High-dimensional variable screening: from single study to multiple studies

Tianzhou (Charles) Ma, Ph.D.

SCHOOL OF
PUBLIC HEALTH

Department of Epidemiology and Biostatistics

Bio3 Biostatistics Colloquium, Georgetown University
Dec 3rd, 2020

The problem of variable selection

- Variable selection or subset selection is one of the most pervasive model selection problems in statistical and machine learning applications.
- Suppose Y is the outcome variable, $X_1, \ldots, X_p$ is a set of potential predictors, consider a linear model:

$$Y = \sum_{j=1}^{p} \beta_j X_j + \epsilon,$$

  the purpose of variable selection is to find a small set of true predictors that contribute to the response (i.e. with nonzero $\beta_j's$).
- Traditionally, when p is small, classical model selection methods using criteria like AIC, BIC or Mallow's C_p, etc. can be applied.

Variable selection in the old days vs. Big Data era

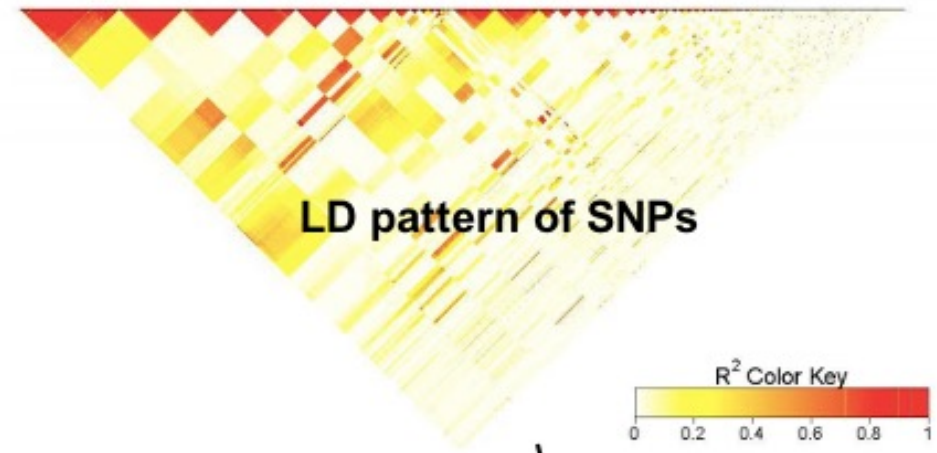
Variable selection in the old days vs. Big Data era

High-dimensional variable selection in biomedical fields

- High dimensional data with much greater number of features than sample size ($p >> n$) has become rule rather than exception nowadays in biomedical fields:
  - Genomics: selection of DNA variants or genes responsible for disease;
  - Neuroimaging: selection of voxels predictive of clinical outcome from brain scans, etc.
- Besides, data from the real world is usually more perplexing than math ...

Variable selection is complex in genomics

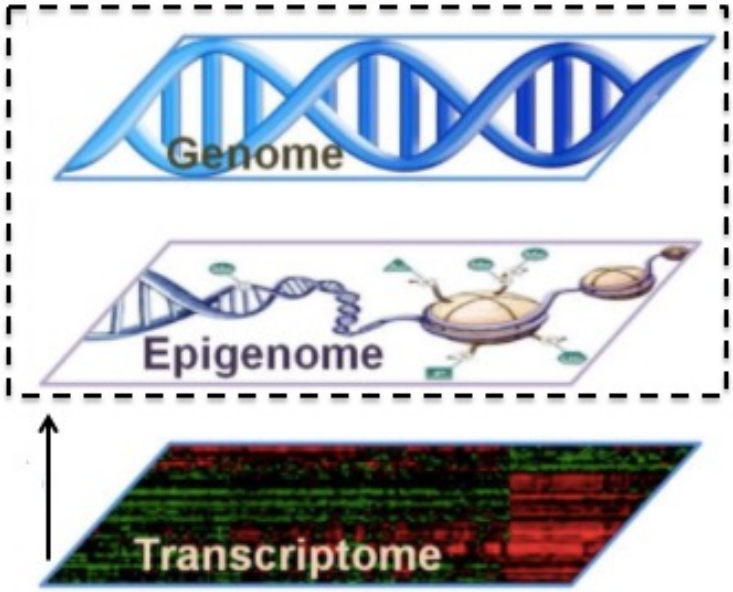
Complex inter-feature structure



$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2}\sum_{s=1}^{S}(y_s - \boldsymbol{x}_s^T\boldsymbol{\beta})^2 + \lambda \sum_{\delta_{ij}=1} |\beta_i - \beta_j| + \gamma\cdot\lambda\sum_{i=1}^{n}|\beta_i|,$$

Mo et al. (2020+). *Genetic Fine-mapping with Dense Linkage Disequilibrium Blocks.*

Multi-omics

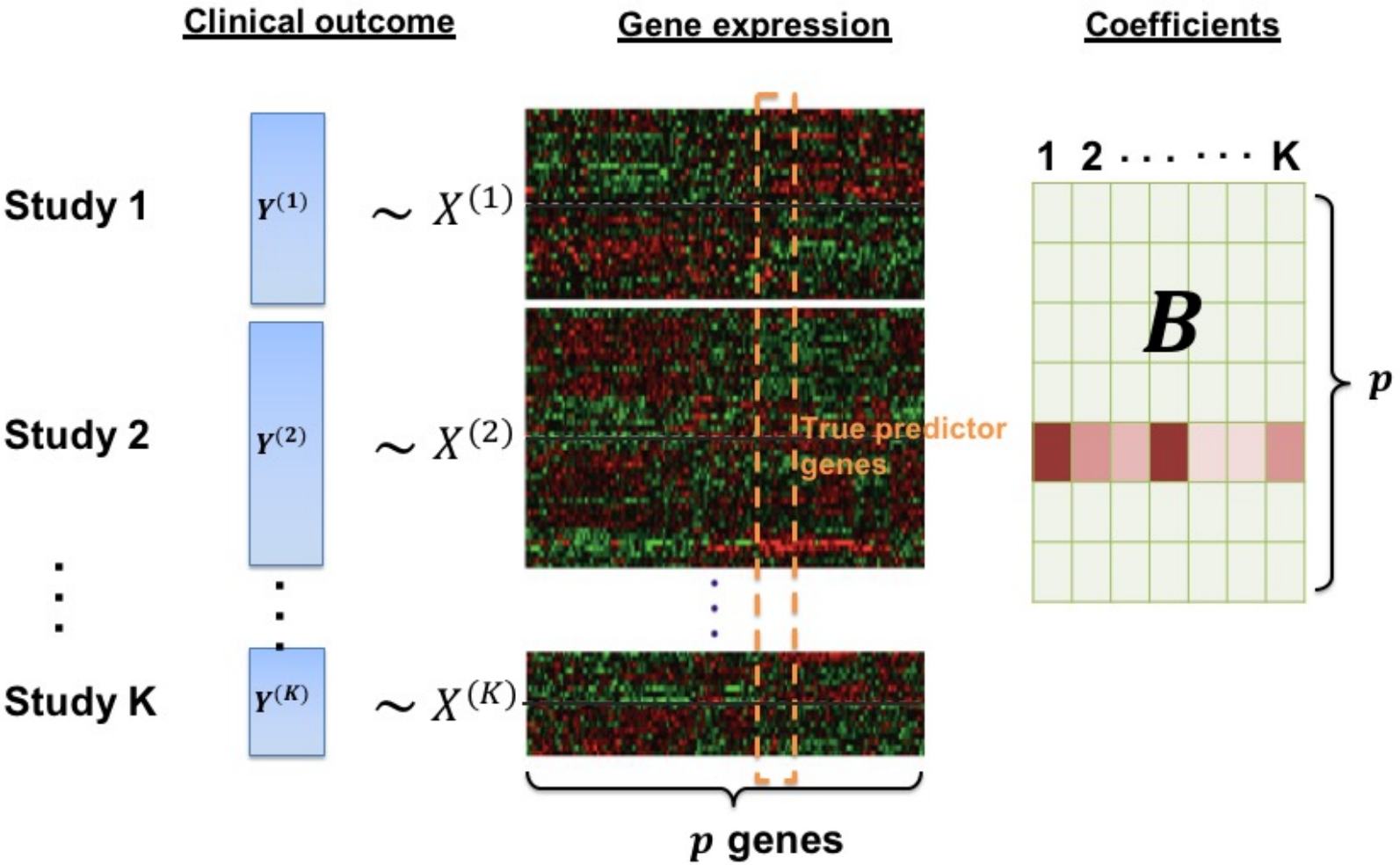


$$\mathbf{G} = \mathbf{G}^M + \mathbf{G}^{\overline{M}}, \mathbf{G}^M = M\Omega,$$

$$\mathbf{Y} = \mathbf{C}\boldsymbol{\gamma}^C + \mathbf{G}^M\boldsymbol{\gamma}^M + \mathbf{G}^{\overline{M}}\boldsymbol{\gamma}^{\overline{M}} + \epsilon,$$

Fang et al. (2019). *Bioinformatics.*
Zhu et al. (2019). *AoAS.*

HighD variable selection from single to multiple studies



HighD variable selection with multiple studies

- Q1: What's good about meta-analysis?
  - Improve power and reproducibility.
- Q2: Can we borrow information across studies to assist variable selection?
  - Hopefully yes.
- Q3: Does there exist between study heterogeneity?
  - Most likely yes.

