Classification with Ultrahigh-Dimensional Features

Yi Li

Department of Biostatistics, University of Michigan

Workshop on Quantitative Methods for Drug Discovery and Development

Singapore, July 2017

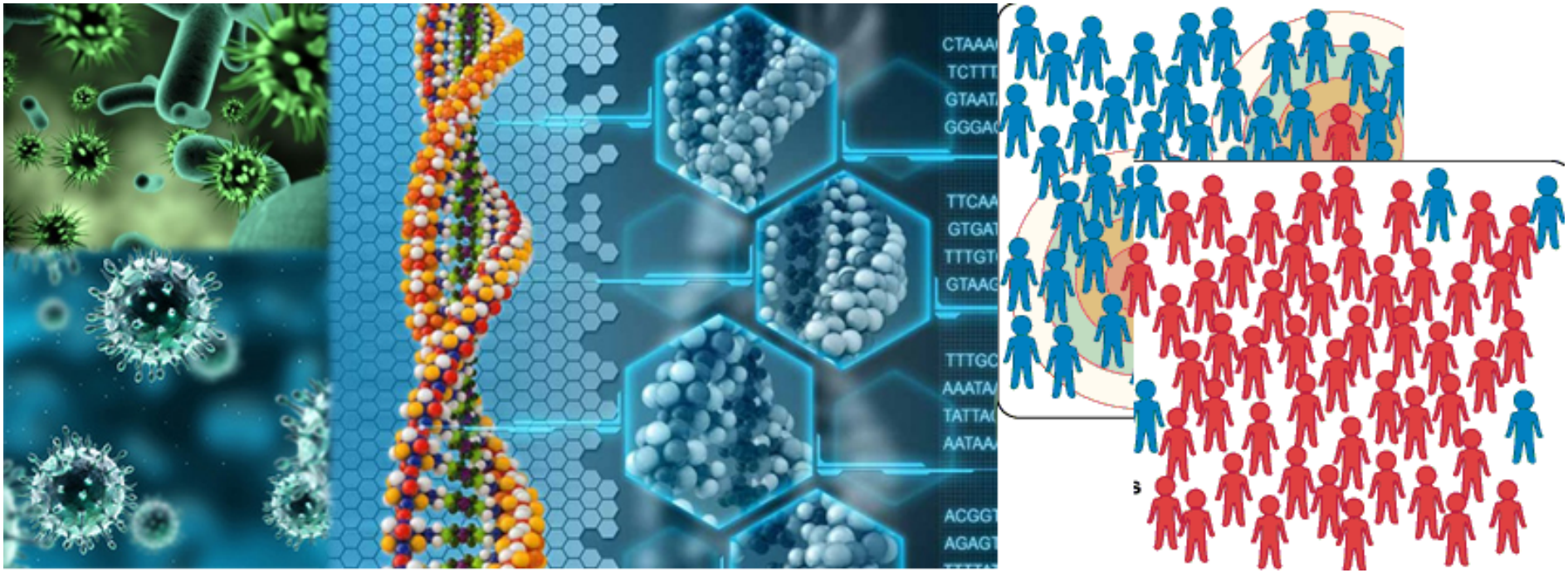
Outline

1. Background
2. Motivation and Challenges
3. Methods
4. Simulation
5. A Data Example
