Statistica Sinica **30** (2020), 925-953
doi:https://doi.org/10.5705/ss.202017.0439

VARIABLE SCREENING WITH MULTIPLE STUDIES

Tianzhou Ma[1], Zhao Ren[2] and George C. Tseng[2]

[1] *University of Maryland and* [2] *University of Pittsburgh*

Abstract: Advancements in technology have generated abundant high-dimensional data, enabling us to integrate multiple relevant studies. In terms of variable selection, the significant computational advantage of variable screening methods based on marginal correlations has resulted in these becoming promising alternatives to the popular regularization methods. However, these screening methods have thus far been limited to single studies. In this study, we consider a general framework for variable screening across multiple related studies. As such, we propose a novel two-step screening procedure, based on a self-normalized estimator, for high-dimensional regression analyses within this framework. Compared with the one-step procedure and rank-based sure independence screening (SIS) procedures, the proposed procedure greatly reduces the false negative rate, while keeping a low false positive rate. From a theoretical perspective, we show that our procedure possesses the sure screening property, with weaker assumptions on the signal strengths, and allows the number of features to grow at an exponential rate with the sample size. In addition, we relax the commonly used normality assumption and allow sub-Gaussian distributions. Simulations and a real transcriptomic application illustrate the advantage of our method over the rank-based SIS method.

Key words and phrases: Multiple studies, partial faithfulness, self-normalized estimator, sure screening property, variable selection.

1. Introduction

In many scientific disciplines, such as omics studies (including genomics, transcriptomics, etc.), biomedical imaging, and signal processing, high-dimensional data with number of features that far larger than the respective sample sizes (i.e., $p \gg n$) have become the rule rather than the exception. For example, biologists may wish to predict certain clinical outcome (e.g., survival) using gene-expression data, where they have far more genes than they do samples. Advancements in technology and a reduction in the price of biomedical research have yielded increasing numbers of experiments being performed on related hypotheses or that explore the same scientific question. Individual studies may have small sample sizes with limited statistical power. Thus integrating the data from multiple studies can improve statistical power, estimation accuracy, and reproducibility. However, the direct merging of data (i.e., a “mega-analysis”) is usually less favored, owing to the inherent discrepancies between studies (Tseng, Ghosh and Feingold (2012)). New statistical methodologies and theories are required to solve high-dimensional problems that integrate multiple related studies.

Various regularization methods are used for feature selection in high-dimensional regression problems. Popular methods include, but are not limited to, the Lasso (Tibshirani (1996)), SCAD (Fan and Li (2001)), elastic net (Zou and Hastie (2005)), and adaptive Lasso (Zou (2006)) methods. When a group structure exists among the variables (e.g., a set of gene features belong to a prespecified pathway), a group version of the regularization methods can be applied (Yuan and Lin (2006); Meier, Van De Geer and Bühlmann (2008); Nardi and Rinaldo (2008)). Refer to Fan and Lv (2010) and Huang, Breheny and Ma (2012) for a detailed overview of variable selection and group selection in high-dimensional models. When the number of features grows significantly larger than the sample size, most regularization methods perform poorly, owing to the simultaneous challenges of computational efficiency, statistical accuracy, and algorithmic stability (Fan, Samworth and Wu (2009)). As an alternative, variable screening methods first reduce the dimension of the problem, and then perform variable regularization. Fan and Lv (2008) proposed a sure independent screening (SIS) method to select features based on their marginal correlations with the response, in the context of linear regression models, showing that their fast selection procedure enjoys the “sure screening property”. Since the development of the SIS method, many screening methods have been proposed for generalized linear models (Fan, Samworth and Wu (2009); Fan and Song (2010); Chang, Tang and Wu (2013)), nonparametric additive models or semiparametric models (Fan, Feng and Song (2011); Chang, Tang and Wu (2016)), quantile linear regressions (Ma, Li and Tsai (2017)), and Gaussian graphical models (Luo, Song and Witten (2014); Liang, Song and Qiu (2015)) that exploit more robust measures for sure screening (Zhu et al. (2011); Li, Zhong and Zhu (2012); Li, Liu and Lou (2017)). However, these screening methods have thus far been limited to single studies.

In this paper, we first propose a general framework for simultaneous variable screening across multiple related studies. Including multiple studies provides additional evidence with which to reduce the dimension and, thus, increase the accuracy and efficiency of removing unimportant features during screening. To the best of our knowledge, ours is the first work to employ multiple studies for variable screening in a high-dimensional linear regression model. Such a framework provides a novel perspective of the screening problem and opens a door to the development of methods using multiple studies to perform screening under different types of models or with different marginal utilities. In this framework, it is natural to apply a screening procedure to each individual study. However, important features with weak signals in some studies may be incorrectly screened out in this case. To avoid such false negative errors and to fully take advantage of multiple studies, we propose a two-step screening procedure. Here, in addition to the traditional one-step procedure, we include a step that combines studies with potential zero correlation as a second check. This procedure can potentially to save those features with weak signals in individual studies, but that have a strong aggregate effect across studies during the screening stage. Compared with the naive multiple study extension of the SIS method, our procedure greatly reduces the false negative error rate, while keeping a low false positive rate. These merits are confirmed by our theoretical analysis. Specifically, we show that our procedure possesses the sure screening property, with weaker assumptions on the signals, and allows the number of features to grow at an exponential rate with the sample size. Furthermore, we require only that the data have a sub-Gaussian distribution using novel self-normalized statistics. Thus, our procedure can be applied to a more general distribution family than the Gaussian distribution, which is considered in Fan and Lv (2008) and Bühlmann, Kalisch and Maathuis (2010) for a related screening procedure under single study scenarios. After screening, we apply two general variable selection algorithms: a multiple study extension of the PC-simple algorithm proposed by Bühlmann, Kalisch and Maathuis (2010), and a two-stage feature selection method, which we use to choose the final model in a lower dimension.

The rest of the paper is organized as follows. In Section 2, we present a framework for variable screening with multiple related studies, as well as the notations used in this paper. Then, we propose our two-step screening procedure in Section 3. Section 4 provides the theoretical properties of our procedure, and demonstrates the benefits of including multiple related studies, as well as the advantages of our procedure. General algorithms for variable selection that follow from our screening procedure are discussed in Section 5. Sections 6 and 7 present the simulation studies and a real-data application based on three breast cancer transcriptomic studies, respectively, which illustrate the advantage of our method over the rank-based SIS method in terms of reducing false negative errors, while retaining important features. We conclude and discuss possible extensions of our procedure in Section 8. Section 9 provides technical proofs of the major theorems.

2. Model and Notation

Suppose we have data from K related studies, each with n observations. Consider a random design linear model in each study $k \in [K]$ ($[K] = 1, \ldots, K$):

$$Y^{(k)} = \sum_{j=1}^{p} \beta_j^{(k)} X_j^{(k)} + \epsilon^{(k)}, \tag{2.1}$$

where each $Y^{(k)} \in \mathbb{R}$; each $X^{(k)} = (X_1^{(k)}, \ldots, X_p^{(k)})^T \in \mathbb{R}^p$, with $E(X^{(k)}) = \mu_X^{(k)}$ and $\text{cov}(X^{(k)}) = \Sigma_X^{(k)}$; each $\epsilon^{(k)} \in \mathbb{R}$, with $E(\epsilon^{(k)}) = 0$ and $\text{var}(\epsilon^{(k)}) = \sigma^2$, such that $\epsilon^{(k)}$ is uncorrelated with $X_1^{(k)}, \ldots, X_p^{(k)}$; and $\beta^{(k)} = (\beta_1^{(k)}, \ldots, \beta_p^{(k)})^T \in \mathbb{R}^p$. We assume implicitly that $E(Y^{(k)2}) < \infty$ and $E\{(X_j^{(k)})^2\} < \infty$, for $j \in [p]$ ($[p] = 1, \ldots, p$).

When p is very large, we usually assume that only a small set of covariates are true predictors that contribute to the response. In other words, we assume most of $\beta_j = (\beta_j^{(1)}, \ldots, \beta_j^{(K)})^T$, where $j \in [p]$, are equal to a zero vector. Here we further assume that $\beta_j^{(k)}$ is either zero or nonzero in all K studies. This framework is partially motivated by an existing high-dimensional linear random effect model considered in the literature (e.g.,Jiang et al. (2016)). More specifically, we have $\beta = (\beta_{(1)}^T, 0^T)^T$, where $\beta_{(1)}$ is the vector of the first s_0 nonzero components of β ($1 \le s_0 \le p$). Consider a random effect model, where only the true predictors of each study are treated as the random effect; that is, $\beta^{(k)} = (\beta_{(1)}^{(k)}, 0)^T$ and $\beta_{(1)}^{(k)}$ is distributed as $N(\beta_{(1)}, \tau^2 I_{s_0})$, where τ^2 is independent of ϵ and X. Consequently, $\beta_j^{(k)}$ is are either zero or nonzero in all K studies, with probability one. In practice, for example, genome-wide association studies (GWAS) usually contain millions of SNPs, but only a few SNPs are important and predictive. The vast majority of SNPs are not associated with the outcome in any study, thus giving consistent sparse patterns across studies. For the few important SNPs, it is possible that signals in other studies have varying strengths, owing to population heterogeneity. Ma, Huang and Song (2011) considered a two-norm group bridge penalty for variable selection with multiple high-dimensional -omics data sets (e.g., gene expression data), where the regression coefficients of the same feature from multiple studies are treated as a group. A group is either selected or not (i.e., “all-in-or-all-out”). The selection of a group leads to nonzero estimated coefficients in all studies, but allows for different strengths of associations in the studies. A different and less constrained model, allowing sparsity across studies, has also been investigated in the literature (Li and Tseng (2011); Li et al. (2014)); however, this is beyond the scope of this study.