Table of contents

1. Introduction to sure screening
2. Screening with multiple studies framework and TSA-SIS method
3. Simulation and real data results
4. Extension to survival analysis: CoxTOTEM

Regularization methods for HighD variable selection

- In the past two decades, many regularization methods, including Lasso, SCAD, elastic net, etc. and their numerous variants, have been developed for HighD variable selection. For linear regression model, the problem can be summarized in the following form:

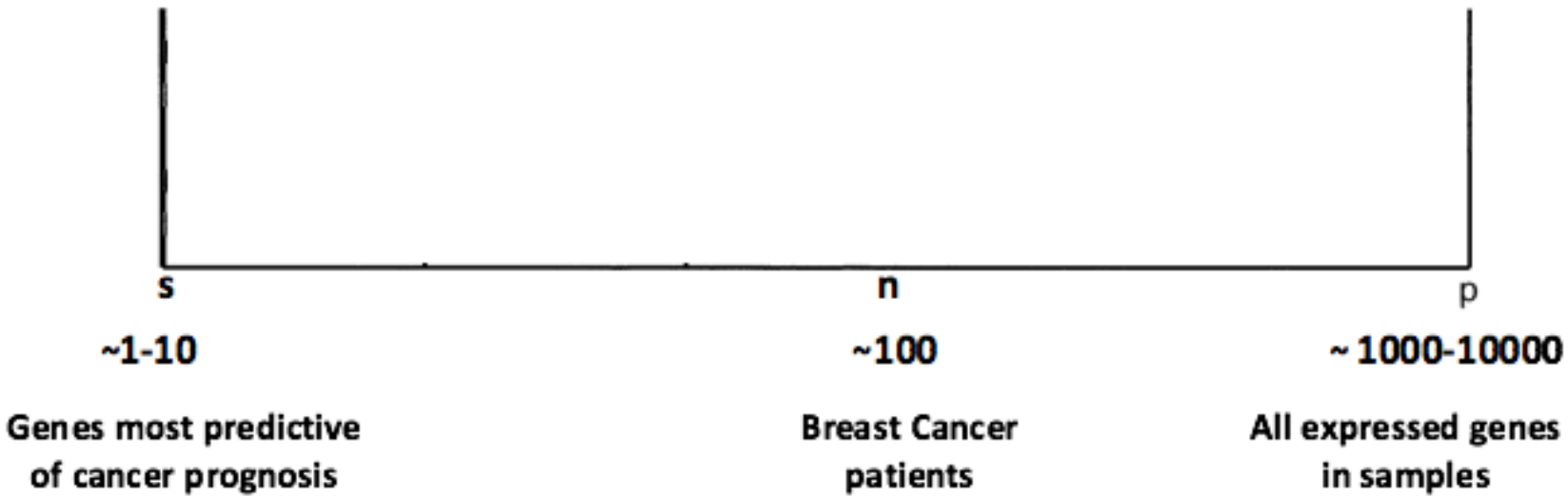
$$\min_{\beta}\{\frac{1}{2n}||Y - X\beta||^2 + \sum_{j=1}^{p} P_\lambda(|\beta_j|)\},$$

  where $P_\lambda(.)$ is the penalty of a specific form, e.g. L_1 norm for Lasso.
- These methods work pretty well when n is large and p not too large. However, as p grows at an exponential rate of n (as seen in most applications), they will fail due to challenges in computation expediency, statistical accuracy and algorithmic stability.

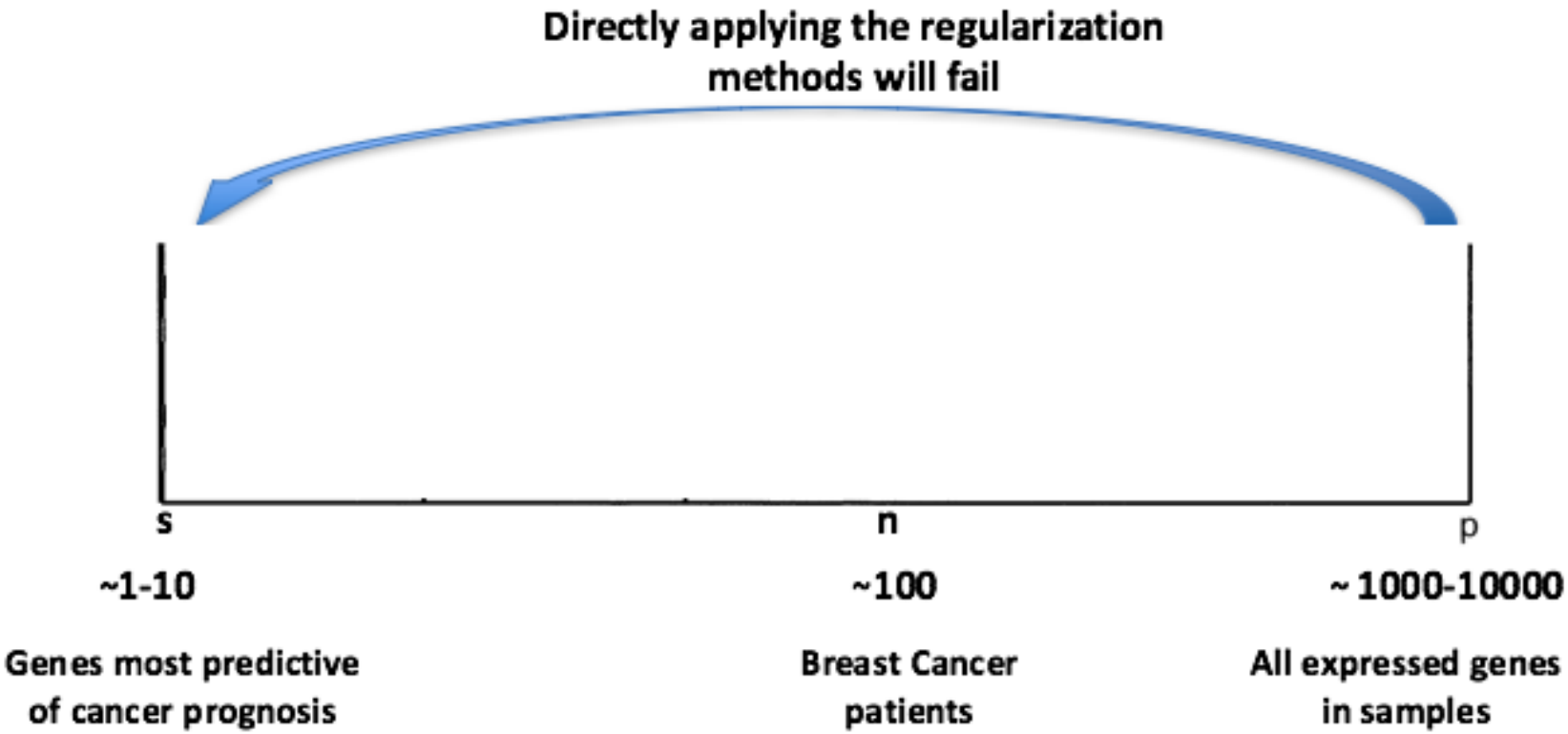
Introduction to sure screening

- Alternatively, a natural way to tackle such a HighD variable selection problem is to consider first reducing dimension to low or moderate and then performing regularization.
- Fan & Lv (2008) first proposed a Sure Independent Screening (SIS) method using the marginal correlation of the features with the response to rank features and keep only top ones which still include the true predictors with large probability ("sure screening property").
- Later, Buhlmann et al. (2010) provided the key theoretical basis of SIS, the partial faithfulness assumption, which states that a zero marginal correlation will imply a zero regression coefficient.