6. Conclusion

Outline

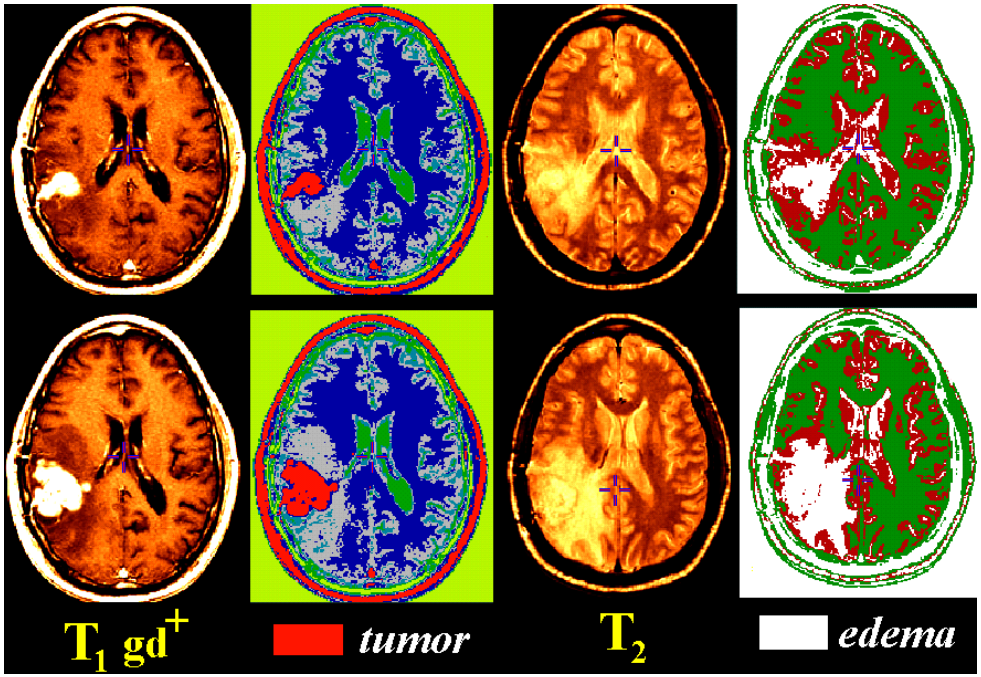
1. Background
2. Motivation and Challenges
3. Methods
4. Simulation
5. A Data Example
6. Conclusion

From Molecular Diversity to Population Classification



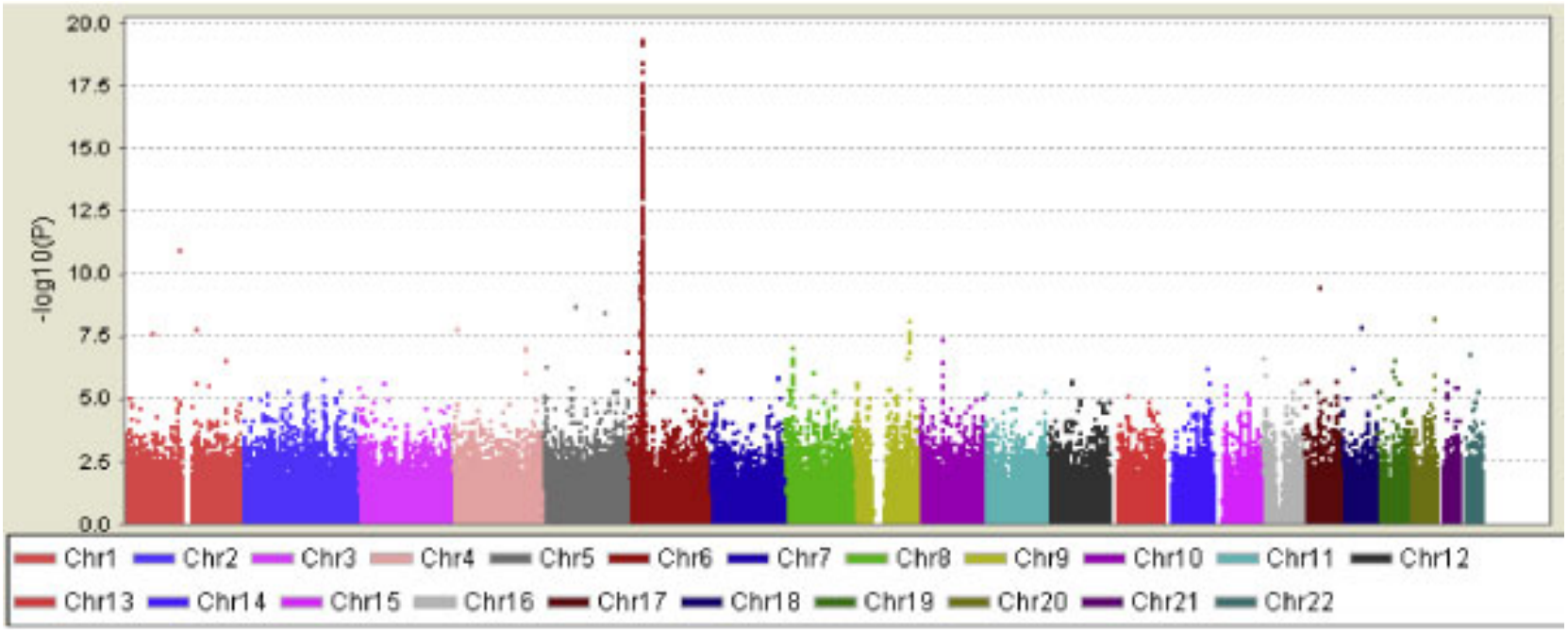
Some Applications for Ultrahigh-Dimensional Classification