With n independent and identically distributed (i.i.d.) observations from model (2.1), our purpose is to identify the nonzero $\beta_{(1)}$. Thus we define the following index sets for active and inactive predictors:

$$\begin{aligned} \mathcal{A} &= \{j \in [p]; \beta_j \neq 0\} = \{j \in [p]; \beta_j^{(k)} \neq 0 \text{ for all } k\}; \\ \mathcal{A}^C &= \{j \in [p]; \beta_j = 0\} = \{j \in [p]; \beta_j^{(k)} = 0 \text{ for all } k\}, \end{aligned} \tag{2.2}$$

where $\mathcal{A}$ is our target. Clearly, under our setting, $\mathcal{A}$ and $\mathcal{A}^C$ are complementary to each other, such that the identification of $\mathcal{A}^C$ is equivalent to the identification of $\mathcal{A}$. Let $|\mathcal{A}| = s_0$, where $|\cdot|$ denotes the cardinality.

3. Screening Procedure for Multiple Studies

3.1. Sure independence screening

For a single study ($K = 1$), Fan and Lv (2008) proposed a variable screening method called sure independence screening (SIS) that ranks the importance of variables according to their marginal correlation with the response. As such, they were able to show its power in preliminary screening and dimension reduction for high-dimensional regression problems. Bühlmann, Kalisch and Maathuis (2010) later introduced a partial faithfulness condition, which states that a zero partial correlation for some separating set S implies a zero regression coefficient, showing that it holds almost surely for a joint normal distribution. In the extreme case, when $S = \emptyset$, it is equivalent to the SIS method.

The purpose of sure screening is to identify a set of moderate size d (with $d \ll p$) that still contains the true set $\mathcal{A}$. Equivalently, we can try to identify $\mathcal{A}^C$, or subsets of $\mathcal{A}^C$, that contain unimportant features that need to be screened out. There are two potential errors that may occur in any sure screening methods (Fan and Lv (2010)):

1. **False Negative (FN):** Important predictors that are marginally uncorrelated, but that are jointly correlated with the response, fail to be selected.
2. **False Positive (FP):** Unimportant predictors that are highly correlated with the important predictors can have a higher priority of being selected than other relatively weaker important predictors.

The current framework for variable screening with multiple studies resolves FP errors significantly. Indeed, we have multiple studies in our model setting. Thus, we have greater evidence with which to exclude noise and reduce FP errors than if we were using a single study only. In addition, sure screening is used to reduce the dimension at the first stage. Therefore, we can include second-stage variable selection methods, such as the Lasso or Dantzig selection, to further refine the set and, thus, reduce FP errors.

The FN errors occur when signals are falsely excluded after screening. Suppose ρ_j is the marginal correlation of the jth feature with the response, which we use to identify the set $\{j : \rho_j = 0\}$ for the screening out process. Under the assumption of partial faithfulness (defined in Section 4.3), these variables have zero coefficients for sure, which means the FN errors are guaranteed to be excluded. However, this might not be true for the empirical version of a marginal correlation. For a single study ($K = 1$), to eliminate the FN errors in the empirical case, it is well known that the signal-to-noise ratio has to be large (at least of order $(\log p/n)^{1/2}$, after a Bonferroni adjustment). In the current setting with multiple studies, the requirement on strong signals remains the same if we naively perform one-step screening in each individual study. However, we propose a novel two-step screening procedure that allows weak signals in individual studies, as long as the aggregate effect is sufficiently strong. Therefore our procedure reduces FN errors in the framework with multiple studies.

Before closing this section, note that to perform a screening test, one usually applies Fisher's Z-transformation to the sample correlation (Bühlmann, Kalisch and Maathuis (2010)). However, this requires a bivariate normality assumption. As an alternative, we propose using the self-normalized estimator of the correlation, which works well, in general, even for non-Gaussian data (Shao (1999)). Similar ideas have been applied to estimations of large covariance matrices (Cai and Liu (2016)).

3.2. Two-step screening procedure for multiple studies

Given multiple studies, we have greater evidence with which to reduce the dimension, where $\rho_j^{(k)} = 0$, for any k, implies a zero coefficient for that feature. On the one hand, it is possible for features with zero β_j to have multiple nonzero $\rho_j^{(k)}$. On the other hand, a nonzero β_j has nonzero $\rho_j^{(k)}$ in all studies. Thus, we aim to identify the following two complementary sets while performing screening using multiple studies:

$$\begin{aligned} \mathcal{A}^{[0]} &= \{j \in [p]; \quad \min_k |\rho_j^{(k)}| = 0\}, \\ \mathcal{A}^{[1]} &= \{j \in [p]; \quad \min_k |\rho_j^{(k)}| \neq 0\}. \end{aligned} \tag{3.1}$$

We know for sure that $\mathcal{A}^{[0]} \subseteq \mathcal{A}^C$ and $\mathcal{A} \subseteq \mathcal{A}^{[1]}$, with the partial faithfulness assumption. For $j \in \mathcal{A}^{[0]}$, the chance of detecting a zero marginal correlation in at least one study greatly increases with increasing K. Thus, unimportant features are more likely be screened out than they are in the single study scenario.

One way to estimate $\mathcal{A}^{[1]}$ is to test $H_0 : \rho_j^{(k)} = 0$, for each k and each feature j. If any of the K tests are not rejected for a feature, we exclude this feature from $\hat{\mathcal{A}}^{[1]}$ (we call this the "one-step sure independence screening" procedure, or "OneStep-SIS"). This can be viewed as an extension of the screening test to a multiple study scenario. However, in reality, it is possible for important features to have weak signals, and thus small $|\rho_j^{(k)}|$, in at least one study. These features might be incorrectly classified as part of $\hat{\mathcal{A}}^{[0]}$ because weak signals can be indistinguishable from null signals in individual testing. This will lead to the serious problem of false excluding important features (FN) from the final set during screening.

This can be significantly improved by adding a second step that combines those studies with potential zero correlation (i.e., failed to reject the null $H_0 : \rho_j^{(k)} = 0$) identified in the first step, and then performs another aggregate test. For features with weak signals in multiple studies, as long as their aggregate test statistics is sufficiently large, they will be retained. This procedure is more conservative when screening features than is the first step alone, but it guarantees a reduction in the false negative rate.

For simplicity, we assume n i.i.d. observations $(X_i^{(k)}, Y_i^{(k)})$, for $i \in [n]$, are obtained from all K studies. It is straightforward to extend the current procedure and analysis to scenarios with different sample sizes across multiple studies; therefore, this is omitted here. Our proposed "two-step aggregation sure independence screening" procedure ("TSA-SIS" for short) is formally described below.

Step 1. Screening in each study

In the first step, we perform a screening test in each study $k \in [K]$; thus, we obtain an estimate of the study set with potential zero correlations $\hat{l}_j$, for each $j \in [p]$, as:

$$\hat{l}_j = \left\{ k; |\hat{T}_j^{(k)}| \leq \Phi^{-1}\left(1 - \frac{\alpha_1}{2}\right) \right\} \quad \text{and} \quad \hat{T}_j^{(k)} = \frac{\sqrt{n}\hat{\sigma}_j^{(k)}}{\sqrt{\hat{\theta}_j^{(k)}}}, \tag{3.2}$$

where $\hat{\sigma}_j^{(k)} = (1/n)\sum_{i=1}^n (X_{ij}^{(k)} - \bar{X}_j^{(k)})(Y_i^{(k)} - \bar{Y}^{(k)})$ is the sample covariance, and $\hat{\theta}_j^{(k)} = (1/n)\sum_{i=1}^n [(X_{ij}^{(k)} - \bar{X}_j^{(k)})(Y_i^{(k)} - \bar{Y}^{(k)}) - \hat{\sigma}_j^{(k)}]^2$. Here, $\hat{T}_j^{(k)}$ is the self-normalized estimator of the covariance between $X_j^{(k)}$ and $Y^{(k)}$, Φ is the CDF of the standard normal distribution, and α_1 is a prespecified significance level.

In each study, we test whether $|\hat{T}_j^{(k)}| > \Phi^{-1}(1 - \alpha_1/2)$; if not, we include study k in $\hat{l}_j$. This step does not screen out any variables, but instead separates potential zero and nonzero study-specific correlations, in preparation for the next step. Define the cardinality of $\hat{l}_j$ as $\hat{\kappa}_j = |\hat{l}_j|$. If $\hat{\kappa}_j = 0$ (i.e., no potential zero correlation), we for sure retain feature j, and do not consider it in step 2; Otherwise, we move on to step 2.

Remark 1. By the scaling property of $\hat{T}_j^{(k)}$, it is sufficient to impose assumptions on the standardized variables: $W^{(k)} = (Y^{(k)} - E(Y^{(k)}))/(\sqrt{\mathrm{var}(Y^{(k)})}), Z_j^{(k)} = (X_j^{(k)} - E(X_j^{(k)}))/(\sqrt{\mathrm{var}(X_j^{(k)})})$. Thus, $\hat{T}_j^{(k)}$ can also be treated as a self-normalized estimator of the correlation. We thus define $\theta_j^{(k)} = \mathrm{var}(Z_j^{(k)} W^{(k)})$ and $\sigma_j^{(k)} = \mathrm{cov}(Z_j^{(k)}, W^{(k)}) = \rho_j^{(k)}$.

Remark 2. In our analysis, the index set in (3.2) is shown to coincide with $l_j (j \in \mathcal{A}^{[0]})$ and $l_j (j \in \mathcal{A}^{[1]})$; see Section 4.

Step 2. Aggregate screening

In the second step, we test whether the aggregate effect of the potential zero correlations in $\hat{l}_j$ identified in step 1 is strong enough to be retained. Define the statistics $\hat{L}_j = \sum_{k \in \hat{l}_j} (\hat{T}_j^{(k)})^2$, which approximately follows a $\chi^2_{\hat{\kappa}_j}$ distribution, with degrees of freedom $\hat{\kappa}_j$ under the null. Thus, we estimate $\hat{\mathcal{A}}^{[0]}$ by:

$$\hat{\mathcal{A}}^{[0]} = \{j \in [p]; \hat{L}_j \leq \varphi_{\hat{\kappa}_j}^{-1}(1 - \alpha_2) \text{ and } \hat{\kappa}_j \neq 0\}, \tag{3.3}$$

or, equivalently, estimate $\hat{\mathcal{A}}^{[1]}$ by:

$$\hat{\mathcal{A}}^{[1]} = \{j \in [p]; \hat{L}_j > \varphi_{\hat{\kappa}_j}^{-1}(1 - \alpha_2) \text{ or } \hat{\kappa}_j = 0\}, \tag{3.4}$$

where $\varphi_{\hat{\kappa}_j}$ is the CDF of the chi-square distribution with degrees of freedom equal to $\hat{\kappa}_j$, and α_2 is the prespecified significance level.