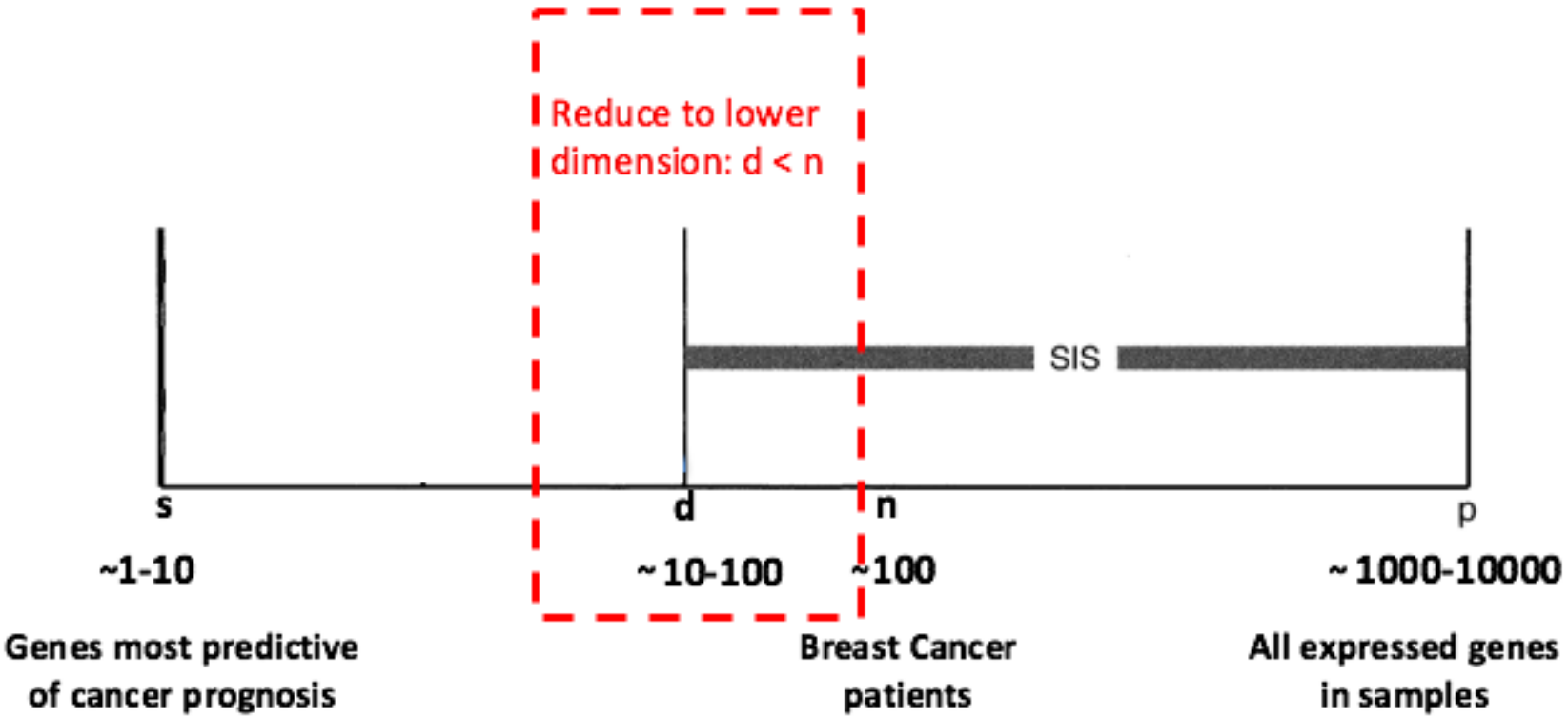
The idea of SIS



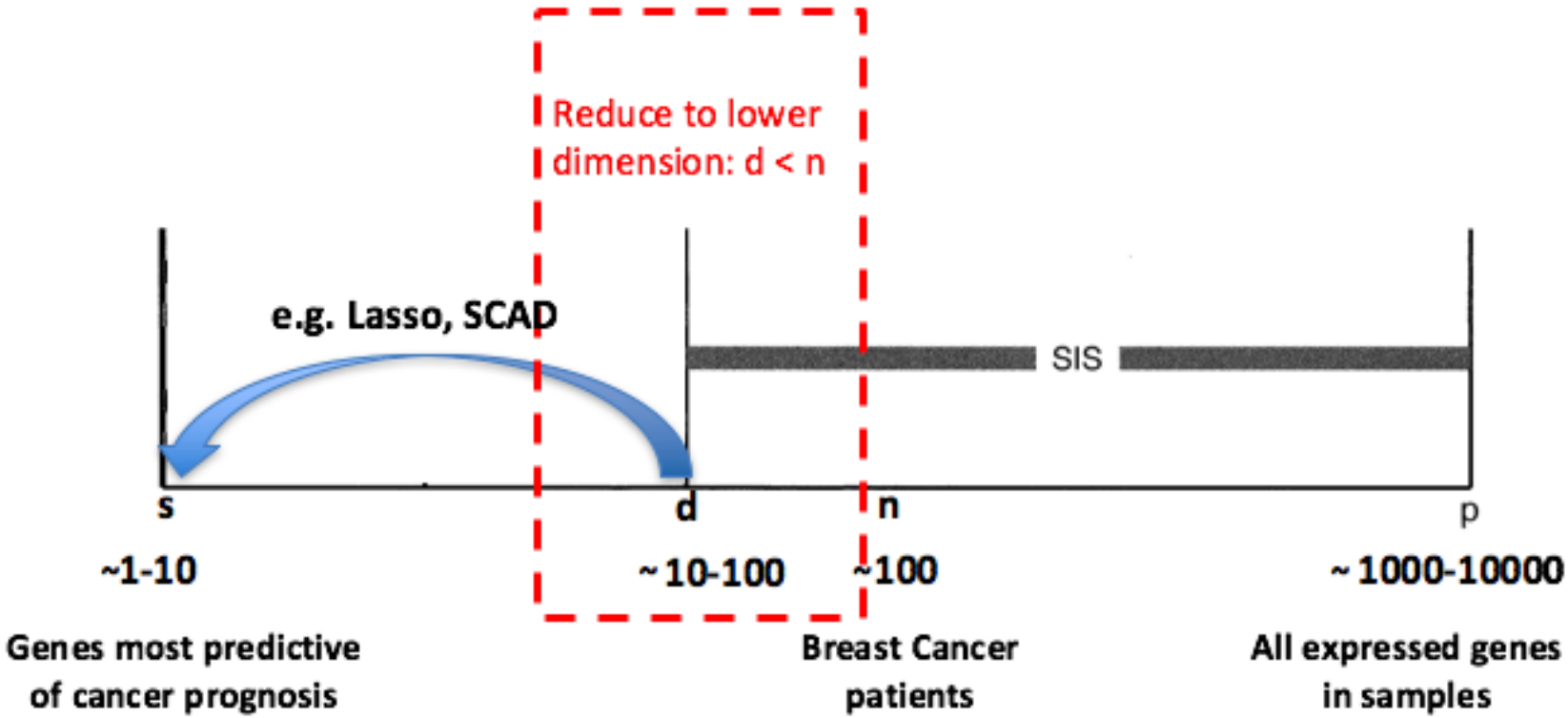
The idea of SIS



The idea of SIS



The idea of SIS



Overview of screening methods

Selected papers	**Key ideas**
Fan & Lv (2008)	Sure independent screening (SIS) for linear model
Fan et al. (2009); Fan & Song (2010)	Sure independent screening for generalized linear model
Buhlmann et al. (2010)	PC-simple algorithm for variable selection in linear model (essentially SIS, but use testing instead of ranking)
Zhu et al. (2011); Li et al. (2012)	Feature screening using more robust measures
Fan et al. (2011)	Nonparametric independence screening for generalized additive model
Mai & Zou (2013); Mai & Zou (2015)	Nonparametric robust screening for classification and regression
Zhao & Li (2013); Song et al. (2014)	Sure screening for censored/survival data
Luo et al. (2014); Liang et al. (2015)	Sure screening for Gaussian graphical model
Ma et al. (2017)	Sure screening for quantile linear regression

Overview of screening methods

Selected papers	Key ideas
Fan & Lv (2008)	Sure independent screening (SIS) for linear model
Fan et al. (2009); Fan & Song (2010)	Sure independent screening for generalized linear model
Buhlmann et al. (2010)	PC-simple algorithm for variable selection in linear model (essentially SIS, but use testing instead of ranking)
Zhu et al. (2011); Li et al. (2012)	Feature screening using more robust measures
Fan et al. (2011)	Nonparametric independence screening for generalized additive model
Mai & Zou (2013); Mai & Zou (2015)	Nonparametric robust screening for classification and regression
Zhao & Li (2013); Song et al. (2014)	Sure screening for censored/survival data
Luo et al. (2014); Liang et al. (2015)	Sure screening for Gaussian graphical model
Ma et al. (2017)	Sure screening for quantile linear regression

- All the methods look at data from only one single study.

Sure screening with multiple studies

- Data from more and more studies target at the same scientific question, effective integration of information from multiple studies can potentially improve the performance of highD variable selection of single study.

Sure screening with multiple studies

- Data from more and more studies target at the same scientific question, effective integration of information from multiple studies can potentially improve the performance of highD variable selection of single study.
- New problem: variable screening with multiple related studies?

Sure screening with multiple studies

- Data from more and more studies target at the same scientific question, effective integration of information from multiple studies can potentially improve the performance of highD variable selection of single study.
- New problem: variable screening with multiple related studies?
- General framework + Two-step screening procedure

A general framework for screening with multiple studies

- Consider a linear model in each study $k \in \{1, \ldots, K\}$:

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

A general framework for screening with multiple studies

- Consider a linear model in each study $k \in \{1, \ldots, K\}$:

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

- Assume β's are either zero or non-zero in all studies, i.e. the true active set $\mathcal{A}^* = \{j \in [p]; \beta_j^{(k)} \neq 0 \text{ for all } k\}$, but the magnitudes of nonzero β's can differ across studies allowing for potential study to study heterogeneity.

A general framework for screening with multiple studies

- Consider a linear model in each study $k \in \{1, \ldots, K\}$:

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

- Assume β's are either zero or non-zero in all studies, i.e. the true active set $\mathcal{A}^* = \{j \in [p]; \beta_j^{(k)} \neq 0 \text{ for all } k\}$, but the magnitudes of nonzero β's can differ across studies allowing for potential study to study heterogeneity.
- With partial faithfulness, marginal correlation $\rho_j^{(k)} = 0$ implies $\beta_j^{(k)} = 0$ in each k. Based on our assumption, we will exclude jth feature whenever $\rho_j^{(k)} = 0$ for any k. This actually helps screen out more features.

A general framework for screening with multiple studies

- Consider a linear model in each study $k \in \{1, \ldots, K\}$:

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

- Assume β's are either zero or non-zero in all studies, i.e. the true active set $\mathcal{A}^* = \{j \in [p]; \beta_j^{(k)} \neq 0 \text{ for all } k\}$, but the magnitudes of nonzero β's can differ across studies allowing for potential study to study heterogeneity.
- With partial faithfulness, marginal correlation $\rho_j^{(k)} = 0$ implies $\beta_j^{(k)} = 0$ in each k. Based on our assumption, we will exclude jth feature whenever $\rho_j^{(k)} = 0$ for any k. This actually helps screen out more features.
- Thus, we aim to keep the following set after performing sure screening with multiple studies: $\mathcal{A}^{[1]} = \{j \in [p]; \min_k |\rho_j^{(k)}| \neq 0\}$.

One step procedure ("OneStep-SIS") may lead to false negative errors

- One way to estimate $\mathcal{A}^{[1]}$ is to test $H_0 : \rho_j^{(k)} = 0$ for each k, whenever a test is not rejected, we will exclude jth feature from $\hat{\mathcal{A}}^{[1]}$.

One step procedure ("OneStep-SIS") may lead to false negative errors

- One way to estimate $\mathcal{A}^{[1]}$ is to test $H_0 : \rho_j^{(k)} = 0$ for each k, whenever a test is not rejected, we will exclude jth feature from $\hat{\mathcal{A}}^{[1]}$.
- However, in reality, it is possible for important features to have weak signals thus small $|\rho_j^{(k)}|$ in at least one study. These features might be incorrectly screened out causing the false negative errors.

One step procedure ("OneStep-SIS") may lead to false negative errors

- One way to estimate $\mathcal{A}^{[1]}$ is to test $H_0 : \rho_j^{(k)} = 0$ for each k, whenever a test is not rejected, we will exclude jth feature from $\hat{\mathcal{A}}^{[1]}$.
- However, in reality, it is possible for important features to have weak signals thus small $|\rho_j^{(k)}|$ in at least one study. These features might be incorrectly screened out causing the false negative errors.
- In screening, false negative is more serious error than false positive since we may include a second stage variable selection to further refine the set and help reduce false positive errors.

Two-Step Aggregation Sure Independence Screening ("TSA-SIS")

Step 1. Perform screening test in each study and collect study set with potential zero correlations:

$$\hat{I}_j = \{k; |\hat{T}_j^{(k)}| \leq \Phi^{-1}(1 - \alpha_1/2)\},$$

where $\hat{T}_j^{(k)}$ is the self-normalized estimator of covariance between $X_j^{(k)}$ and $Y^{(k)}$ serving as our test statistics.

