law, M. H., Bishop, D. T., Lee, J. E., Brossard, M., Martin, N. G., Moses, E. K., et al. (2015). Genome-wide meta-analysis identifies
five new susceptibility loci for cutaneous malignant melanoma. Nature genetics, 47(9):987-995.
46. Li, R., Zhong, W., and Zhu, L. (2012a). Feature screening via distance correlation learning. Journal of the American Statistical Association, 107(499):1129-1139. PMID: 25249709.
47. Li, Z., Gopal, V., Li, X., Davis, J., and Casella, G. (2012b). Simultaneous snp identification in association studies with missing data. The Annals of Applied Statistics, 6(2):432-456.
48. Liew, A., Law, N., and Yan, H. (2011). Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in Bioinformatics, 12(5):498-513.
49. Little, R. and Rubin, D. (2002). Statistical analysis with missing data. Wiley series in probability and statistics.
50. Liu, J., Li, R., and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh-dimensional covariates. Journal of the American Statistical Association, 109(505):266-274.
51. Liu, Y., Wang, Y., Feng, Y., and Wall, M. (2016). Variable selection and prediction with incomplete high-dimensional data. Annals of Applied Statistics, 10(1):418-450.
52. Long, Q. and Johnson, B. (2015). Variable selection in the presence of missing data: resampling and imputation. Biostatistics, 16(3):596-610.
53. Lu, J. and Lin, L. (2017). Model-free conditional screening via conditional distance correlation. Statistical Papers.
54. Luo, M., Gong, C., Chen, C., Hu, H., Huang, P., Zheng, M., et al. (2015). The rab2a gtpase promotes breast cancer stem cells and tumorigenesis via erk signaling activation. Cell Reports, 11(1):111-124.
55. Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics, 39:906-913.
56. Mills, I. (2014). Hoxb13, rfx6 and prostate cancer risk. Nature Genetics, 46:94-95.
57. Nagy, R., Boutin, T. S., Marten, J., Human, J. E., Kerr, S. M., Campbell, A., et al. (2017). Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 generation scotland participants. Human genetics, 9(1):23.
58. Neykov, N. M., Filzmoser, P., and Neytchev, P. N. (2014). Ultrahigh dimensional variable selection through the penalized maximum trimmed likelihood estimator. Statistical Papers, 55(1):187-207.
59. Paik, M. C. and Tsai, W. (1997). On using cox proportional hazard model with missing covariates. Biometrika, 84:579-593.
60. Pencik, J., Schlederer, M., Gruber, W., Unger, C., Walker, S. M., Chalaris, A., et al. (2015). Stat3 regulated arf expression suppresses prostate cancer metastasis. Nature Communications, 6:7736.
61. Pilie, P., Giri, V., and Cooney, K. (2016). Hoxb13 and other high penetrant genes for prostate cancer. Asian Journal of Andrology, 18(4):530-532.

62. Pritchard, C. C., Mateo, J., Walsh, M. F., De Sarkar, N., Abida, W., Beltran, H., et al. (2016). Inherited dna-repair gene mutations in men with metastatic prostate cancer. New England Journal of Medicine, 375(5):443-453. PMID: 27433846.
63. Rabier, C.-E., Azas, J.-M., Elsen, J.-M., and Delmas, C. (2016). Chi-square processes for gene mapping in a population with family structure. Statistical Papers.
64. Rahaman, M., Kumarasiri, M., Mekonnen, L., Yu, M., Diab, S., Albrecht, H., et al. (2016). Targeting cdk9: a promising therapeutic opportunity in prostate cancer. Endocrine-Related Cancer, 23(12):T211-T226.
65. Rubin, D. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley series in probability and mathematical statistics.
66. Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics, Wiley Series in Probability and Statistics. John Wiley & Sons Inc., New York.
67. Shen, C.-W. and Chen, Y.-H. (2012). Model selection for generalized estimating equations accommodating dropout missingness. Biometrics, 68:1046-1054.
68. Suhre, K., Arnold, M., Bhagwat, A. M., Cotton, R. J., Engelke, R., Raer, J., et al. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. Nature communications, 8:14357.
69. Tang, N., Xia, L., and Yan, X. (2018). Feature screening in ultrahighdimensional partially linear models with missing responses at random. Computational Statistics & Data Analysis.
70. Tibshirani, R. (1996). Regression shrinkage and selection via the lasoo. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267-288.
71. Tomlins, S. A., Laxman, B., Dhanasekaran, S. M., Helgeson, B. E., Cao, X., Morris, D. S., et al. (2007). Distinct classes of chromosomal rearrangements create oncogenic ets gene fusions in prostate cancer. Nature, 448:595-599.
72. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., et al. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17:520-525.
73. Trust, W. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3000 shared controls. Nature, 447:661-678.
74. Wang, Q. and Li, Y. (2018). How to make model-free feature screening approaches for full data applicable to the case of missing response? Scandinavian Journal of Statistics, 45(2):324-346.
75. Wang, S., Nan, B., Rosset, S., and Zhu, J. (2011). Random lasso. Annals of Applied Statistics, 5:468-485.
76. Wang, X., Inzunza, H., Chang, H., Qi, Z., Hu, B., Malone, D., et al. (2013). Mutations in the hedgehog pathway genes smo and ptch1 in human gastric tumors. PLoS ONE.
77. Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. The Annals of Statistics, 37(5A):2178-2201.
78. Yan, Q., Brehm, J., Pino-Yanes, M., Forno, E., Lin, J., Oh, S. S., et al. (2017). A meta-analysis of genome-wide association studies of asthma in puertoricans. The European respiratory journal, 49(5).
79. Yang, H., Guo, C., and Lv, J. (2016). Variable selection for generalized varying coefficient models with longitudinal data. Statistical Papers, 57(1):115-132.
80. Yang, H. and Liu, H. (2016). Penalized weighted composite quantile estimators with missing covariates. Statistical Papers, 57(1):69-88.
81. Yang, X., Belin, T. R., and Boscardin, W. J. (2005). Imputation and variable selection in linear regression models with missing covariates. Biometrics, 61:498-506.
82. Yoon, D., Lee, E., and Park, T. (2007). Robust imputation method for missing values in mocroarray data. BMC Bioinformatics, 8 Supple 2:S6.
83. Zambom, A. Z. and Akritas, M. G. (2018). Hypothesis testing sure independence screening for nonparametric regression. Electron. J. Statist., 12(1):767-792.
84. Zhao, Y. and Long, Q. (2016). Multiple imputation in the presence of highdimensional data. Statistical Methods in Medical Research, 25(5).
85. Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 101:1418-1429.