The second step takes the sum of the squares of $\hat{T}_j^{(k)}$ from studies with potential zero correlation as the test statistic. For each feature j, we test whether $\sum_{k \in \hat{l}_j} (\hat{T}_j^{(k)})^2 > \varphi_{\hat{\kappa}_j}^{-1}(1 - \alpha_2)$. If rejected, we conclude that the aggregate effect is strong and the feature needs to be retained; otherwise, we screen it out. This step performs a second check in addition to the individual testing in step 1, and potentially saves those important features with weak signals in individual studies, but that have a strong aggregate effect.

The procedure proposed here involves two tuning parameters: α_1 and α_2. Because the actual screening test is performed in the second step, commonly used

Table 1. Toy example to demonstrate the strength of the two-step screening procedure.

		S1 (signal)	S2 (signal)	N1 (noise)
	$k = 1$	$\lvert\hat{T}_1^{(1)}\rvert = 3.71$	$\lvert\hat{T}_2^{(1)}\rvert = 3.70$	$\lvert\hat{T}_3^{(1)}\rvert = 0.42$
	$k = 2$	$\lvert\hat{T}_1^{(2)}\rvert = 3.16$	$\lvert\hat{T}_2^{(2)}\rvert = 2.71$	$\lvert\hat{T}_3^{(2)}\rvert = 0.54$
	$k = 3$	$\lvert\hat{T}_1^{(3)}\rvert = 3.46$	$\lvert\hat{T}_2^{(3)}\rvert = 2.65$	$\lvert\hat{T}_3^{(3)}\rvert = 0.56$
	$k = 4$	$\lvert\hat{T}_1^{(4)}\rvert = 3.63$	$\lvert\hat{T}_2^{(4)}\rvert = 2.68$	$\lvert\hat{T}_3^{(4)}\rvert = 0.12$
	$k = 5$	$\lvert\hat{T}_1^{(5)}\rvert = 3.24$	$\lvert\hat{T}_2^{(5)}\rvert = 1.94$	$\lvert\hat{T}_3^{(5)}\rvert = 0.69$
TSA-SIS	$\hat{l}_j$	$\emptyset$	$\{2, 3, 4, 5\}$	$\{1, 2, 3, 4, 5\}$
	$\hat{\kappa}_j$	0	4	5
	$\hat{L}_j$	-	$25.31 > \varphi_4(0.95)$	$1.27 < \varphi_5(0.95)$
	$\hat{\mathcal{A}}^{[0]}$	N	N	Y
	$\hat{\mathcal{A}}^{[1]}$	Y	Y	N
OneStep-SIS	$\hat{\mathcal{A}}^{[0]}$	N	Y	Y
	$\hat{\mathcal{A}}^{[1]}$	Y	N (FN)	N

significance levels such as $\alpha_2 = 0.05$, are recommended to reduce false negative errors, following Bühlmann, Kalisch and Maathuis (2010). In general, there is a trade-off between false negative errors and false positive errors, determined by the choice of α_1. To further reduce the rate of false negative errors during screening, we recommend using a small α_1 (e.g., 1e-4) in practical applications. A sensitivity analysis on the choices of these two parameters is performed in Section 6; the results supported our recommendation.

In Table 1, we use a toy example to demonstrate our idea and compare the two approaches (OneStep-SIS vs. TSA-SIS). Suppose we have five studies ($K = 5$) and three features (two signals and one noise). S1 is a strong signal, with $\beta = 0.8$ in all studies, S2 is a weak signal, with $\beta = 0.4$ in all studies, and N1 is noise, with $\beta = 0$. In THE hypothesis tests, both small β and zero β yield a small marginal correlation, and are sometimes indistinguishable. Suppose $T = 3.09$ is used as the threshold (corresponding to $\alpha_1 = 0.001$). For the strong signal S1, all studies have large marginal correlations; thus both OneStep-SIS and TSA-SIS include it correctly. The weak signal S2 has small correlations in many studies. As a result, it is incorrectly screened out by OneStep-SIS (FN). However, the TSA-SIS procedure saves it in the second step (with $\alpha_2 = 0.05$). Both methods tend to remove the noise N1 after screening.

4. Theoretical Properties

4.1. Assumptions and conditions

We impose the following conditions to establish the model selection consistency of our procedure:

(C1) (Sub-Gaussian Condition) There exist constants $M_1 > 0$ and $\eta > 0$, such that, for all $|t| \leq \eta$, $j \in [p]$, $k \in [K]$:

$$E\{\exp(tZ_j^{(k)2})\} \leq M_1, \quad E\{\exp(tW^{(k)2})\} \leq M_1.$$

In addition, there exist $\tau_0 > 0$, such that $\min_{j,k} \theta_j^{(k)} \geq \tau_0$.

(C2) The number of studies $K = O(p^b)$, for some constant $b \geq 0$. The dimension satisfies $\log^3(p) = o(n)$ and $\kappa_j \log^2 p = o(n)$, where κ_j is defined next.

(C3) For $j \in \mathcal{A}^{[0]}$, $l_j(j \in \mathcal{A}^{[0]}) = \{k; \rho_j^{(k)} = 0\}$ and $\kappa_j = |l_j|$. If $k \notin l_j$, then $|\rho_j^{(k)}| \geq C_3\sqrt{\log p/n}\sqrt{1.01\theta_j^{(k)}}$, where $C_3 = 3(L+1+b)$.

(C4) For $j \in \mathcal{A}^{[1]}$, $l_j(j \in \mathcal{A}^{[1]}) = \{k; |\rho_j^{(k)}| < C_1\sqrt{\log p/n}\sqrt{0.99\theta_j^{(k)}}\}$ and $\kappa_j = |l_j|$, where $C_1 = L+1+b$. If $k \notin l_j$, then $|\rho_j^{(k)}| \geq C_3\sqrt{\log p/n}\sqrt{1.01\theta_j^{(k)}}$. In addition, we require $\sum_{k \in l_j} |\rho_j^{(k)}|^2 \geq (C_2(\log^2 p + \sqrt{\kappa_j \log p}))/n$, where C_2 is some large positive constant.

The first condition (C1) assumes that each standardized variable $Z_j^{(k)}$ or $W^{(k)}$, for $j \in [p]$, $k \in [K]$, marginally follows a sub-Gaussian distribution in each study. This condition relaxes the normality assumption in (Fan and Lv (2008); Bühlmann, Kalisch and Maathuis (2010)). The second part of (C1) assumes there always exists some positive τ_0 not greater than the minimum variance of $Z_j^{(k)}W^{(k)}$. In particular, if $(X_j^{(k)}, Y^{(k)})$ jointly follows a multivariate normal distribution, then $\theta_j^{(k)} = 1 + \rho_j^{(k)2} \geq 1$; thus, we can always pick $\tau_0 = 1$.

The second condition (C2) allows the dimension p to grow at an exponential rate with the sample size n, which is a fairly standard assumption in high-dimensional analyses. Many sure screening methods (e.g. SIS, DC-SIS, TPC) use this assumption (Fan and Lv (2008); Li, Zhong and Zhu (2012); Li, Liu and Lou (2017)). Although the PC-simple algorithm (Bühlmann, Kalisch and Maathuis (2010)) assumes a polynomial growth of p_n as a function of n, this can be relaxed to assume exponential growth with n. We further require that the product $\kappa_j \log^2 p$ be small. This is used to control the errors in the second step of our screening procedure, and is always true if $K \log^2 p = o(n)$.

Condition (C3) assumes a lower bound on nonzero correlations (i.e., $k \notin l_j$) for features from $\mathcal{A}^{[0]}$. In other words, if the marginal correlation $|\rho_j^{(k)}|$ is not zero, then it needs to be larger than the signal-to-noise ratio. Although this is a key assumption for a single study in many sure screening methods (Fan and Lv (2008); Bühlmann, Kalisch and Maathuis (2010); Li, Zhong and Zhu (2012); Li, Liu and Lou (2017)), we only impose this assumption for $j \in \mathcal{A}^{[0]}$, rather than on all $j \in [p]$. This condition is used to control for type-II errors in step 1 for features from $\mathcal{A}^{[0]}$.

Condition (C4) gives assumptions on features from $\mathcal{A}^{[1]}$. We assume the correlations are small for $k \in l_j$, and large for $k \notin l_j$, such that studies with strong or weak signals can be well identified in the first step. For studies in l_j, we further require that the sum of the squares of their correlations be greater than a threshold; this controls for type-II errors in step 2. This condition is different to those of methods based on single studies, which they usually assume a lower bound on each marginal correlation for features from $\mathcal{A}^{[1]}$ as in (C3). We relax this condition, placing restriction on their L_2 norm only. This allows features from $\mathcal{A}^{[1]}$ to have weak signals in each study, but a strong combined signal. To appreciate this change, we compare the minimal requirements with and without step 2. For each $j \in \mathcal{A}^{[1]}$, in order to detect this feature, we need $|\rho_j^{(k)}| \geq C(\log p/n)^{1/2}$, with some large constant C, for all $k \in l_j$ and, thus, at least $\sum_{k \in l_j} |\rho_j^{(k)}|^2 \geq C^2 \kappa_j \log p/n$. By comparison, the assumption in (C4) is much weaker in reasonable settings $\kappa_j \gg \log p$.

4.2. Consistency of the two-step screening procedure

The first theorem addresses the consistency of the screening in step 1.