Step 2. Test the aggregate effect of potential zero correlations, if strong, we will keep the feature, i.e.:

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

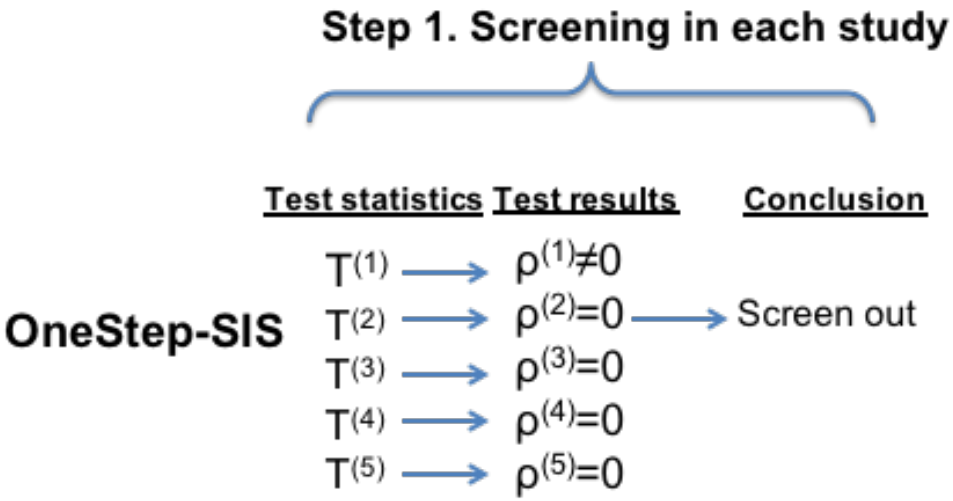
where the aggregate $\hat{L}_j = \sum_{k \in \hat{I}_j} (\hat{T}_j^{(k)})^2$ follows a $\chi^2_{\hat{\kappa}_j}$ with $\hat{\kappa}_j = |\hat{I}_j|$.

Two-Step Aggregation Sure Independence Screening ("TSA-SIS")

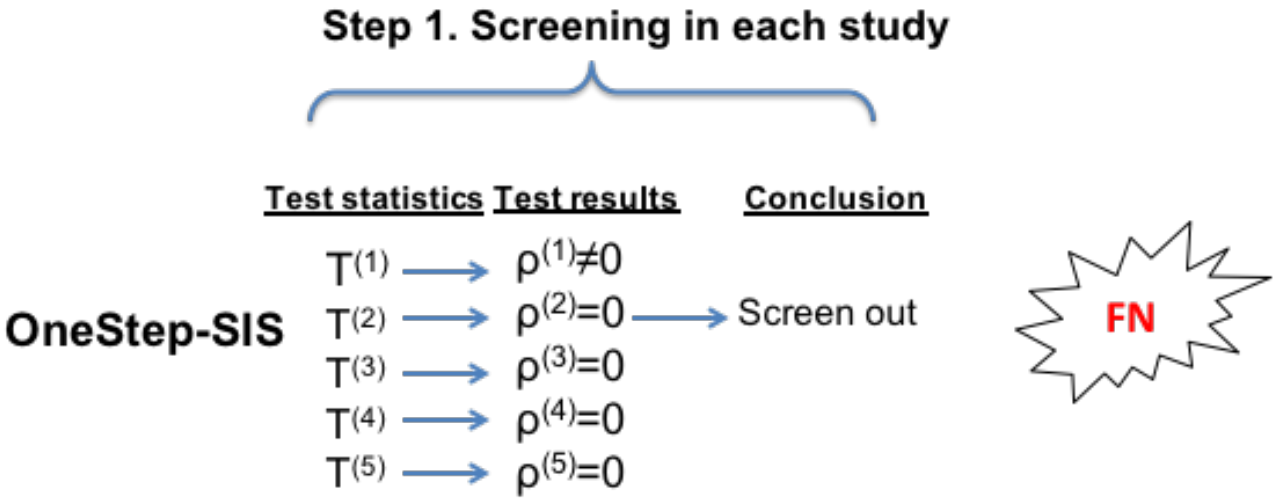
Advantages:

1. low false negative errors as compared to one-step procedure.
2. self-normalized estimator relaxes the Gaussian assumption.
3. sure screening property with weaker assumptions on signal strengths.

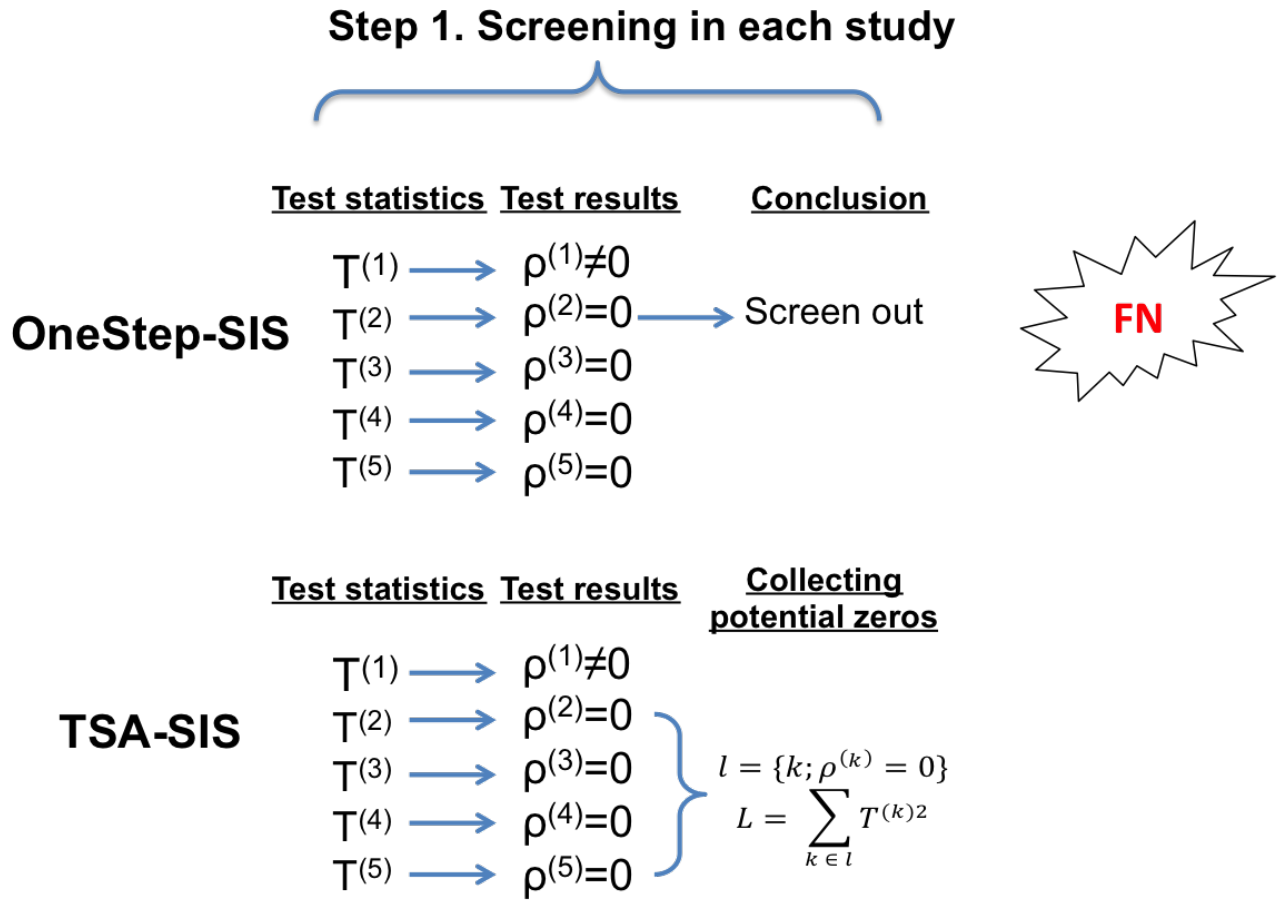
OneStep-SIS vs. TSA-SIS for weak signals



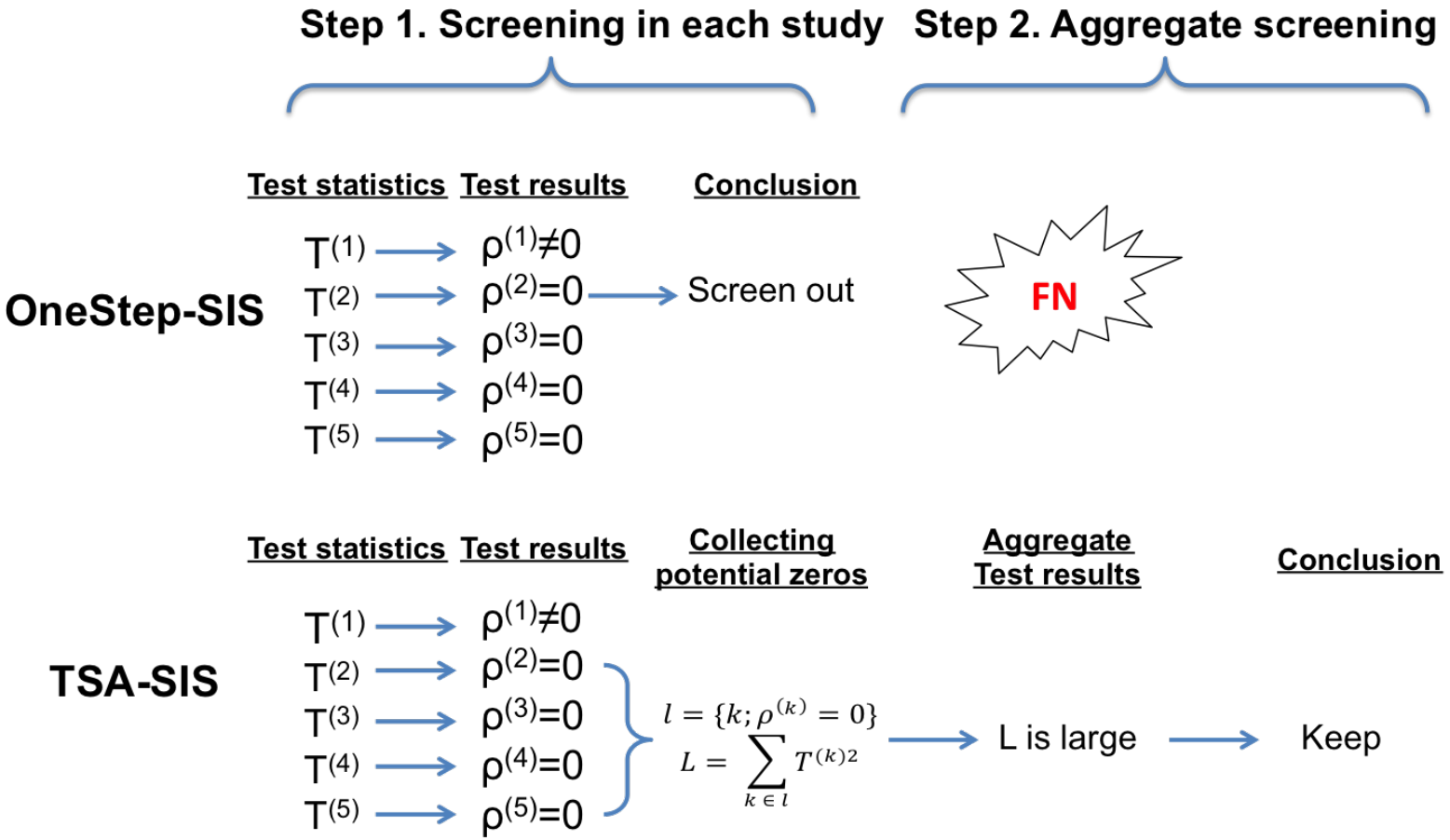
OneStep-SIS vs. TSA-SIS for weak signals