- Distinguish tumor and normal brain tissues



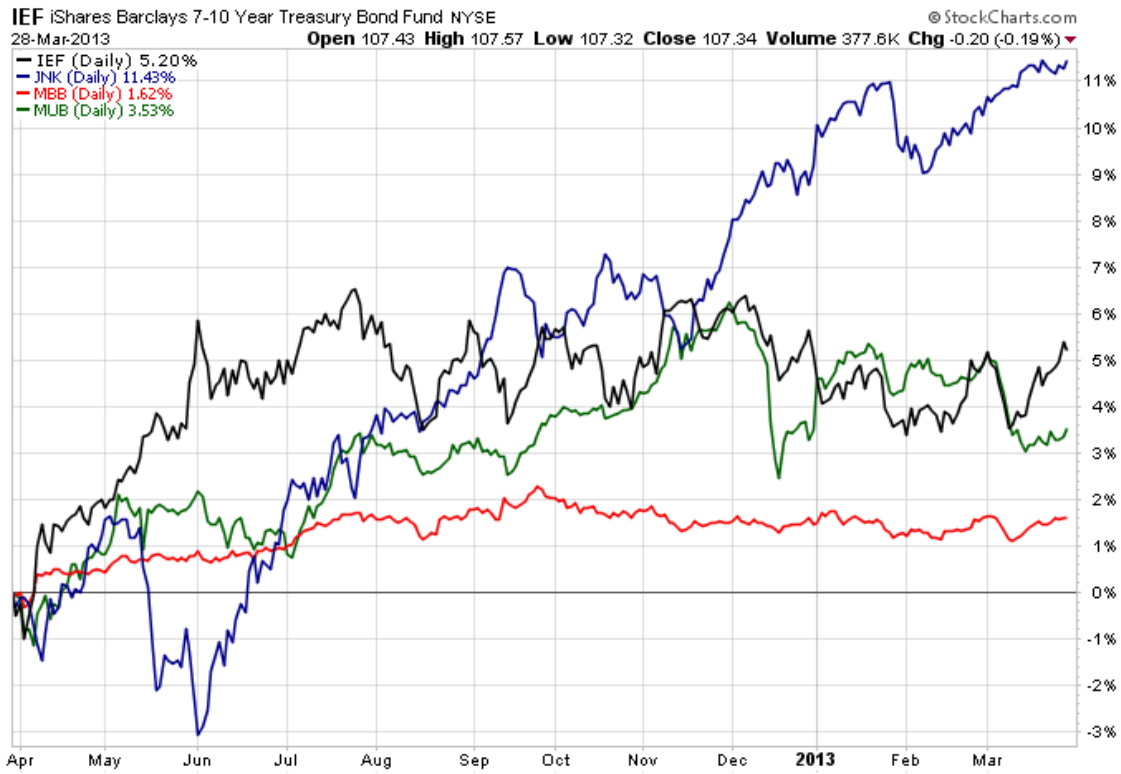
Applications for Ultrahigh-Dimensional Classification