Theorem 1. *Consider a sequence of linear models, as in* (2.1)*, that satisfy Assumptions and Conditions* (C1)—(C4)*, and define the event* $A = \{\hat{l}_j = l_j$ *for all* $j \in [p]\}$*. Then, there exists a sequence* $\alpha_1 = \alpha_1(n,p) \to 0$ *as* $(n,p) \to \infty$*, where* $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$*, with* $\gamma = 2(L+1+b)$*, such that:*

$$P(A) = 1 - O(p^{-L}) \to 1 \text{ as } (n,p) \to \infty. \tag{4.1}$$

The proof of Theorem 1 can be found in Section 9. This theorem states that the screening in our first step correctly identifies the set l_j for features in both $\mathcal{A}^{[0]}$ and $\mathcal{A}^{[1]}$ (in which strong and weak signals are well separated), and that the chance of incorrect assignment is low. Given the results in Theorem 1, we can now show the main theorem for the consistency of the two-step screening procedure.

Theorem 2. *Consider a sequence of linear models, as in* (2.1)*, that satisfy Assumptions and Conditions* (C1)*—*(C4)*. We know there exists a sequence* $\alpha_1 = \alpha_1(n,p) \to 0$ *and* $\alpha_2 = \alpha_2(n,p) \to 0$ *as* $(n,p) \to \infty$*, where* $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ *with* $\gamma = 2(L+1+b)$*,* $\alpha_2 = 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$ *with* $\gamma_{\kappa_j} = \kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p})$*, and some constant* $C_4 > 0$*, such that:*

$$P\{\hat{\mathcal{A}}^{[1]}(\alpha_1,\alpha_2) = \mathcal{A}^{[1]}\} = 1 - O(p^{-L}) \to 1 \text{ as } (n,p) \to \infty. \tag{4.2}$$

The proof of Theorem 2 can be found in Section 9. The result shows that the two-step screening procedure enjoys the model selection consistency property, and identifies the model specified in (3.1) with high probability. The significance levels that yield consistency are $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ and $\alpha_2 = 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$.

Remark 3. Condition (C3) is not needed if our goal is to obtain $P\{\hat{\mathcal{A}}^{[1]}(\alpha_1,\alpha_2) \supset \mathcal{A}^{[1]}\} = 1 - O(p^{-L})$, rather than the model selection consistency. In addition, the separation requirement in Condition (C4), $|\rho_j^{(k)}| \geq C_3\sqrt{\log p/n}\sqrt{1.01\theta_j^{(k)}}$, for all $k \notin l_j$, can be removed if we are willing to assume stronger conditions on the sample size, with an additional sample-splitting procedure (Wasserman and Roeder (2009)). To make our procedure and analysis transparent, we impose such a mild separation requirement in Theorems 1—2.

4.3. Partial faithfulness and the sure screening property

Bühlmann, Kalisch and Maathuis (2010) were the first to derive the partial faithfulness assumption, which theoretically justifies the use of a marginal correlation or a partial correlation in screening, as follows:

$$\rho_{j|S} = 0 \text{ for some } S \subseteq \{j\}^C \text{ implies } \beta_j = 0, \tag{4.3}$$

where S is the set of variables conditioned on. For independence screening, $S = \emptyset$.

Under two conditions (the positive-definiteness of Σ_X, and nonzero regression coefficients being realized from some common absolutely continuous distribution), they showed that partial faithfulness holds almost surely (Theorem 1 in Bühlmann, Kalisch and Maathuis (2010)). Because the random effect model described in Section 2 also satisfies the two conditions, the partial faithfulness condition holds almost surely in each study.

Thus, we can readily extend their Theorem 1 to a scenario with multiple studies, as follows.

Corollary 1. *Consider a sequence of linear models, as in* (2.1)*, that satisfy the* *partial faithfulness condition in each study, and are true active and inactive sets, as defined in* (2.2). *Then, the following holds for every* $j \in [p]$*:*

$$\rho_{j|S}^{(k)} = 0 \text{ for some } k \text{ for some } S \subseteq \{j\}^C \text{ implies } \beta_j = 0. \tag{4.4}$$

The proof is straightforward, and is thus omitted: if $\rho_{j|S}^{(k)} = 0$, for some study k, then with partial faithfulness, we have $\beta_j^{(k)} = 0$ for that particular k. Because we only consider features with zero or nonzero $\beta_j^{(k)}$ in all studies in (2.2), we have $\beta_j = 0$. In the case of independence screening (i.e., $S = \emptyset$), $\rho_j^{(k)} = 0$, for some k implies a zero β_j.

With the model selection consistency in Theorem 2 and the extended partial faithfulness condition in Corollary 1, the sure screening property of our two-step screening procedure follows immediately.

Corollary 2. *Consider a sequence of linear models, as in* (2.1)*, that satisfy Assumptions and Conditions* (C1)—(C4)*, as well as the extended partial faithfulness condition in Corollary* 1*. Then, there exists sequences* $\alpha_1 = \alpha_1(n,p) \to 0$ *and* $\alpha_2 = \alpha_2(n,p) \to 0$*, as* $(n,p) \to \infty$*, where* $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ *with* $\gamma = 2(L + 1 + b)$*, and* $\alpha_2 = 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$ *with* $\gamma_{\kappa_j} = \kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p})$*, such that:*

$$P\{\mathcal{A} \subseteq \hat{\mathcal{A}}^{[1]}(\alpha_1, \alpha_2)\} = 1 - O(p^{-L}) \to 1 \text{ as } (n,p) \to \infty. \tag{4.5}$$

The proof of this corollary simply combines the results of Theorem 2 and the extended partial faithfulness and, thus, is omitted.

5. Algorithms for Variable Selection with Multiple Studies

Usually, performing sure screening once may not remove enough unimportant features. In our case, because we have data from multiple studies, we expect our two-step screening procedure to remove more unimportant features than if we had data from a single study only. If the dimension is still high after applying our two-step screening procedure, we can readily extend our procedure to an iterative variable selection algorithm by testing the partial correlation with a gradually increasing size of the conditional set S. Because this method is a multiple study extension of the PC-simple algorithm in Bühlmann, Kalisch and Maathuis (2010), we call it the "Multi-PC" algorithm (Section 5.1).

On the other hand, if the dimension has been greatly reduced by the two-step screening procedure, we can simply add a second-stage group-based feature-selection technique to select the final set of variables (Section 5.2).

5.1. Multi-PC algorithm

We start from $S = \emptyset$, (i.e., our two-step screening procedure) to build a first set of candidate active variables:

$$\hat{\mathcal{A}}^{[1,1]} = \hat{\mathcal{A}}^{[1]} = \{j \in [p]; \hat{L}_j > \varphi_{\hat{\kappa}_j}^{-1}(1-\alpha_2) \text{ or } \hat{\kappa}_j = 0\}. \tag{5.1}$$

We call this set the $stage_1$ active set, where the first index in $[,]$ corresponds to the stage of our algorithm, and the second index corresponds to whether the set contains active variables ($[,1]$) or inactive variables ($[,0]$). If the dimensionality has already been decreased by a large amount, we can directly apply group-based feature selection methods, such as the group Lasso, to the remaining variables (introduced in Section 5.2).

However, if the dimension is still very high, we can reduce it further by increasing the size of S and considering partial correlations, given the variables in $\hat{\mathcal{A}}^{[1,1]}$. We follow a similar two-step procedure, but now use a partial correlation of order one instead of the marginal correlation, which yields a smaller $stage_2$ active set:

$$\hat{\mathcal{A}}^{[2,1]} = \{j \in \hat{\mathcal{A}}^{[1,1]}; \hat{L}_{j|q} > \varphi_{\hat{\kappa}_{j|q}}^{-1}(1-\alpha_2) \text{ or } \hat{\kappa}_{j|q} = 0, \text{ for all } q \in \hat{\mathcal{A}}^{[1,1]} \backslash \{j\}\}, \tag{5.2}$$

where each self-normalized estimator of the partial correlation is computed using the residuals from the regression over the variables in the conditional set.

We continue screening high-order partial correlations, resulting in a nested sequence of m active sets:

$$\hat{\mathcal{A}}^{[m,1]} \subseteq \cdots \subseteq \hat{\mathcal{A}}^{[2,1]} \subseteq \hat{\mathcal{A}}^{[1,1]}. \tag{5.3}$$

Note that the active and inactive sets at each stage are nonoverlapping, and that the union of active and inactive sets at a stage m is the active set in the previous stage $m-1$; that is, $\hat{\mathcal{A}}^{[m,1]} \cup \hat{\mathcal{A}}^{[m,0]} = \hat{\mathcal{A}}^{[m-1,1]}$. This is very similar to the original PC-simple algorithm, but we now perform the two-step procedure at each order-level. The algorithm can stop at any stage m when the dimension of $\hat{\mathcal{A}}^{[m,1]}$ drops to a low-to-moderate level, and other common group-based feature selection techniques can be used to select the final set. Alternatively, we can continue the algorithm until the candidate active set no longer changes. The algorithm is summarized as follows:

```
Algorithm 1: Multi-PC algorithm for variable selection.
  Step 1. Set $m = 1$, and perform the two-step screening procedure to construct
          the $stage_1$ active set:
              $\hat{\mathcal{A}}^{[1,1]} = \{j \in [p]; \hat{L}_j > \varphi^{-1}_{\hat{\kappa}_j}(1-\alpha_2) \text{ or } \hat{\kappa}_j = 0\}$.
  Step 2. Set $m = m + 1$. Construct the $stage_m$ active set:
              $\hat{\mathcal{A}}^{[m,1]} = \{j \in \hat{\mathcal{A}}^{[m-1,1]}; \hat{L}_{j|S} > \varphi^{-1}_{\hat{\kappa}_{j|S}}(1-\alpha_2) \text{ or } \hat{\kappa}_{j|S} = 0,$
              $\text{for all } S \subseteq \hat{\mathcal{A}}^{[m-1,1]} \backslash \{j\} \text{ with } |S| = m-1\}$.
  Step 3. Repeat Step 2 until $m = \hat{m}_{reach}$, where
          $\hat{m}_{reach} = \min\{m : |\hat{\mathcal{A}}^{[m,1]}| \leq m\}$.
```

5.2. Two-stage feature selection

As an alternative to the "Multi-PC" algorithm for variable selection, we introduce here a two-stage feature selection algorithm that combines our two-step screening procedure with other regular feature selection methods. For a single study, Fan and Lv (2008), for example, perform sure independence screening in the first stage, and then apply model selection techniques, including the adaptive Lasso, Dantzig Selector, and SCAD, which they refer to as SIS-AdaLasso, SIS-DS and SIS-SCAD, respectively.