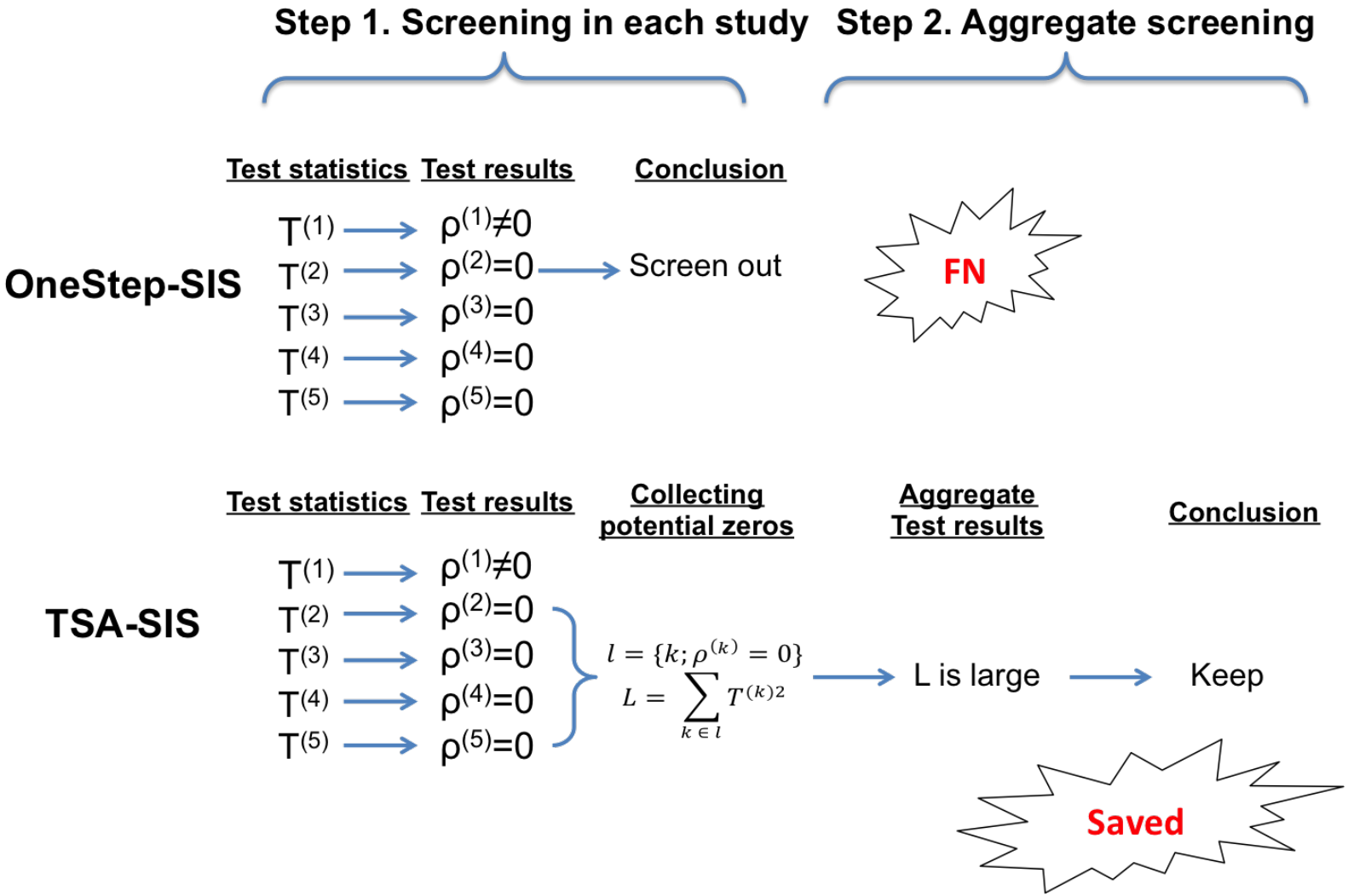
OneStep-SIS vs. TSA-SIS for weak signals



OneStep-SIS vs. TSA-SIS for weak signals



OneStep-SIS vs. TSA-SIS for weak signals



Conditions

(C1) (Sub-Gaussian Condition) There exist some 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 some $\tau_0 > 0$ 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 $j \in \mathcal{A}^{[0]}$, $I_j(j \in \mathcal{A}^{[0]}) = \{k; \rho_j^{(k)} = 0\}$ and $\kappa_j = |I_j|$. If $k \notin I_j$, then $|\rho_j^{(k)}| \geq C_3 \sqrt{\frac{\log p}{n}} \sqrt{1.01\theta_j^{(k)}}$, where $C_3 = 3(L + 1 + b)$.

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

Consistency and sure screening property

Theorem 1: Consistency

Consider a sequence of linear models which 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 $\gamma = 2(L + 1 + b)$ and $\alpha_2 = 1 - \varphi_{\kappa_j}(\gamma_{\kappa_j})$ with $\gamma_{\kappa_j} = \kappa_j + C_4(\log^2 p + \sqrt{\kappa_j \log p})$ and some constant $C_4 > 0$ such that:

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

With partial faithfulness, the sure screening property immediately follows:

$$P\{\mathcal{A}_n^* \subseteq \hat{\mathcal{A}}_n^{[1]}(\alpha_{1,n}, \alpha_{2,n})\} = 1 - O(p^{-L}).$$

Guideline on the choice of α_1 and α_2

- The procedure proposed here involves two tuning parameters: α_1 and α_2.

Guideline on the choice of α_1 and α_2

- The procedure proposed here involves two tuning parameters: α_1 and α_2.
- Since the actual screening test is performed in the second step, commonly used significance level such as $\alpha_2 = 0.05$ is recommended to reduce false negative errors following Buhlmann et al. (2010).

Guideline on the choice of α_1 and α_2

- The procedure proposed here involves two tuning parameters: α_1 and α_2.
- Since the actual screening test is performed in the second step, commonly used significance level such as $\alpha_2 = 0.05$ is recommended to reduce false negative errors following Buhlmann et al. (2010).
- In general, there is a tradeoff between false negative errors and false positive error determined by the choice of α_1.

Guideline on the choice of α_1 and α_2

- The procedure proposed here involves two tuning parameters: α_1 and α_2.
- Since the actual screening test is performed in the second step, commonly used significance level such as $\alpha_2 = 0.05$ is recommended to reduce false negative errors following Buhlmann et al. (2010).
- In general, there is a tradeoff between false negative errors and false positive error determined by the choice of α_1.
- To further reduce the serious false negative errors in screening, we recommend using a small α_1 (e.g. 1e-4) in practical application.

Sensitivity analysis on the choice of α_1 and α_2

Sensitivity/Specificity	$\alpha_2 = 0.15$	0.05	0.01	0.001
α_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

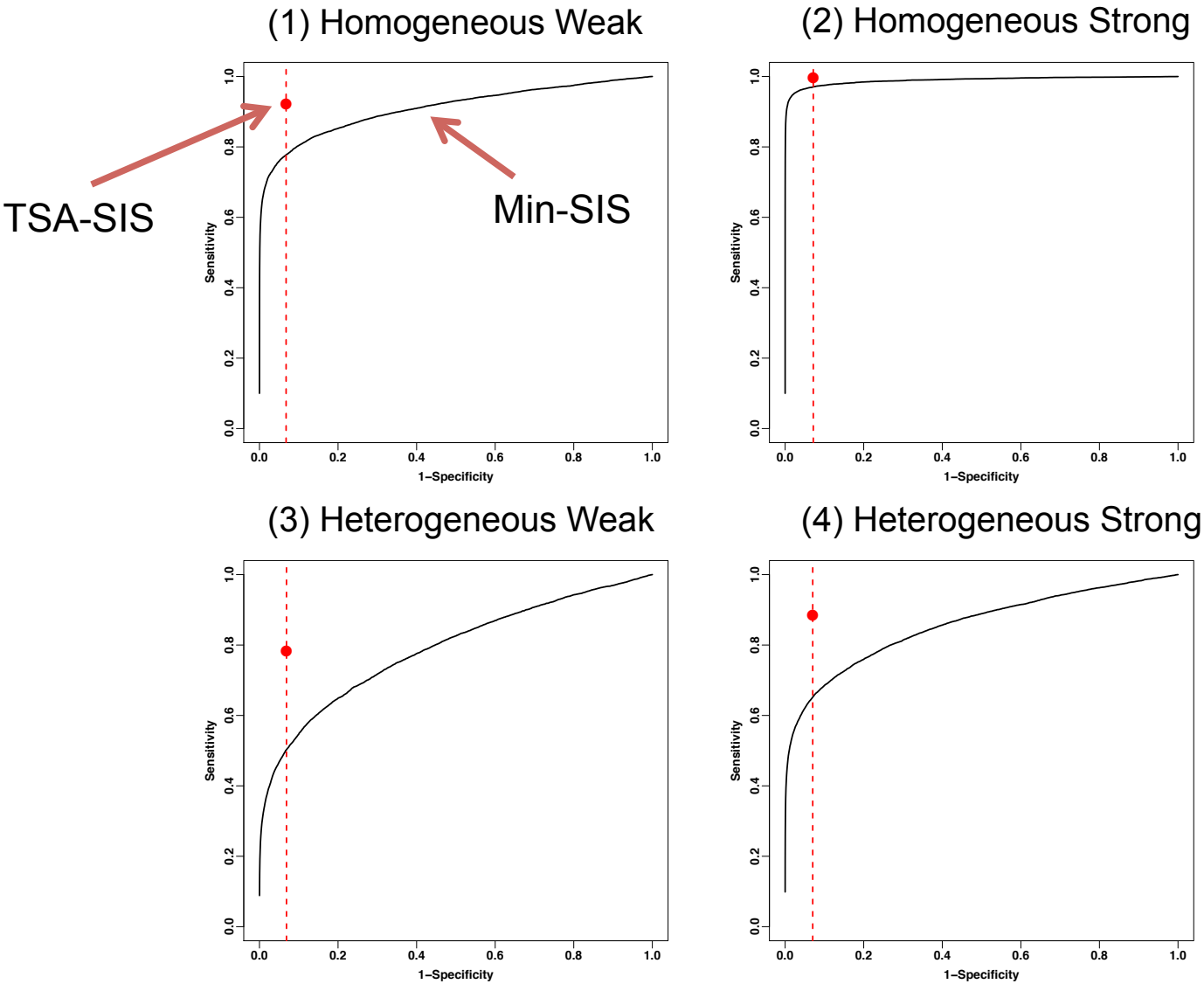
Note: All value are based on average results from $B = 1000$ replications.