- Predict disease status based on genomic profiles



Applications for Ultrahigh-Dimensional Classification

- Classify portfolios based on returns



Applications for Ultrahigh-Dimensional Classification

- Facial recognition

Outline

1 Background

2 Motivation and Challenges

3 Methods

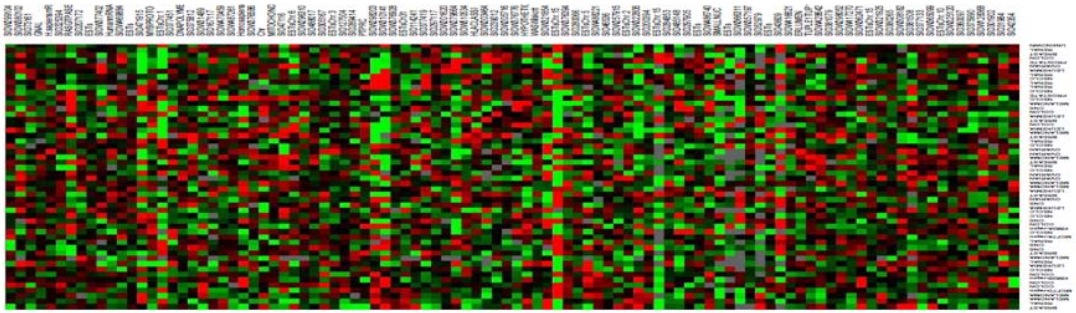
4 Simulation

5 A Data Example

6 Conclusion

A Motivating Study

- Goal: Classify kidney transplant patients with gene expression profiles
- To distinguish different types of kidney transplant recipients is crucial for
  - balancing immuno-suppression to prevent rejection
  - minimizing drug-introduced toxicities
  - risk prediction for new patients and facilitating personalized transplant policy making
- Dataset: Post-transplant rejection project [Flencher et al., 2004] contains 62 kidney transplant patients with 4 post-transplantation rejection types and their genome-wide gene profiles across 12,625 genes



Introduction

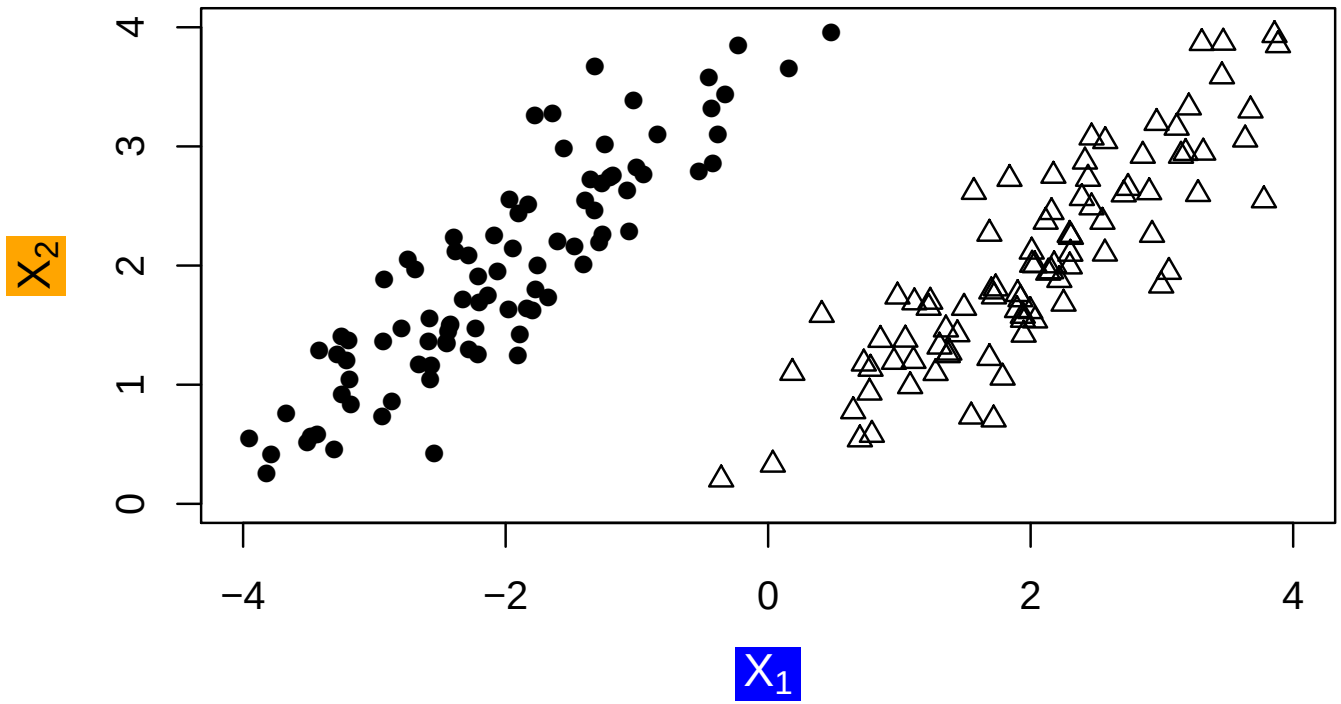
- A common goal is to identify a small number of relevant genetic variants or genes for further investigation
- Screening reduces ultrahigh dimensionality to a handleable scale
- Currently identified genetic variants or genes only explain a small proportion of disease heritability. More relevant features should be included, especially those low-risk susceptibility variants

Introduction (Cont.)

- Most screening methods focus on detecting marginally strong signals. Weak signal detection has drawn much attention recently
- Many classification rules use the independence rule and ignore inter-feature correlations, failing to detect Marginally Uninformative but Jointly Informative (MUJI) signals

Are MUJI Features Famous?