In our case, because the feature selection is group based, we adopt a model selection technique that uses a group Lasso penalty in the second stage:

$$\min_{\beta} \sum_{k=1}^{K} \|y^{(k)} - X^{(k)}_{\hat{\mathcal{A}}^{[1]}} \beta^{(k)}_{\hat{\mathcal{A}}^{[1]}}\|_2^2 + \lambda \sum_{j \in \hat{\mathcal{A}}^{[1]}} \|\beta_j\|_2, \tag{5.4}$$

where $\hat{\mathcal{A}}^{[1]}$ is the active set identified from our two-step screening procedure, and the tuning parameter λ can be chosen using cross-validation or the BIC in practice, just as in a regular group Lasso problem. We call this two-stage feature selection algorithm TSA-SIS-groupLasso.

In addition, if the dimension drops to a moderate level at any stage while running the Multi-PC algorithm, the group Lasso-based feature selection techniques can take over to select the final set of variables.

6. Numerical Evidence

In this section, we demonstrate the advantage of the TSA-SIS procedure by comparing it with the multiple study extension of SIS (called "Min-SIS"), which ranks features by the minimum absolute correlation between all studies. We simulated data according to the linear model in (2.1), including p covariates with a zero mean and covariance matrix $\Sigma_{i,j}^{(k)} = r^{|i-j|}$, where $\Sigma_{i,j}^{(k)}$ denotes the (i,j)th entry of $\Sigma_X^{(k)}$.

In the first part of the simulation, we fixed the sample size $n = 100$, $p = 1{,}000$, and the number of studies $K = 5$, and performed $B = 1{,}000$ replications in each setting. We assume that the true active set consists of just 10 variables, and that all other variables have zero coefficients (i.e., $s_0 = 10$). The indices of nonzero coefficients are evenly spaced between 1 and p. The variance of the random error term in the linear model is fixed as 0.5^2. We randomly drew r from $\{0, 0.2, 0.4, 0.6\}$ and allowed r to vary across studies. We considered the following four settings:

1. Homogeneous weak signals across all studies: nonzero β_j generated from Unif(0.1, 0.3) and $\beta_j^{(1)} = \beta_j^{(2)} = \cdots = \beta_j^{(K)} = \beta_j$.
2. Homogeneous strong signals across all studies: nonzero β_j generated from Unif(0.7, 1) and $\beta_j^{(1)} = \beta_j^{(2)} = \cdots = \beta_j^{(K)} = \beta_j$.
3. Heterogeneous weak signals across all studies: nonzero β_j generated from Unif(0.1, 0.3) and $\beta_j^{(k)} \sim N(\beta_j, 0.5^2)$.
4. Heterogeneous strong signals across all studies: nonzero β_j generated from Unif(0.7, 1) and $\beta_j^{(k)} \sim N(\beta_j, 0.5^2)$.

We evaluated the performance of Min-SIS using receiver operating characteristic (ROC) curves, which measure the accuracy of the variable selection independently of choosing good tuning parameters (for Min-SIS, the tuning parameter is the top number of features d). The OneStep-SIS procedure is actually a special case of the Min-SIS procedure (thresholding at α_1). In presenting our TSA-SIS procedure, we fixed $\alpha_1 = 0.0001$ and $\alpha_2 = 0.05$, so the result was just one point on the sensitivity vs. 1-specificity plot. We also performed a sensitivity analysis on the two cutoffs, based on the first simulation (see Table 2), and found the two values to be optimal because they both had high sensitivity and high specificity. Thus, we suggest fixing these two values in all simulations.

Figure 1 shows the results of simulation 1—4. When the signals are homogeneously weak in all studies, as in (1), TSA-SIS clearly outperforms the Min-SIS procedure (it lies above its ROC curve). The TSA-SIS procedure reached about 90% sensitivity with controlled false positive errors (specificity $\sim$ 95%). In order to reduce false negatives, Min-SIS has to sacrifice specificity and increase the rate

Table 2. Sensitivity analysis on the choice of α_1 and α_2 in the simulation (Sensitivity/Specificity).

Sensitivity/Specificity	$\alpha_2 = 0.15$	0.05	0.01	0.001
$\alpha_1 = 0.01$	0.793/0.901	0.525/0.984	0.210/0.999	0.142/1.000
0.001	0.947/0.826	0.864/0.943	0.691/0.990	0.373/0.999
0.0001	0.966/0.816	0.922/0.932	0.840/0.985	0.681/0.998

Note: All values are based on the average results from $B = 1{,}000$ replications.

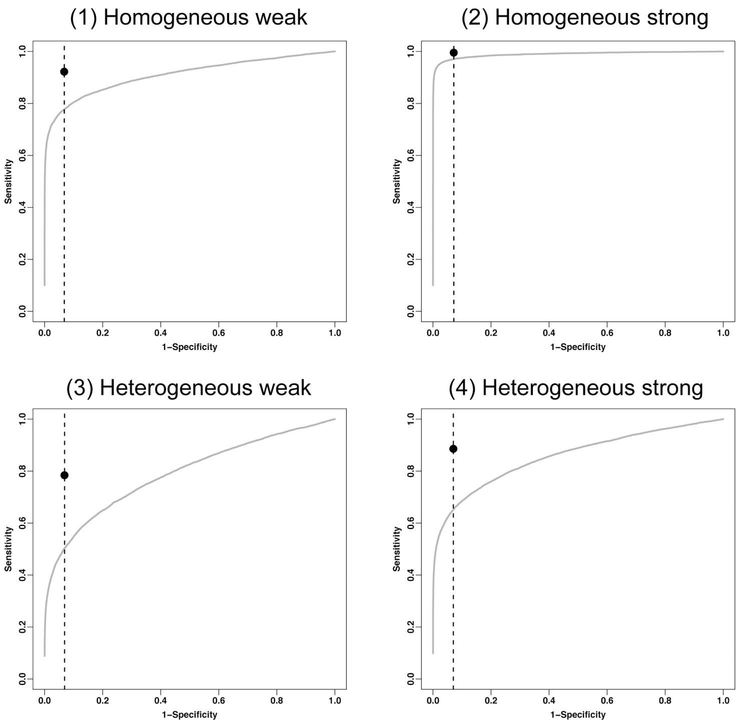

Figure 1. Simulation results 1-4: the ROC curve applies to Min-SIS, and the black point denotes our TSA-SIS using $\alpha_1 = 0.0001$ and $\alpha_2 = 0.05$.

of false positives, thus losing the benefits of performing screening (i.e. it keeps too many features). When the signals became strong, as in (2), both procedures performed equally well. This fits our motivation and theory, and shows the strength of our two-step procedure in saving weak signals, without increasing the false positive rate significantly. When the signals become heterogeneous, as in (3) and (4), both procedures perform worse than before. However, the Min-SIS procedure never outperforms the TSA-SIS procedure, because it only examines the minimum correlation between all studies, whereas the two-step procedure also considers the aggregate statistics.

7. Real Data Application

We next demonstrate our method using three microarray data sets of triple-negative breast cancer (TNBC, sometimes called basal-like), an aggressive subtype of breast cancer, usually with a poor prognosis. Previous studies have shown that the tumor suppressor protein "p53" plays an important role in breast cancer prognosis, and its expression is associated with both disease-free survival and overall survival in TNBC (Yadav, Chanana and Jhamb (2015)). Our purpose is to identify the genes most relevant and predictive to the response, namely, the expression level of the *TP53* gene, which encodes the p53 protein. The three data sets are publicly available on the authors' website or at the GEO repository, including METABRIC (a large cohort consisting of roughly 2000 primary breast tumours), GSE25066, and GSE76250 (Curtis et al. (2012); Itoh et al. (2014); Liu et al. (2016)). We filter the data to focus on TNBC cases only, which yielded 275, 178, and 165 TNBC samples for the three data sets, respectively. After routine preprocessing and filtering by including genes sufficiently expressed and with enough variation, a total of 3,377 genes remained in common for the analysis.

We applied our Multi-PC algorithm and compared the results with those of the Min-SIS method using $d = n/\log(n) = 49$ (as suggested by their paper). We used $\alpha_1 = 0.0001$ and $\alpha_2 = 0.05$ (as determined by the sensitivity analysis in the simulation); the Multi-PC algorithm ran up to the first order only (i.e., $m = 2$), and stopped with six features. This again shows the power of screening using multiple studies. After the feature selection, we fit the linear model in each study to obtain the coefficient estimates and adjusted R^2. Table 3 shows the coefficient estimates and standard errors of the final set of six genes selected by our procedure. We have added two columns to indicate whether they were also retained by the Min-SIS procedure and their relative rank, respectively. As we can see from the table, all six genes selected by our procedure were missed by Min-SIS. These genes typically had weak signals in one or more studies, and thus were very likely to be incorrectly excluded by a one-step screening procedure. Because the METABRIC study had a larger sample size, the coefficients all appear to be more significant than for the other two studies. Furthermore, the final Min-SIS model with 49 features had a much larger BIC (mean BIC = -189.17) than that of the model in our procedure with only six features (mean BIC = -313.07), showing the advantage of our model selection procedure.

The gene *EXOC1* and p53 are both components of the Ras signaling pathway, which is responsible for cell growth and division, and can ultimately lead to

Table 3. The six genes selected by our TSA-SIS procedure.

Gene	METABRIC Est (SE)	GSE25066 Est (SE)	GSE76250 Est (SE)	Min-SIS $d = 49$	Rank in Min-SIS
Intercept	7.600 (1.502)	0.213 (0.553)	−1.783 (0.971)	-	-
EXOC1	0.251 (0.081)**	0.278 (0.157)$^{\cdot}$	0.293 (0.167)$^{\cdot}$	N	164
ITGB1BP1	−0.134 (0.045)**	0.003 (0.111)	−0.178 (0.194)	N	123
RBM23	0.168 (0.078)*	0.144 (0.167)	0.367 (0.168)*	N	152
SETD3	−0.166 (0.081)*	0.366 (0.184)*	−0.080 (0.175)	N	101
SQSTM1	−0.114 (0.050)*	0.029 (0.099)	0.245 (0.183)	N	98
TRIOBP	−0.126 (0.062)*	0.084 (0.118)	0.628 (0.261)*	N	91
Adjusted-R^2	0.151	0.522	0.359		

Note: "." indicates a significance level of 0.1, "$*$" denotes a level of 0.05, "$**$" denotes a level of 0.01.

cancer (Rajalingam et al. (2007)). *RBM23* encodes for an RNA-binding protein implicated in the regulation of estrogen-mediated transcription, and has been found to be associated with p53 indirectly via a heat shock factor (Asano et al. (2016)). *ITGB1BP1* encodes for an integrin protein that is essential for cell adhesion and other downstream signaling pathways that are also modulated by p53 (Brakebusch et al. (2002)).