Simulation setting

1. Homogeneous weak signals: $\beta_j^{(1)} = \ldots = \beta_j^{(K)} = \beta_j \sim \mathit{Unif}(0.1, 0.3)$.
2. Homogeneous strong signals: $\beta_j^{(1)} = \ldots = \beta_j^{(K)} = \beta_j \sim \mathit{Unif}(0.7, 1)$.
3. Heterogeneous weak signals: $\beta_j \sim \mathit{Unif}(0.1, 0.3)$, $\beta_j^{(k)} \sim N(\beta_j, 0.5^2)$.
4. Heterogeneous strong signals: $\beta_j \sim \mathit{Unif}(0.7, 1)$, $\beta_j^{(k)} \sim N(\beta_j, 0.5^2)$.

With $n = 100$, $p = 1000$, $K = 5$, $B = 1000$ replications, fixed $\alpha_1 = 0.0001$, $\alpha_2 = 0.05$. Compare to multiple study extension of SIS which ranks the features by the minimum correlation in all studies (min-SIS).

Simulation results



Application to TNBC example

- The tumor suppressor protein “p53", encoded by *TP53* gene, is essential in breast cancer prognosis.
- Three microarray datasets of triple-negative breast cancer (TNBC), aim to identify genes most predictive of the expression level of *TP53* gene.
- After filtering, a total of 3377 genes remained in common for the analysis.
- There are 275/178/165 TNBC samples in each dataset, respectively.
- Compare our TSA-SIS procedure to OneStep-SIS as well as the Min-SIS (by choosing the top d features, where $d = n/\log(n) = 49$).

TSA-SIS procedure helps save the weak signals

Gene	METABRIC Est (SE)	GSE25066 Est (SE)	GSE76250 Est (SE)	Min-SIS d=49	Rank in Min-SIS	OneStep-SIS $\|S\|$=25
Intercept	7.600 (1.502)	0.213 (0.553)	-1.783 (0.971)	-	-	-
EXOC1	0.251 (0.081)**	0.278 (0.157)·	0.293 (0.167)·	N	164	N
ITGB1BP1	-0.134 (0.045)**	0.003 (0.111)	-0.178 (0.194)	N	123	N
RBM23	0.168 (0.078)*	0.144 (0.167)	0.367 (0.168)*	N	152	N
SETD3	-0.166 (0.081)*	0.366 (0.184)*	-0.080 (0.175)	N	101	N
SQSTM1	-0.114 (0.050)*	0.029 (0.099)	0.245 (0.183)	N	98	N
TRIOBP	-0.126 (0.062)*	0.084 (0.118)	0.628 (0.261)*	N	91	N
Adjusted-R^2	0.151	0.522	0.359			

Note: "." indicates significant level of 0.1, "$*$" for level of 0.05, "$**$" for level of 0.01. , "$***$" for level of 0.001.

Interim summary

- We provide a general framework for simultaneous variable screening with multiple related studies and open a door to the development of methods using multiple studies to perform screening under different types of models (e.g. different outcomes) or with different marginal utilities.

Interim summary

- We provide a general framework for simultaneous variable screening with multiple related studies and open a door to the development of methods using multiple studies to perform screening under different types of models (e.g. different outcomes) or with different marginal utilities.
- In most cases, screening is still the first step that leaves us with many false positives, we need to further refine the pool using regularization methods. With multiple studies, how should we proceed?

Interim summary

- We provide a general framework for simultaneous variable screening with multiple related studies and open a door to the development of methods using multiple studies to perform screening under different types of models (e.g. different outcomes) or with different marginal utilities.
- In most cases, screening is still the first step that leaves us with many false positives, we need to further refine the pool using regularization methods. With multiple studies, how should we proceed?
- With that said, in the second part of the talk, I will introduce a two-stage highD variable selection method for censored/survival data with multiple studies/populations, which extends our current screening framework and includes a second stage regularization method.

Marginal screening in survival analysis: a short review

- Song et al. (2014) proposed a nonparametric censored rank independence screening method for survival data.

Marginal screening in survival analysis: a short review

- Song et al. (2014) proposed a nonparametric censored rank independence screening method for survival data.
- Gorst-Rasmussen & Scheike (2013) proposed a "feature aberration at survival times" (FAST) statistics as a natural survival equivalent of correlation screening in linear model.

Marginal screening in survival analysis: a short review

- Song et al. (2014) proposed a nonparametric censored rank independence screening method for survival data.
- Gorst-Rasmussen & Scheike (2013) proposed a "feature aberration at survival times" (FAST) statistics as a natural survival equivalent of correlation screening in linear model.
- Zhao & Li (2012) proposed a Principled Cox Sure Independence Screening (PSIS) to perform screening in Cox model.

Marginal screening in survival analysis: a short review

- Song et al. (2014) proposed a nonparametric censored rank independence screening method for survival data.
- Gorst-Rasmussen & Scheike (2013) proposed a "feature aberration at survival times" (FAST) statistics as a natural survival equivalent of correlation screening in linear model.
- Zhao & Li (2012) proposed a Principled Cox Sure Independence Screening (PSIS) to perform screening in Cox model.
- Intuitively, the framework can be fit to survival models other than the Cox model.

Real world motivation for high-dimensional variable selection with multiple survival studies

- Pan-cancer prognosis analysis: e.g. genes associated with survival in a variety of cancer types.
- Meta-analysis of multiple Gene-Wide Association Studies (GWAS) on survival outcomes.
- "Big" survival data, e.g. sparse high-dimensional massive sample size (sHDMSS) data such as the pediatric trauma mortality data from the National Trauma Databank (NTDB), covariates include ICD9 Codes, AIS codes, etc.

Detect survival-associated biomarkers in multiple studies

- Assume a Cox model within each study k:

$$\lambda^{(k)}(t|x^{(k)}) = \lambda_0^{(k)}(t)\exp(X^{(k)T}\beta^{(k)}).$$

Detect survival-associated biomarkers in multiple studies

- Assume a Cox model within each study k:

$$\lambda^{(k)}(t|x^{(k)}) = \lambda_0^{(k)}(t)\exp(X^{(k)T}\beta^{(k)}).$$

- As in our previous framework, we assume the same sparsity pattern (i.e., $\beta_j^{(k)} = 0$ for all k or $\beta_j^{(k)} \neq 0$ for all k) but varied signal strength.

Detect survival-associated biomarkers in multiple studies

- Assume a Cox model within each study k:

$$\lambda^{(k)}(t|x^{(k)}) = \lambda_0^{(k)}(t)\exp(X^{(k)T}\beta^{(k)}).$$

- As in our previous framework, we assume the same sparsity pattern (i.e., $\beta_j^{(k)} = 0$ for all k or $\beta_j^{(k)} \neq 0$ for all k) but varied signal strength.
- Allowing for baseline hazard rates and censoring rates to vary across different studies.

Detect survival-associated biomarkers in multiple studies

- Assume a Cox model within each study k:

$$\lambda^{(k)}(t|x^{(k)}) = \lambda_0^{(k)}(t)\exp(X^{(k)T}\beta^{(k)}).$$

- As in our previous framework, we assume the same sparsity pattern (i.e., $\beta_j^{(k)} = 0$ for all k or $\beta_j^{(k)} \neq 0$ for all k) but varied signal strength.
- Allowing for baseline hazard rates and censoring rates to vary across different studies.
- Perform screening using aggregate statistics from marginal Cox regression coefficients for each covariate, followed by a novel group lasso penalty in the partial log-likelihood of Cox models to identify a common set of features across multiple studies.

Detect survival-associated biomarkers in multiple studies

- Assume a Cox model within each study k:

$$\lambda^{(k)}(t|x^{(k)}) = \lambda_0^{(k)}(t)\exp(X^{(k)T}\beta^{(k)}).$$

- As in our previous framework, we assume the same sparsity pattern (i.e., $\beta_j^{(k)} = 0$ for all k or $\beta_j^{(k)} \neq 0$ for all k) but varied signal strength.
- Allowing for baseline hazard rates and censoring rates to vary across different studies.
- Perform screening using aggregate statistics from marginal Cox regression coefficients for each covariate, followed by a novel group lasso penalty in the partial log-likelihood of Cox models to identify a common set of features across multiple studies.
- Cox model based TwO-sTage variable sElection for Multiple studies ("Cox-TOTEM")