What is a MUJI Feature?



X_1 : Marginally Informative

X_2 : Marginally Uninformative but Jointly Informative with X_1

A Quick Literature Review of High-Dimensional Classification

Methods	Ultrahigh-dimensional screening method other than regularization method?	Consider inter-feature correlation?	Able to detect marginally weak features?
Guo 2010	X	X	X
Witten & Tibshirani 2011	X	X	X
Clemmensen 2011	X	X	X
Fan & Lv 2008	√	X	X
Xu et al. 2014	X	√	√
Cai & Sun 2014	√	X	X
Cai & Liu 2011	X	√	√
Fan, Hong & Tong 2012	X	√	√
Mai et al. 2012	X	√	√
Gaynanova et al 2016	X	√	√

Proposed Covariance-Insured Screening and Classification Method

- Can handle ultrahigh-dimensional cases
- Utilize inter-feature correlations
- Can detect MUJI signals that would be missed by marginal screening methods
- Can be used to classify features from a much wider class of distributions than Gaussian distribution
- Computationally efficient
- Recovers all informative features with a large probability
- Has optimal post-screening misclassification error

Outline

1 Background

2 Motivation and Challenges

3 Methods

4 Simulation

5 A Data Example

6 Conclusion

Two-Class Classification

- Y_i: a class label taking values in $\{1, 2\}$
- Conditional on Y_i, $\mathbf{X}_i$ follows a multivariate sub-Gaussian distribution with a mean vector $\boldsymbol{\mu}_k$ and a marginal variance proxy σ^2

$$\mathbf{X}_i | \{Y_i = k\} \sim SG(\boldsymbol{\mu}_k, \sigma^2), \quad i = 1, \cdots, n \text{ and } k = 1, 2.$$

- Sub-Gaussian distribution contains a wide class of distributions such as Gaussian, binary and all bounded random variables
- Assume that the two classes have a common covariance matrix

$$\boldsymbol{\Sigma} = E[(\mathbf{X}^{(k)} - \boldsymbol{\mu}_k)(\mathbf{X}^{(k)} - \boldsymbol{\mu}_k)']$$

- Most marginal screening processes assume a diagonal $\boldsymbol{\Sigma}$ and ignore the inter-feature correlations

Non-informative features

- Let $\omega_k = \mathbb{P}(Y_i = k), k = 1, 2$. For multivariate Gaussian distributed features, consider the pairwise difference

$$\log \mathbb{P}(Y_i = 1|\mathbf{X}_i = \mathbf{x}) - \log \mathbb{P}(Y_i = 2|\mathbf{X}_i = \mathbf{x}) = \log \omega_1 - \log \omega_2$$
$$-\frac{1}{2}\sum_{j=1}^{p}(\mu_{1j} + \mu_{2j})\sum_{j'=1}^{p}\Omega_{jj'}(\mu_{1j'} - \mu_{2j'}) + \sum_{j=1}^{p}x_j\left\{\sum_{j'=1}^{p}\Omega_{jj'}(\mu_{1j'} - \mu_{2j'})\right\}$$

- Variable j is non-informative for distinguishing the two classes if and only if

$$\sum_{j'=1}^{p}\Omega_{jj'}(\mu_{1j'} - \mu_{2j'}) = 0$$

where $\Omega_{jj'}$ is the (j, j')'th entry of the precision matrix (Xu et al. 2014)

MUJI features

Definition

Feature X_j is called a MUJI signal, if and only if it satisfies the following two conditions:

$$\mu_{1j} - \mu_{2j} = 0 \tag{A}$$

$$\sum_{j'=1}^{p} \Omega_{jj'}(\mu_{1j'} - \mu_{2j'}) \neq 0. \tag{B}$$

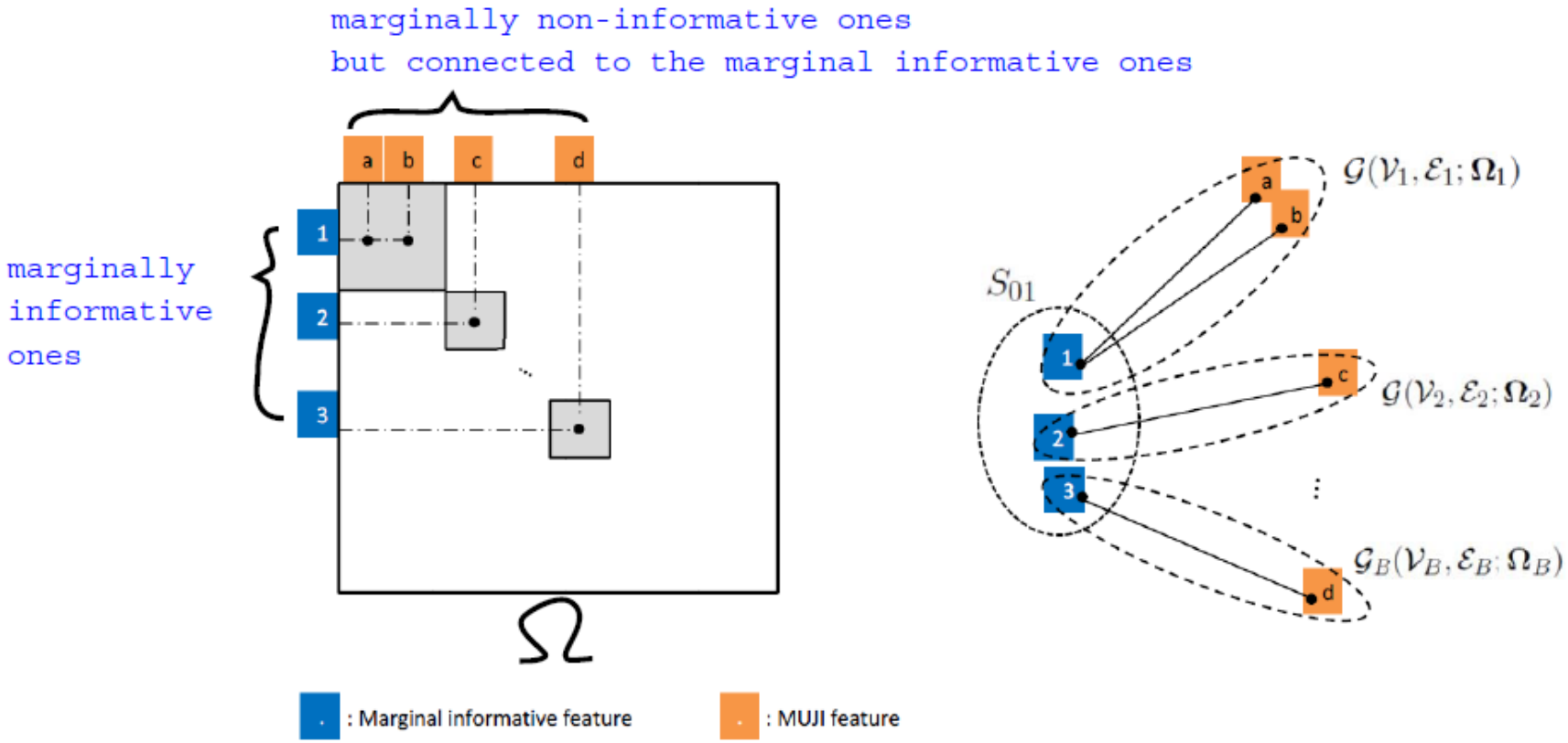
A MUJI feature must be partially correlated with at least one marginally informative feature!

Challenges in Finding MUJI Features

- Marginal methods do not consider the off-diagonal elements of $\boldsymbol{\Omega}$
- Estimating $\boldsymbol{\Omega}$ is infeasible when p is large

Our Proposal: Covariance-insured Screening (CIS)

- Identify marginally strong features
- Identify marginally weak/non-informative features that are (partially) correlated with them
- Detecting connected components of marginally informative features in a graph is the Key!



Detecting Connected Components of a Marginally Informative Feature

- Detect connected components of marginally informative features via thresholding sample covariance matrix $\tilde{\mathbf{\Sigma}}^{\alpha} \equiv \hat{\mathbf{\Sigma}} \mathbb{1}\{|\hat{\mathbf{\Sigma}}| \geq \alpha\}$

Theorem

Let $\mathcal{G}(\mathcal{V}_j, \mathcal{E}_j; \mathbf{\Omega}_j)$ be the connected component of a marginally informative feature j in $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{\Omega})$, and $\mathcal{G}(\tilde{\mathcal{V}}_j, \tilde{\mathcal{E}}_j; \tilde{\mathbf{\Sigma}}_j^{\alpha})$ be the connected component of feature j in $\mathcal{G}(\mathcal{V}, \tilde{\mathcal{E}}; \tilde{\mathbf{\Sigma}}^{\alpha})$, then for sufficiently large n and $\alpha = O(\sqrt{n^{\xi-1}})$, we have

$$P\left(\mathcal{G}(\mathcal{V}_j, \mathcal{E}_j; \mathbf{\Omega}_j) \subseteq \mathcal{G}(\tilde{\mathcal{V}}_j, \tilde{\mathcal{E}}_j; \tilde{\mathbf{\Sigma}}_j^{\alpha})\right) \geq 1 - C_1 \exp(-C_2 n^{\xi}),$$

where C_1, C_2 are some positive constants and $0 < \xi < 1$.