8. Discussion

In this paper, we proposed a two-step screening procedure for a high-dimensional regression analysis of multiple related studies. In a fairly general framework, with weaker assumptions on the signal strength, we showed that our procedure possesses the sure screening property for exponentially growing dimensionality, without requiring the normality assumption. We have shown through simulations that our procedure consistently outperforms the rank-based SIS procedure, independent of their tuning parameter d. To the best of our knowledge, our study is the first to perform variable screening in a high-dimensional regression when there are multiple related studies. In addition, we introduced two variable selection algorithms that follow the two-step screening procedure.

In our procedure, we used the self-normalized estimator of the correlation to perform the screening test in order to relax the Gaussian assumption to sub-Gaussian assumptions. This relaxation is especially beneficial compared with Fisher's Z-transformation, because in many real scenarios, the normality assumption is violated. Cai and Liu (2016) discuss the same self-normalized sample correlation as that in our procedure to perform a large-scale correlation test. In some other scenarios, self-normalized estimators enjoy the normal approximation property only under certain weak moment conditions and, thus, are more robust. We refer interested readers to Peña, Lai and Shao (2008) for more general results and examples.

Variable selection in regressions with multiple studies have been studied in a subfield of machine learning called multi-task learning (MTL). The general procedure is to apply regularization methods by including a group Lasso penalty, fused Lasso penalty, or trace norm penalty (Argyriou, Evgeniou and Pontil (2007); Zhou et al. (2012); Ji and Ye (2009)). However, at ultrahigh dimensions, such regularization methods usually fail, owing to challenges related to computation efficienciency, statistical accuracy, and algorithmic stability. Instead, sure screening can be used as a fast algorithm for preliminary feature selection, and as long as it exhibits comparable statistical performance both theoretically and empirically, its computational advantages make it a good choice (Genovese et al. (2012)). Our method provides an alternative for high-dimensional multi-task learning problems. Our scenario is related to meta-analysis, and the procedure in a broader sense can be regarded as a two-stage meta-analysis-based variable screening method. Here, we first compute the statistics for each study with initial screening and then combine the results for the aggregate test. From a variable selection point of view, our general framework is a homogeneous meta-analysis setting that assumes that the coefficients are either zeros or nonzeros in all studies. However, the magnitudes of the nonzero coefficients may still vary across studies, allowing for potential study-to-study heterogeneity.

The current two-step screening procedure is based on a linear model, but relaxes the Gaussian assumption to a sub-Gaussian distribution. We can apply a modified Fisher's Z-transformation estimator rather than our self-normalized estimator to readily accommodate general elliptical distribution families (Li, Liu and Lou (2017)). In biomedical applications, noncontinuous outcomes, such as categorical, count, or survival outcomes, are more commonly observed. Fan and Song (2010) extended the SIS and proposed a more general independent learning approach for generalized linear models by ranking the maximum marginal likelihood estimates. Fan, Feng and Song (2011) further extended the correlation learning to marginal nonparametric learning for screening in ultrahigh-dimensional additive models. Other researchers have exploited more robust measures for the correlation screening (Zhu et al. (2011); Li, Zhong and Zhu (2012); Balasubramanian, Sriperumbudur and Lebanon (2013)). These measures are all potential extensions to our method by modifying the marginal utility used in the screening procedure. In addition, the idea of performing screening with multiple studies is quite general, and is applicable to relevant statistical models other than the regression model (e.g. a Gaussian graphical model for multiple studies). We leave these interesting problems for future research.

9. Proofs

We start by introducing three technical lemmas that are essential for the proofs of the main results. By the scaling property of $\hat{T}_j^{(k)}$ and Remark 1, without loss of generality, we assume $E(X_j^{(k)}) = E(Y^{(k)}) = 0$ and $\mathrm{var}(X_j^{(k)}) = \mathrm{var}(Y^{(k)}) = 1$, for all $k \in [K]$, $j \in [p]$. Therefore in the proof we do not distinguish between $\sigma_j^{(k)}$ and $\rho_j^{(k)}$. The first lemma describes the concentration inequalities of the self-normalized covariance and $\hat{\theta}_j^{(k)}$.

Lemma 1. *Under the Assumptions* (C1) *and* (C2)*, for any $\delta \geq 2$ and $M > 0$, we have:*

(i) $P(\max_{j,k} |(\hat{\sigma}_j^{(k)} - \sigma_j^{(k)})/((\hat{\theta}_j^{(k)})^{1/2})| \geq \delta\sqrt{(\log p)/n}) = O((\log p)^{-1/2} p^{-\delta+1+b})$,

(ii) $P(\max_{j,k} |\hat{\theta}_j^{(k)} - \theta_j^{(k)}| \geq C_\theta \sqrt{(\log p)/n}) = O(p^{-M})$,

where C_θ is a positive constant depending on M_1, η, and M only.

The second and third lemmas, which are used in the proof of Theorem 2, describe the concentration behaviors of $\hat{H}_j^{(k)} := ((1/\sqrt{n}) \sum_{i=1}^n [(X_{ij}^{(k)} - \bar{X}_j^{(k)})(Y_i^{(k)} - \bar{Y}^{(k)}) - \rho_j^{(k)}])/\sqrt{\theta_j^{(k)}} = \hat{T}_j^{(k)} \sqrt{\hat{\theta}_j^{(k)}/\theta_j^{(k)}} - (\sqrt{n}\rho_j^{(k)})/\sqrt{\theta_j^{(k)}}$ and $\check{H}_j^{(k)} := ((1/\sqrt{n}) \sum_{i=1}^n (X_{ij}^{(k)} Y_i^{(k)} - \rho_j^{(k)}))/\sqrt{\theta_j^{(k)}}$.

Lemma 2. *There exists some constant $c > 0$, such that,*

$$P\left(\left|\sum_{k \in l_j} [\check{H}_j^{(k)2} - 1]\right| > t\right) \leq 2\exp\left(-c \min\left[\frac{t^2}{\kappa_j}, t^{1/2}\right]\right),$$

where c depends on M_1 and η only.

Lemma 3. *There exists some constant $C_H > 0$, such that,*

$$P\left(\max_{j,k} |\check{H}_j^{(k)} - \hat{H}_j^{(k)}| > C_H \sqrt{\frac{\log^2 p}{n}}\right) = O(p^{-M}),$$

$$P\left(\max_{j,k} |\check{H}_j^{(k)2} - \hat{H}_j^{(k)2}| > C_H \sqrt{\frac{\log^3 p}{n}}\right) = O(p^{-M}),$$

where C_H depends on M_1, η, M, and τ_0 only.

The proofs of the three lemmas are provided in the Supplementary Material.

Proof of Theorem 1. We first define the following error events:

$$\begin{aligned}
E_{j,k}^{I,\mathcal{A}^{[0]}} &= \left\{|\hat{T}_j^{(k)}| > \Phi^{-1}\left(1-\frac{\alpha_1}{2}\right) \text{ and } j \in \mathcal{A}^{[0]}, k \in l_j\right\},\\
E_{j,k}^{II,\mathcal{A}^{[0]}} &= \{|\hat{T}_j^{(k)}| \le \Phi^{-1}\left(1-\frac{\alpha_1}{2}\right) \text{ and } j \in \mathcal{A}^{[0]}, k \notin l_j\},\\
E_{j,k}^{I,\mathcal{A}^{[1]}} &= \left\{|\hat{T}_j^{(k)}| > \Phi^{-1}\left(1-\frac{\alpha_1}{2}\right) \text{ and } j \in \mathcal{A}^{[1]}, k \in l_j\right\},\\
E_{j,k}^{II,\mathcal{A}^{[1]}} &= \left\{|\hat{T}_j^{(k)}| \le \Phi^{-1}\left(1-\frac{\alpha_1}{2}\right) \text{ and } j \in \mathcal{A}^{[1]}, k \notin l_j\right\}.
\end{aligned}$$

To show Theorem 1 that $P(A) = 1 - O(p^{-L})$, it suffices to show that

$$P\left\{\bigcup_{j,k}(E_{j,k}^{I,\mathcal{A}^{[0]}} \cup E_{j,k}^{II,\mathcal{A}^{[0]}})\right\} = O(p^{-L}), \tag{9.1}$$

and

$$P\left\{\bigcup_{j,k}(E_{j,k}^{I,\mathcal{A}^{[1]}} \cup E_{j,k}^{II,\mathcal{A}^{[1]}})\right\} = O(p^{-L}). \tag{9.2}$$