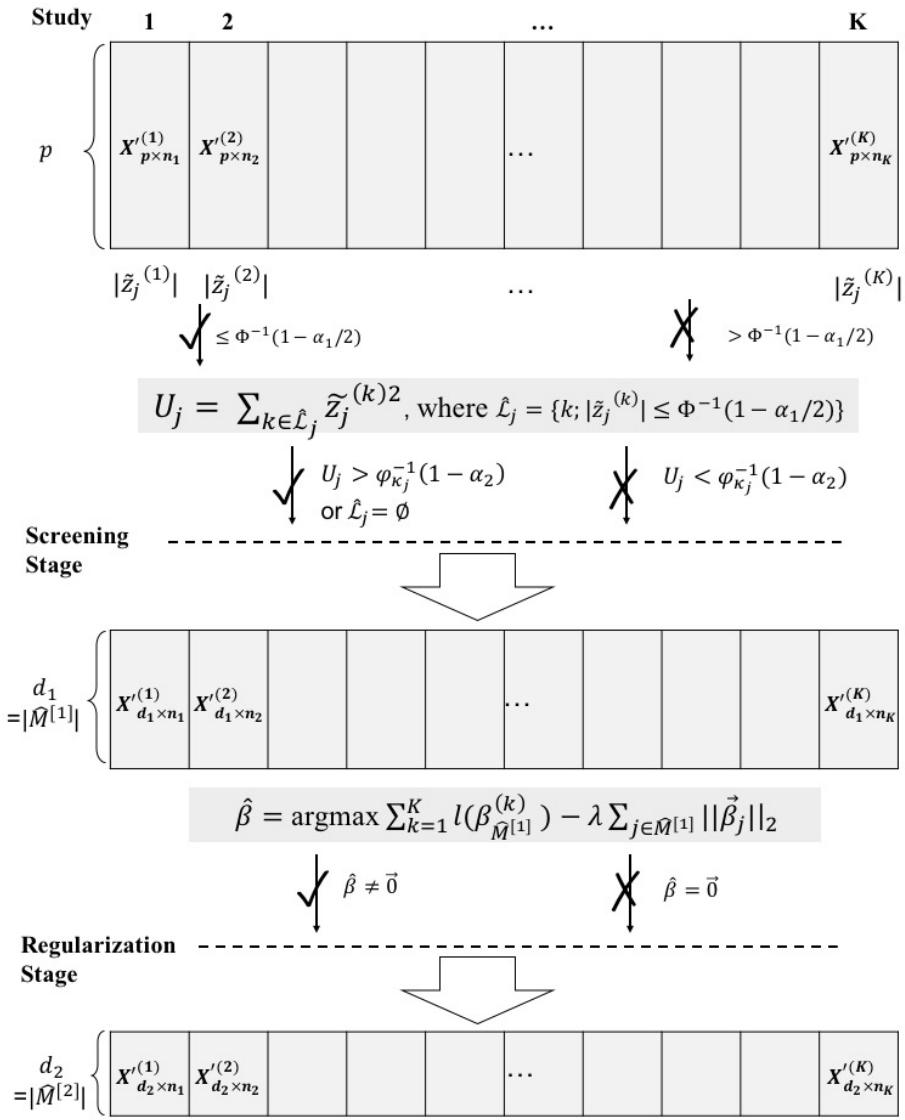


Figura 1: Flowchart of "Cox-TOTEM"

1. Screening stage

- Perform a two-step screening as we did before:

  Step 1. $\hat{\mathcal{L}}_j = \{k; |\tilde{z}_j^{(k)}| \leq \Phi^{-1}(1-\alpha_1/2)\}$ with $\tilde{z}_j^{(k)} = \mathsf{I}(\tilde{\beta}_j^{(k)})^{1/2}|\tilde{\beta}_j^{(k)}|$;

  Step 2. $\hat{\mathcal{A}}^{[1]} = \{j \in [p]; U_j > \varphi_{\hat{\kappa}_j}^{-1}(1-\alpha_2)$ or $\hat{\kappa}_j = 0\}$ with $\hat{\kappa}_j = |\hat{\mathcal{L}}_j|$ and $U_j = \sum_{k \in \hat{\mathcal{L}}_j} (\tilde{z}_j^{(k)})^2$,

  where $\tilde{\beta}_j^{(k)}$ is the marginal partial likelihood estimator for the jth feature in the kth study, $\mathsf{I}(\tilde{\beta}_j^{(k)}) = 1/\widehat{var}(\tilde{\beta}_j^{(k)})$ is the observed information matrix.

2. Regularization stage

- Since our framework assumes a common sparsity pattern across all studies, we employ group lasso penalty in the partial log-likelihoods in the second stage, where we treat the same feature in all studies as one group:

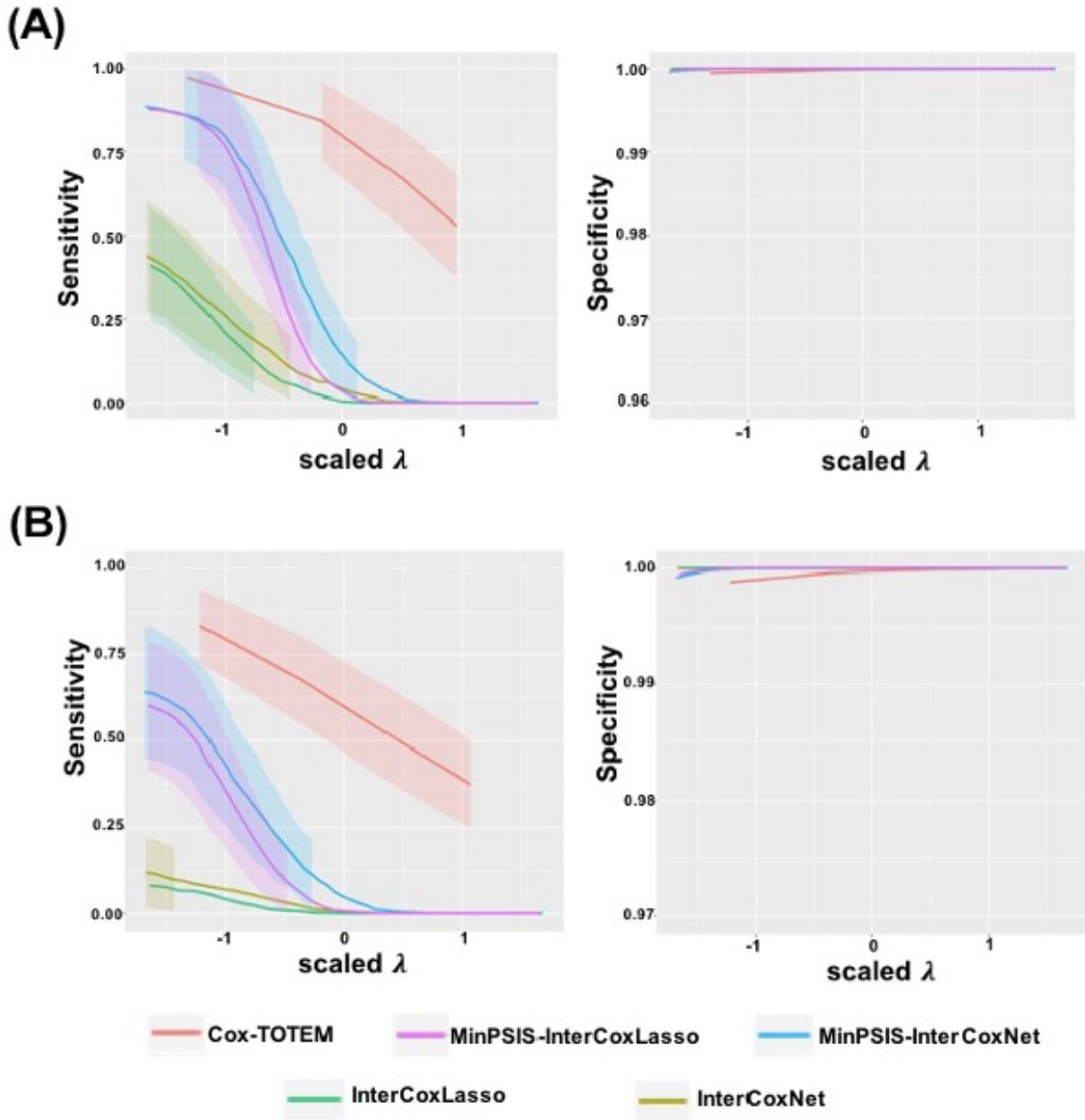
$$\min_{\beta^{(k)},k=1,2,\ldots K} -\sum_{k=1}^{K} l(\beta^{(k)}_{\hat{\mathcal{M}}^{[1]}}) + \lambda \sum_{j\in\hat{\mathcal{M}}^{[1]}} ||\vec{\beta}_j||_2 \quad ,$$

  where $l(\beta^{(k)}) = \log \prod_{i=1}^{n_k} \{ \frac{\exp(\beta^{(k)T} x_i^{(k)})}{\sum_{j\in R(Y_i^{(k)})} \exp(\beta^{(k)T} x_j^{(k)})} \}^{\Delta_i^{(k)}}$ is the partial likelihood and $\vec{\beta}_j = (\beta_j^{(1)}, \ldots, \beta_j^{(K)})$.
- An **ADMM** algorithm is used to solve the above optimization problem.

Methods for comparison

- One-stage: InterCoxLasso, InterCoxNet
  - Apply CoxLasso or CoxNet in each study and then take the intersection.
- Two-stage: MinPSIS-InterCoxLasso, MinPSIS-InterCoxNet
  - Multiple study extension of Zhao et al. (2012)'s "PSIS" for sure screening in the Cox model followed by InterCoxLasso or InterCoxNet.

Simulation results



Simulation results

Table 1. Comparison of variable selection and parameter estimation under the four different scenarios with $n = 100$, $p = 2000$ and $s = 10$ true predictors. Mean results of 100 replications are reported with standard errors shown in the parentheses.

Simulation scenarios	Methods	Sensitivity	Specificity	Average number of features selected	SSE
Homo-S	Cox-TOTEM	0.985 (0.004)	0.999 (0)	12.26 (0.194)	4.227 (0.269)
	MinPSIS-InterCoxLasso	0.856 (0.018)	1 (0)	9.4 (0.173)	11.36 (1.02)
	MinPSIS-InterCoxNet	0.881 (0.013)	0.997 (0.001)	14.36 (0.244)	10.79 (0.772)
	InterCoxLasso	0.37 (0.018)	1(0)	3.7 (0.183)	37.421 (0.775)
	InterCoxNet	0.53 (0.014)	1(0)	5.31 (0.143)	30.823 (0.684)
Homo-W	Cox-TOTEM	0.977 (0.049)	0.997 (0)	16.6 (0.385)	3.321 (0.203)
	MinPSIS-InterCoxLasso	0.464 (0.028)	0.999 (0)	5.17 (0.347)	7.646 (0.311)
	MinPSIS-InterCoxNet	0.618 (0.019)	0.997 (0)	13.07 (0.384)	6.804 (0.214)
	InterCoxLasso	0.049 (0.007)	1(0)	0.49 (0.073)	12.032 (0.073)
	InterCoxNet	0.143 (0.011)	1(0)	1.43 (0.11)	11.094 (0.117)
Hetero-S	Cox-TOTEM	0.968 (0.006)	0.998 (0)	13.64 (0.261)	7.969 (0.437)
	MinPSIS-InterCoxLasso	0.723 (0.023)	1(0)	7.98 (0.226)	20.758 (1.27)
	MinPSIS-InterCoxNet	0.774 (0.015)	0.997 (0.001)	13.41 (0.26)	18.986 (0.943)
	InterCoxLasso	0.258 (0.017)	1(0)	2.58 (0.167)	45.579 (0.716)
	InterCoxNet	0.393 (0.015)	1(0)	3.93 (0.146)	39.842 (0.698)
Hetero-W	Cox-TOTEM	0.885 (0.01)	0.993 (0)	22.68 (0.66)	14.163 (0.163)
	MinPSIS-InterCoxLasso	0.211 (0.017)	1(0)	2.93 (0.241)	13.118 (0.276)
	MinPSIS-InterCoxNet	0.367 (0.016)	0.997 (0)	10.01 (0.4)	11.813 (0.251)
	InterCoxLasso	0.037 (0.006)	1(0)	0.37 (0.06)	15.738 (0.123)
	InterCoxNet	0.083 (0.008)	1(0)	0.83 (0.075)	14.932 (0.154)

Application to Pan-cancer transcriptomic studies

- Transcriptomic data from TCGA Pan-cancer project studying four gynecological cancer types plus breast cancer (together known as Pan-Gyn cancers): OV, UCEC, CESC, UCS and BRCA (K=5).
- Previous studies in TCGA identified shared molecular features among these five cancer type.
- The purpose of this application is to identify the survival associated genes common in these five cancer types.
- RNA-seq data (in TPM) from five studies merged by quantile normalization, filtered and log2 transformed.