How Does It Work?

The goal is to estimate $\mathbf{\Omega}$ or the conditional dependency graph $\mathcal{E}$

- Given an estimate $\hat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$ (which is easy to get!), call variables i, j conditionally dependent $\hat{\Sigma}_{ij} \geq \alpha$ (Luo et al., 2014)
- This would give us an estimate of edge set $\mathcal{E}$
- "From a distance" $\mathbf{\Omega}$ resembles $\mathbf{\Sigma}$, which makes sense intuitively: the inverse of a block-diagonal matrix is also block-diagonal
- The results hold for sub-Gaussian data

Covariance-insured Screening and Classification ($K = 2$)

Screening step

1. Select marginally informative signals

$$\left|\bar{X}_{\cdot j}^{(1)} - \bar{X}_{\cdot j}^{(2)}\right| = \left|\frac{1}{n_1}\sum_{Y_i=1} X_{ij} - \frac{1}{n_2}\sum_{Y_i=2} X_{ij}\right| > \tau$$

2. Construct the thresholded sample covariance matrix

$$\tilde{\Sigma}_{jj'}^{\alpha} = \hat{\Sigma}_{jj'}\mathbb{1}\{|\hat{\Sigma}_{jj'}| \geq \alpha\},\ 1 \leq j, j' \leq p$$

3. For each selected marginally informative signal, detect its connected component $\tilde{\mathbf{\Sigma}}_k^{\alpha}$ and set the estimated precision matrix

$$\hat{\Omega}^{\alpha,\tau} = \mathrm{diag}((\tilde{\mathbf{\Sigma}}_{S_1}^{\alpha})^{-1}, \cdots, (\tilde{\mathbf{\Sigma}}_{S_{k(\alpha,\tau)}}^{\alpha})^{-1})$$

4. Select informative features by

$$S^{\mathrm{CIS}}(\alpha, \tau; \nu_n) = \left\{ j \in \cup_{g=1}^{k(\alpha,\tau)} S_g : \left| \sum_{j' \in S_{g(j)}} (\tilde{\Sigma}_{S_{g(j)}}^{\alpha})_{jj'}^{-1} (\bar{X}_{\cdot j'}^{(1)} - \bar{X}_{\cdot j'}^{(2)}) \right| \geq \nu_n \right\}$$

Post-screening classification

5. Classify a new observation $\mathbf{X}_{\text{new}}$ to class 1 if

$$(\mathbf{X}_{\text{new}} - \hat{\boldsymbol{\mu}}^{\text{CIS}})' \hat{\boldsymbol{\Omega}}^{\text{CIS}} (\hat{\boldsymbol{\mu}}_1^{\text{CIS}} - \hat{\boldsymbol{\mu}}_2^{\text{CIS}}) > 0,$$

where $\hat{\boldsymbol{\mu}}_1 = \sum_{\{i:Y_i=1\}} \mathbf{X}_i/n_1$, $\hat{\boldsymbol{\mu}}_2 = \sum_{\{i:Y_i=2\}} \mathbf{X}_i/n_2$, $\hat{\boldsymbol{\mu}} = (\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_2)/2$, $\hat{\boldsymbol{\Omega}} = \hat{\boldsymbol{\Omega}}^{\alpha,\tau}$ and superscript "CIS" restricts features to $S^{\text{CIS}}(\alpha, \tau; \nu_n)$

Rank Importance of Variables

- Importance Score (IS) evaluates the importance of a feature j:

$$IS_j = \left| \sum_{j' \in S_{g(j)}} (\tilde{\Sigma}^{\alpha}_{S_{g(j)}})^{-1}_{jj'} (\bar{X}^{(1)}_{\cdot j'} - \bar{X}^{(2)}_{\cdot j'}) \right|$$

- IS incorporates inter-feature correlations. When $\mathbf{\Omega}$ is diagonal, it reduces to the absolute mean differences