We can apply Lemma 1 to bound each component in (9.1) and (9.2), with $\alpha_1 = 2\{1 - \Phi(\gamma\sqrt{\log p})\}$ and $\gamma = 2(L+1+b)$. Specifically, we obtain that,

$$\begin{aligned}
P\left(\bigcup_{j,k} E_{j,k}^{I,\mathcal{A}^{[0]}}\right) &= P\left(\max_{j\in\mathcal{A}^{[0]},k\in l_j} |\hat{T}_j^{(k)}| \ge \gamma\sqrt{\log p}\right)\\
&= O\left(\frac{1}{\sqrt{\log p}}p^{-\gamma+1+b}\right) = o(p^{-L}),
\end{aligned} \tag{9.3}$$

where the second equality follows from Lemma 1 (i) with $\delta = \gamma$, noting that $\sigma_j^{(k)} = 0$ and $\hat{T}_j^{(k)} = \sqrt{n}\hat{\sigma}_j^{(k)}/\sqrt{\hat{\theta}_j^{(k)}}$. In addition, we have that

$$\begin{aligned}
P\left(\bigcup_{j,k} E_{j,k}^{I,\mathcal{A}^{[1]}}\right) &= P\left\{\max_{j\in\mathcal{A}^{[1]},k\in l_j} |\hat{T}_j^{(k)}| \ge \gamma\sqrt{\log p}\right\}\\
&\le P\left(\max_{j\in\mathcal{A}^{[1]},k\in l_j} \left|\frac{\hat{\sigma}_j^{(k)} - \rho_j^{(k)}}{(\hat{\theta}_j^{(k)})^{1/2}}\right| \ge (\gamma - C_1)\sqrt{\frac{\log p}{n}}\right) + O(p^{-L})\\
&= O\left(\frac{1}{\sqrt{\log p}}p^{-(\gamma-C_1)+1+b}\right) + O(p^{-L})
\end{aligned}$$

$$= O(p^{-L}), \tag{9.4}$$

where the inequality on the second line is the result of Assumption (C4) on l_j for $j \in \mathcal{A}^{[1]}$, Lemma 1 (ii) with $M = L$, and Assumption (C1) $\min_{j,k} \theta_j^{(k)} \geq \tau_0$; that is, $\hat{\theta}_j^{(k)} \geq \theta_j^{(k)} - C_\theta(\log p/n)^{1/2} \geq 0.99\theta_j^{(k)}$. The equality on the third line follows from Lemma 1 (i), where $\delta = \gamma - C_1 = L + 1 + b$. Finally, we obtain that,

$$\begin{aligned}
&P\left\{\bigcup_{j,k}\left(E_{j,k}^{II,\mathcal{A}^{[0]}} \cup E_{j,k}^{II,\mathcal{A}^{[1]}}\right)\right\} \\
&= P\left(\max_{j,k\notin l_j} |\hat{T}_j^{(k)}| < \gamma\sqrt{\log p}\right) \\
&\leq P\left(\max_{j,k\notin l_j}\left|\frac{\hat{\sigma}_j^{(k)} - \rho_j^{(k)}}{(\hat{\theta}_j^{(k)})^{1/2}}\right| \geq (C_3 - \gamma)\sqrt{\frac{\log p}{n}}\right) + O(p^{-L}) \\
&= O\left(\frac{1}{\sqrt{\log p}} p^{-(C_3-\gamma)+1+b}\right) + O(p^{-L}) \\
&= O(p^{-L}),
\end{aligned} \tag{9.5}$$

where the inequality on the second line follows from Assumptions (C3) and (C4) on l_j, Lemma 1 (ii) with $M = L$, and Assumption (C1) on sub-Gaussian distributions; that is, $\hat{\theta}_j^{(k)} \leq \theta_j^{(k)} + C_\theta(\log p/n)^{1/2} \leq 1.01\theta_j^{(k)}$. In particular, we have implicitly used the fact that $\max_{j,l} \theta_j^{(k)}$ is upper bounded by a constant depending on M_1 and η only. The equality on the third line follows from Lemma 1 (i), where $\delta = C_3 - \gamma = L + 1 + b$.

Finally, we complete the proof by combining (9.3)–(9.5) to show (9.1)–(9.2).

Proof of Theorem 2. We first define the following error events:

$$\begin{aligned}
E_j^{\mathcal{A}^{[0]},2} &= \{|\hat{L}_j| > \varphi^{-1}(1-\alpha_2) \text{ or } \hat{\kappa}_j = 0\} \text{ for } j \in \mathcal{A}^{[0]}, \\
E_j^{\mathcal{A}^{[1]},2} &= \{|\hat{L}_j| < \varphi^{-1}(1-\alpha_2) \text{ and } \hat{\kappa}_j \neq 0\} \text{ for } j \in \mathcal{A}^{[1]}.
\end{aligned}$$

To prove Theorem 2, we only need to show that

$$P\left(\bigcup_{j\in\mathcal{A}^{[0]}} E_j^{\mathcal{A}^{[0]},2}\right) = O(p^{-L}) \quad \text{and} \quad P\left(\bigcup_{j\in\mathcal{A}^{[1]}} E_j^{\mathcal{A}^{[1]},2}\right) = O(p^{-L}), \tag{9.6}$$

with $\alpha_{2,\kappa_j} := 1 - \varphi_{\kappa_j}[\kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p})] := 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$.

Recall the event A defined in Theorem 1. Thus, we have that

$$P\left\{\left(\cup_{j\in\mathcal{A}^{[0]}} E_j^{\mathcal{A}^{[0]},2}\right)\bigcup\left(\cup_{j\in\mathcal{A}^{[1]}} E_j^{\mathcal{A}^{[1]},2}\right)\right\}$$

$$\leq P(A^C) + p \max_{j\in\mathcal{A}^{[0]}} P\left(\sum_{k\in l_j} \hat{T}_j^{(k)2} > \gamma_{\kappa_j}\right) + p \max_{j\in\mathcal{A}^{[1]},\kappa_j\neq 0} P\left(\sum_{k\in l_j} \hat{T}^{(k)2} < \gamma_{\kappa_j}\right).$$

Therefore, given the results in Theorem 1, it suffices to show that

$$P\left(\sum_{k\in l_j} \hat{T}^{(k)2} > \gamma_{\kappa_j}\right) = O(p^{-L-1}) \text{ for any } j \in \mathcal{A}^{[0]}, \tag{9.7}$$

and

$$P\left(\sum_{k\in l_j} \hat{T}^{(k)2} < \gamma_{\kappa_j}\right) = O(p^{-L-1}) \text{ for any } j \in \mathcal{A}^{[1]} \text{ and } \kappa_j > 0. \tag{9.8}$$

We first prove equation (9.7). Because $j \in \mathcal{A}^{[0]}$, we have $\hat{H}_j^{(k)} = \hat{T}_j^{(k)} \sqrt{\hat{\theta}_j^{(k)}/\theta_j^{(k)}}$. We are ready to bound the probability of $\sum_{k\in l_j} \hat{T}_j^{(k)2} > \gamma_{\kappa_j}$ below:

$$\begin{aligned}
&P\left(\sum_{k\in l_j} \hat{T}_j^{(k)2} > \gamma_{\kappa_j}\right)\\
&\leq P\left(\sum_{k\in l_j} \hat{H}_j^{(k)2} > \left(1 - \frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}}\right)\gamma_{\kappa_j}\right) + O(p^{-L-1})\\
&\leq P\left(\sum_{k\in l_j} (\check{H}_j^{(k)2} - 1) > \left(1 - \frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}}\right)\gamma_{\kappa_j} - \kappa_j - \kappa_j C_H \sqrt{\frac{\log^3 p}{n}}\right)\\
&\quad + O(p^{-L-1})\\
&= P\left(\sum_{k\in l_j} (\check{H}_j^{(k)2} - 1) > \kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p}) - \frac{C_\theta}{\tau_0}\sqrt{\frac{\kappa_j^2 \log p}{n}}\right.\\
&\quad \left. - \frac{C_\theta C_4}{\tau_0}\left(\sqrt{\frac{\log^5 p}{n}} + \sqrt{\frac{\kappa_j \log^2 p}{n}}\right) - \kappa_j - \kappa_j C_H \sqrt{\frac{\log^3 p}{n}}\right) + O(p^{-L-1})\\
&\leq P\left(\sum_{k\in l_j} (\check{H}_j^{(k)2} - 1) > C_2'(\log^2 p + \sqrt{\kappa_j \log p})\right) + O(p^{-L-1})\\
&= O(p^{-L-1}).
\end{aligned}$$

The inequality on the second line follows from Assumption (C1) that $\min_{j,k} \theta_j^{(k)} \geq \tau_0 > 0$, and from Lemma 1 (ii) with $M = L + 1$. The inequality on the third line follows from Lemma 3 with $M = L + 1$. The inequality on the fifth line is the result of the choice of γ_{κ_j}, with a sufficiently large $C_4 > 0$, and Assumption (C2) that $\log^3 p = o(n)$ and $\kappa_j \log^2 p = o(n)$. The last equality follows from Lemma 2.

Lastly, we prove (9.8) as follows:

$$
\begin{aligned}
&P\left(\sum_{k\in l_j} \hat{T}_j^{(k)2} < \gamma_{\kappa_j}\right) = P\left(\sum_{k\in l_j}\left(\hat{H}_j^{(k)} + \frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}}\right)^2 \frac{\theta_j^{(k)}}{\hat{\theta}_j^{(k)}} < \gamma_{\kappa_j}\right)\\
&\le P\left(\sum_{k\in l_j}\left(\hat{H}_j^{(k)} + \frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}}\right)^2 \le \left(1+\frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}}\right)\gamma_{\kappa_j}\right) + O(p^{-L-1})\\
&\le P\left(\sum_{k\in l_j}(\check{H}_j^{(k)2}-1) \le \kappa_j C_H\sqrt{\frac{\log^3 p}{n}} - \kappa_j + \left(1+\frac{C_\theta}{\tau_0}\sqrt{\frac{\log p}{n}}\right)\gamma_{\kappa_j}\right.\\
&\qquad \left. -C_m n\sum_{k\in l_j}\rho_j^{(k)2} - 2\sum_{k\in l_j}\check{H}_j^{(k)}\frac{\sqrt{n}\rho_j^{(k)}}{\sqrt{\theta_j^{(k)}}} + 2C_H\sqrt{\frac{\log^2 p}{n}}\sum_{k\in l_j}\frac{\sqrt{n}|\rho_j^{(k)}|}{\sqrt{\theta_j^{(k)}}}\right)\\
&\qquad + O(p^{-L-1}). \qquad (9.9)
\end{aligned}
$$

The first inequality follows from Assumption (C1) that $\min_{j,k}\theta_j^{(k)} \ge \tau_0 > 0$, and Lemma 1 (ii) with $M = L+1$. The last inequality follows from Lemma 3 (both equations) and $\min_{j,k}(\theta_j^{(k)})^{-1} := C_m > 0$, guaranteed by the sub-Gaussian assumption in Assumption (C1).