Tabella 1: Overview of the samples from the five Pan-Gyn cancer types

	OV	UCEC	CESC	UCS	BRCA
Sample size	304	170	301	57	981
Censoring proportion	39%	81%	76%	39%	89%

- The censoring distributions vary greatly across the different cancer types, implying the heterogeneity in censoring patterns among studies commonly seen in real datasets.

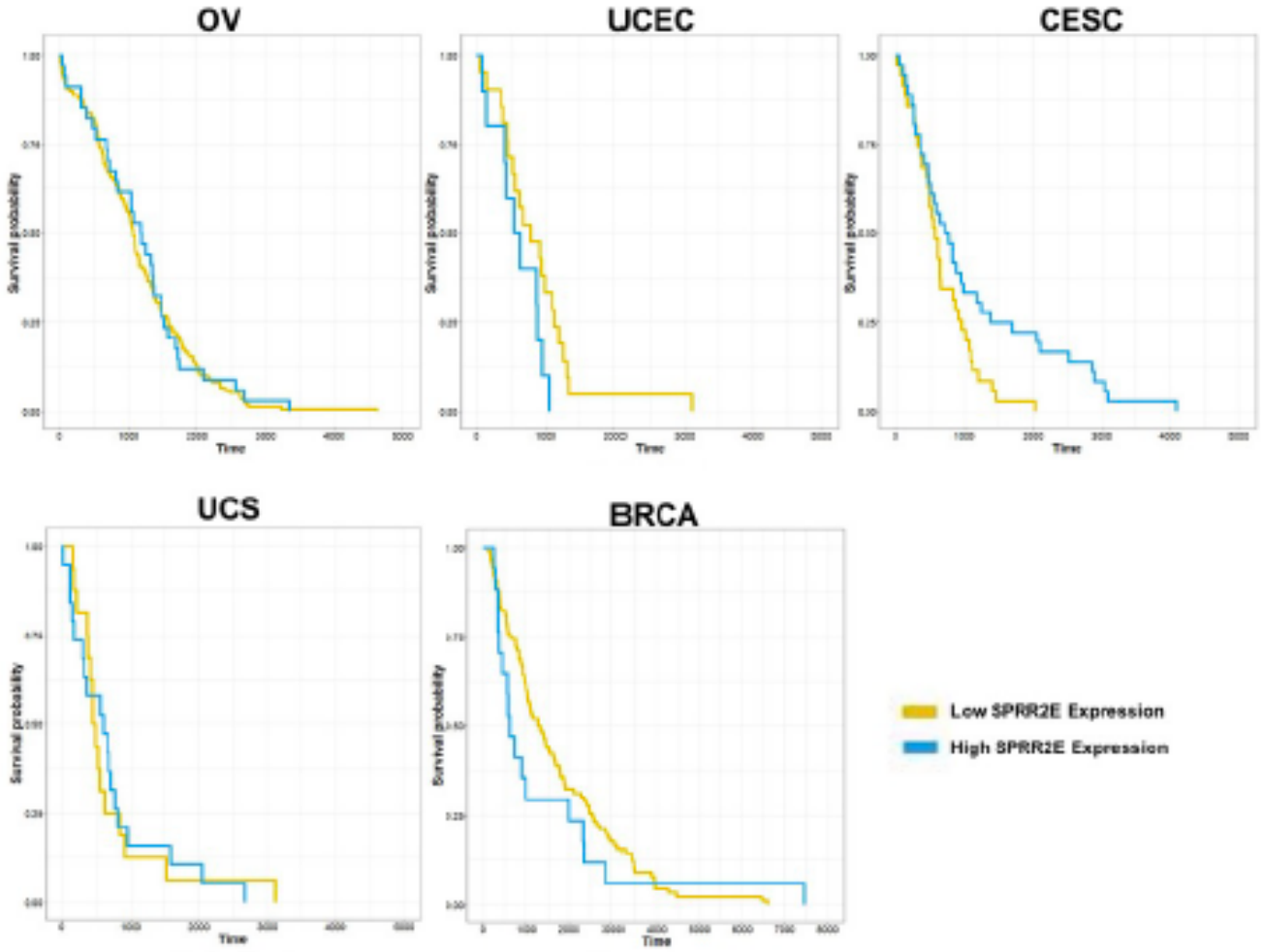
Pan-Gyn application results

Table 3. List of the top five genes from the 29 genes selected by Cox-TOTEM with the largest absolute coefficient estimates and their corresponding coefficient estimates and p-values (in parentheses) when fitting a Cox model in each cancer type. None of those genes were selected by MinPSIS-InterCoxLasso or MinPSIS-InterCoxNet.

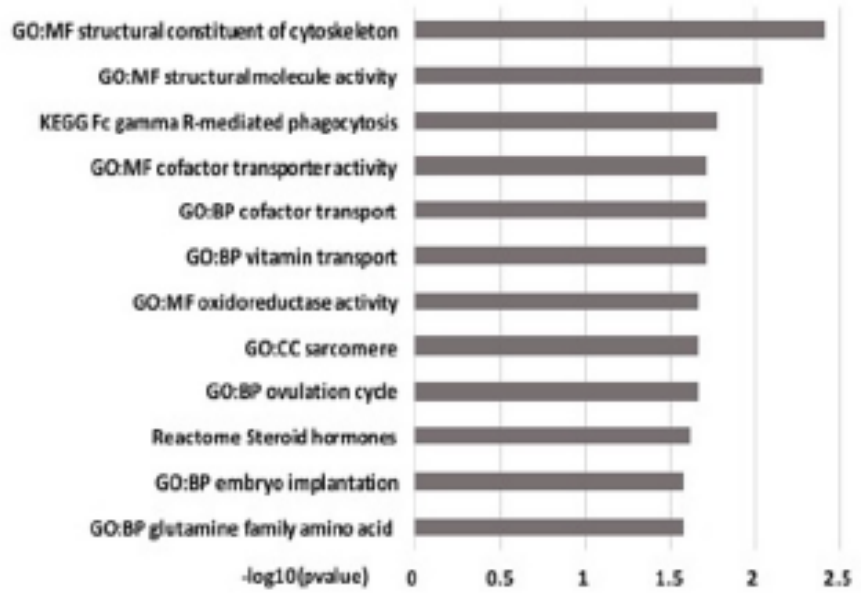
	Cox model coefficient estimate (p-value)					MinPSIS-InterCoxLasso	MinPSIS-InterCoxNet
	OV	UCEC	CESC	UCS	BRCA		
PPAP2C	-.214 (.007**)	-.291 (.131·)	-.130 (.244)	.155 (.455)	-.074 (.594)	N.S.	N.S.
SLC19A1	.240 (.006**)	.180 (.368)	-.232 (.089*)	.158 (.521)	.115 (.330)	N.S.	N.S.
SPRR2E	-.058 (.619)	2.002 (.009**)	-.300 (.095*)	.072 (.654)	.128 (.079*)	N.S.	N.S.
EIF4E3	-.177 (.041*)	-.079 (.632)	-.234 (.119·)	-.240 (.215)	-.201 (.074*)	N.S.	N.S.
IFRD2	-.133 (.118·)	-.274 (.237)	.349 (.029*)	-.188 (.473)	.129 (.197)	N.S.	N.S.

Significant code: < 0.01**, < 0.1*, < 0.15·; N.S.: Not Selected

Pan-Gyn application results



Pan-Gyn application results



- Pathways of important biological processes specific to female physiology such as GO:BP ovulation cycle and GO:BP embryo implantation.

Summary

- We proposed a general framework and an efficient procedure for sure screening with multiple studies, opening a door to the development of meta-analysis methods for highD variable screening and selection.
- Such framework and procedure borrowed the information across studies to help screen out more false positive noises while in the same time keep most of the true signals.
- We also extended the framework to model survival data from multiple studies and include a second stage regularization using group lasso to to perform highD variable selection in survival analysis.
- The study to study heterogeneity has been carefully considered in the methods, future direction will consider heterogeneous sparsity pattern across studies.

Acknowledgment

- Zhao Ren, University of Pittsburgh
- George C. Tseng, University of Pittsburgh
- Takumi Saegusa, University of Maryland
- Zhiwei Zhao, University of Maryland
- Hongjie Ke, University of Maryland
- Zhenyao Ye, University of Maryland
- Shuo Chen, University of Maryland

Our research group

Visit us at `https://www.umdbright.com/`

Paper reference

- Ma T., Ren Z. and Tseng G. (2020). Variable screening with multiple studies. *Statistica Sinica*. in press.
  `https://matianzhou.github.io/files/preprints/TSA-SIS.pdf`

- Saegusa T., Zhao Z., Ke H., Ye Z., Xu Z., Chen S. and Ma T. (2020+). Detecting survival associated biomarkers in heterogeneous populations. Under revision in *Scientific Reports*. Software available at `https://github.com/kehongjie/CoxTOTEM`

Thanks for your attention !
Any questions?

Hope everyone stays safe and healthy !