We can upper bound the term $2C_H\sqrt{(\log^2 p)/n}\sum_{k\in l_j}(\sqrt{n}|\rho_j^{(k)}|)/(\sqrt{\theta_j^{(k)}})$ in (9.9) as follows:

$$
\begin{aligned}
2C_H\sqrt{\frac{\log^2 p}{n}}\sum_{k\in l_j}\frac{\sqrt{n}|\rho_j^{(k)}|}{\sqrt{\theta_j^{(k)}}} &\le 2C_H\sqrt{\frac{\log^2 p}{n}}\frac{\sqrt{n}}{\sqrt{\tau_0}}\sqrt{\kappa_j}\sqrt{\sum_{k\in l_j}\rho_j^{(k)2}}\\
&= o\left(\sqrt{n\sum_{k\in l_j}\rho_j^{(k)2}}\right). \qquad (9.10)
\end{aligned}
$$

The first inequality follows from the Cauchy—Schwarz inequality and Assumption (C1), and the second equality follows from (C2) that $\kappa_j \log^2 p = o(n)$.

We next upper bound the term $-2\sum_{k\in l_j}\check{H}_j^{(k)}(\sqrt{n}\rho_j^{(k)})/\sqrt{\theta_j^{(k)}}$ with a high probability. Note that $\theta_j^{(k)}$ is bounded below and above; that is, $\tau_0 \le \theta_j^{(k)} \le C_m^{-1}$ by Assumption (C1). In addition, $\check{H}_j^{(k)}$ has a zero mean and is sub-exponential with bounded constants, by Assumption (C1). From the Bernstein inequality

(Proposition 5.16 in Vershynin (2010)), we have, with some constant $c' > 0$,

$$\begin{aligned}
&P\left(|2\sum_{k\in l_j}\left|\check{H}_j^{(k)}\frac{\sqrt{n}|\rho_j^{(k)}|}{\sqrt{\theta_j^{(k)}}}\right| > t\right)\\
&\quad\le 2\exp\left(-c'\min\left[\frac{t^2}{n\sum_{k\in l_j}\rho_j^{(k)2}}\right], \frac{t}{\max_{k\in l_j}\sqrt{n}|\rho_j^{(k)}|}\right). \qquad (9.11)
\end{aligned}$$

We select $t = C_B\sqrt{n\sum_{k\in l_j}\rho_j^{(k)2}\log^2 p}$ with a large constant C_B in the inequality above, and apply (9.10) to reduce (9.9), as follows:

$$\begin{aligned}
&P\left(\sum_{k\in l_j}\hat{T}_j^{(k)2} < \gamma_{\kappa_j}\right)\\
&\le P\Bigg(\sum_{k\in l_j}(\check{H}_j^{(k)2} - 1) \le -C_m n\sum_{k\in l_j}\rho_j^{(k)2} + 2C_B\sqrt{n\sum_{k\in l_j}\rho_j^{(k)2}\log^2 p}\\
&\qquad + 2C_4\sqrt{\kappa_j\log p} + 2C_4\log^2 p\Bigg) + O(p^{-L-1})\\
&\le P\Bigg(\sum_{k\in l_j}(\check{H}_j^{(k)2} - 1) \le -C_m C_2(\log^2 p + \sqrt{\kappa_j\log p})\\
&\qquad + 2C_B\sqrt{C_2\log^2 p(\log^2 p + \sqrt{\kappa_j\log p})}\\
&\qquad + 2C_4\sqrt{\kappa_j\log p} + 2C_4\log^2 p\Bigg) + O(p^{-L-1})\\
&\le P\Bigg(\sum_{k\in l_j}(\check{H}_j^{(k)2} - 1) \le -C_2'(\log^2 p + \sqrt{\kappa_j\log p})\Bigg) + O(p^{-L-1})\\
&= O(p^{-L-1}).
\end{aligned}$$

The inequality on the first line is obtained by the choice of γ_{κ_j}, with the chosen $C_4 > 0$ and Assumption (C2) that $\kappa_j\log^2 p = o(n)$. The inequalities on the second and third lines follows from Assumption (C4) that $\sum_{k\in l_j}|\rho_j^{(k)}|^2 \ge (C_2(\log^2 p + \sqrt{\kappa_j\log p}))/n$, for a sufficiently large $C_2 > 0$. The last equality follows from Lemma 2.

This completes the proof of (9.7) and (9.8), which yields

$$P\Big\{\Big(\cup_{j\in\mathcal{A}^{[0]}} E_j^{\mathcal{A}^{[0]},2}\Big)\bigcup\Big(\cup_{j\in\mathcal{A}^{[1]}} E_j^{\mathcal{A}^{[1]},2}\Big)\Big\} = O(p^{-L}),$$

with the results from Theorem 1. Therefore we have completed the proof of

Theorem 2.

Supplementary Material

The online Supplementary Material contains the proofs of the three lemmas.

Acknowledgments

Z.R. was supported by the National Science Foundation under award number DMS-1812030, and T.M. and G.C.T. were supported by the NCI of the National Institutes of Health under award number R01CA190766.

References

Argyriou, A., Evgeniou, T. and Pontil, M. (2007). Multi-task feature learning. In: *Advances in Neural Information Processing Systems*, 41–48.

Asano, Y., Kawase, T., Okabe, A., Tsutsumi, S., Ichikawa, H., Tatebe, S., Kitabayashi, I., Tashiro, F., Namiki, H., Kondo, T. and et al. (2016). IER5 generates a novel hypophosphorylated active form of HSF1 and contributes to tumorigenesis. *Scientific Reports* **6**, 19174.

Balasubramanian, K., Sriperumbudur, B. and Lebanon, G. (2013). Ultrahigh dimensional feature screening via rkhs embeddings. In: *Artificial Intelligence and Statistics*. 126–134.

Brakebusch, C., Bouvard, D., Stanchi, F., Sakai, T. and Fassler, R. (2002). Integrins in invasive growth. *The Journal of Clinical Investigation* **109**, 999–1006.

Bühlmann, P., Kalisch, M. and Maathuis, M. H. (2010). Variable selection in high-dimensional linear models: partially faithful distributions and the PC-simple algorithm. *Biometrika* **97**, 261–278.

Cai, T. T. and Liu, W. (2016). Large-scale multiple testing of correlations. *Journal of the American Statistical Association* **111**, 229–240.

Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *The Annals of Statistics* **41**, 2123–2148.

Chang, J., Tang, C. Y. and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *The Annals of Statistics* **44**, 515–539.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y. and et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352.

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

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

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

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

Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research* **10**, 2013–2038.

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

Genovese, C. R., Jin, J., Wasserman, L. and Yao, Z. (2012). A comparison of the lasso and marginal regression. *Journal of Machine Learning Research* **13**, 2107–2143.

Huang, J., Breheny, P. and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **27**, 481–499.

Itoh, M., Iwamoto, T., Matsuoka, J., Nogami, T., Motoki, T., Shien, T., Taira, N., Niikura, N., Hayashi, N., Ohtani, S. and et al. (2014). Estrogen receptor (er) mrna expression and molecular subtype distribution in er-negative/progesterone receptor-positive breast cancers. *Breast Cancer Research and Treatment* **143**, 403–409.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization. In: *Proceedings of the 26th Annual International Conference on Machine Learning*, 457–464. ACM.

Jiang, J., Li, C., Paul, D., Yang, C. and Zhao, H. (2016). On high-dimensional misspecified mixed model analysis in genome-wide association study. *The Annals of Statistics* **44**, 2127–2160.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics* **5**, 994–1019.

Li, Q., Wang, S., Huang, C.-C., Yu, M. and Shao, J. (2014). Meta-analysis based variable selection for gene expression data. *Biometrics* **70**, 872–880.

Li, R., Liu, J. and Lou, L. (2017). Variable selection via partial correlation. *Statistica Sinica* **27**, 983–996.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139.

Liang, F., Song, Q. and Qiu, P. (2015). An equivalent measure of partial correlation coefficients for high-dimensional Gaussian graphical models. *Journal of the American Statistical Association* **110**, 1248–1265.

Liu, Y.-R., Jiang, Y.-Z., Xu, X.-E., Hu, X., Yu, K.-D. and Shao, Z.-M. (2016). Comprehensive transcriptome profiling reveals multigene signatures in triple-negative breast cancer. *Clinical Cancer Research* **22**, 1653–1662.

Luo, S., Song, R. and Witten, D. (2014). Sure screening for Gaussian graphical models. *arXiv preprint arXiv:1407.7819.*

Ma, S., Huang, J. and Song, X. (2011). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* **12**, 763–775.

Ma, S., Li, R. and Tsai, C.-L. (2017). Variable screening via quantile partial correlation. *Journal of the American Statistical Association* **112**, 1–14.

Meier, L., Van De Geer, S. and Bühlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 53–71.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics* **2**, 605–633.

Peña, V. H., Lai, T. L. and Shao, Q.-M. (2008). *Self-Normalized Processes: Limit Theory and Statistical Applications.* Springer Science & Business Media.

Rajalingam, K., Schreck, R., Rapp, U. R. and Albert, S. (2007). Ras oncogenes and their downstream targets. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* **1773**, 1177–1195.

Shao, Q.-M. (1999). A Cramér type large deviation result for Student's t-statistic. *Journal of Theoretical Probability* **12**, 385–398.

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

Tseng, G. C., Ghosh, D. and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research* **40**, 3785–3799.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027.*

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics* **37**, 2178–2201.

Yadav, B. S., Chanana, P. and Jhamb, S. (2015). Biomarkers in triple negative breast cancer: A review. *World Journal of Clinical Oncology* **6**, 252–263.

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

Zhou, J., Liu, J., Narayan, V. A. and Ye, J. (2012). Modeling disease progression via fused sparse group lasso. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1095–1103. ACM.

Zhu, L.-P., Li, L., Li, R. and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

Department of Epidemiology and Biostatistics, University of Maryland, College Park, MD 20742, USA.

E-mail: tma0929@umd.edu

Department of Statistics, University of Pittsburgh, Pittsburgh, PA 15261, USA.

E-mail: zren@pitt.edu

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA.

E-mail: ctseng@pitt.edu

(Received October 2017; accepted June 2018)