Theoretical Framework

- $S_0 = \{j : \sum_{j'=1}^{p} \Omega_{jj'}(\mu_{1j'} - \mu_{2j'}) \neq 0, j = 1, \cdots, p\}$ be the true informative set
- $S_{01} = \{j : \mu_{1j} - \mu_{2j} \neq 0, j = 1, \cdots, p\}$ is the true marginally informative set
- $S_{02} = \{j : \mu_{1j} - \mu_{2j} = 0, \sum_{j'=1}^{p} \Omega_{jj'}(\mu_{1j'} - \mu_{2j'}) \neq 0, j = 1, \cdots, p\}$ is the MUJI set
- $S_0 = S_{01} \cup S_{02}$

A Sparse and Weak Feature Model

- Assume that $\boldsymbol{\mu_2}$ is only different from $\boldsymbol{\mu_1}$ in ϵ fraction of coordinates, i.e., $|S_{01}| = \epsilon p$.
- Let $l_1, l_2, \cdots, l_p$ be p samples from Bernoulli(ϵ), and

$$\mu_{2j} = \mu_{1j} + \delta_j l_j, \quad 1 \leq j \leq p,$$

- Set $\tau = \min\{|\delta_j| : 1 \leq j \leq p, l_j = 1\} = \sqrt{r \log p}$ for some $r \in (0, 1)$
- Larger r, stronger signals
- Set $\epsilon = p^{-\beta}$ for some $0 < \beta < 1$. Larger β, sparser the marginally informative signals
- Assume $|S_{02}| = p^{1-\gamma}$, for some $0 < \gamma < 1$. Larger γ, sparser the MUJI signals

Sure Screening Property

Let $S_0 = \{j : \sum_{j'=1}^{p} \mathbf{\Omega}_{jj'}(\mu_{1j'} - \mu_{2j'}) \neq 0, j = 1, \cdots, p\}$ be the true informative set (including marginally informative and MUJI features)

Proposition

With $\tau = O(\sqrt{r \log p})$, $\alpha = O(\sqrt{n^{\xi - 1}})$ *and* $\nu_n = O\left(\sqrt{r \log p}\right)$ *and any* $\iota > 0$, $S_{CIS}^{\alpha,\tau}$ *satisfies*

$$P(|S_0 \cap S_{CIS}^{\alpha,\tau}| \geq (1 - \iota)|S_0|) \to 1 \text{ as } n \to \infty$$

False Positive Control Property

Proposition

For any $\gamma_n = O(p^{\varsigma})$ *with* $0 < \varsigma < 1$, $S_{CIS}^{\alpha,\tau}$ *satisfies*

$$P(|S_{CIS}^{\alpha,\tau} \cap S_0^c| \leq \gamma_n^{-1}|S_0^c|) \to 1 \text{ as } n \to \infty$$

Post-screening Misclassification Rate

- R_{OPT}: the misclassification rate of the oracle rule assuming that the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known
- $R_{\text{CIS\_PSC}}$: the misclassification rate for the CIS classification rule $\mathbf{1}(\hat{\psi}^{CIS} \geq 0)$ with $\hat{\psi}^{CIS} = (\mathbf{X}_{\text{new}}^{CIS} - \hat{\boldsymbol{\mu}}^{CIS})^T \hat{\boldsymbol{\Omega}}^{CIS} (\hat{\boldsymbol{\mu}}_1^{CIS} - \hat{\boldsymbol{\mu}}_2^{CIS})$, then

Proposition

When classifying $\mathbf{X}_{new}$ based on features selected from the screening step, we have

$$R_{CIS\_PSC} - R_{OPT} \to_P 0$$

Discovery and Classification Boundaries

- A **discovery boundary** $d(\beta)$ divides the $\beta - r$ plane:

  (a) If $r > d(\beta)$, it is possible to find a decision rule ψ such that at least some features in S_0, if not all, can be discovered by $\psi(\tau)$

  (b) If $r < d(\beta)$, it is impossible to find a decision rule ψ to find any informative features in S_0

- A **classification boundary** $c(\beta)$ divides the $\beta - r$ plane:

  (a) If $r > c(\beta)$, it is possible to find a decision rule ψ such that the two classes can be reliably separated

  (b) If $r < c(\beta)$, it is impossible to find a decision rule ψ to reliably separate the two classes

Discovery and Classification Boundaries (Cont.)

Proposition (**Discovery Boundary**)

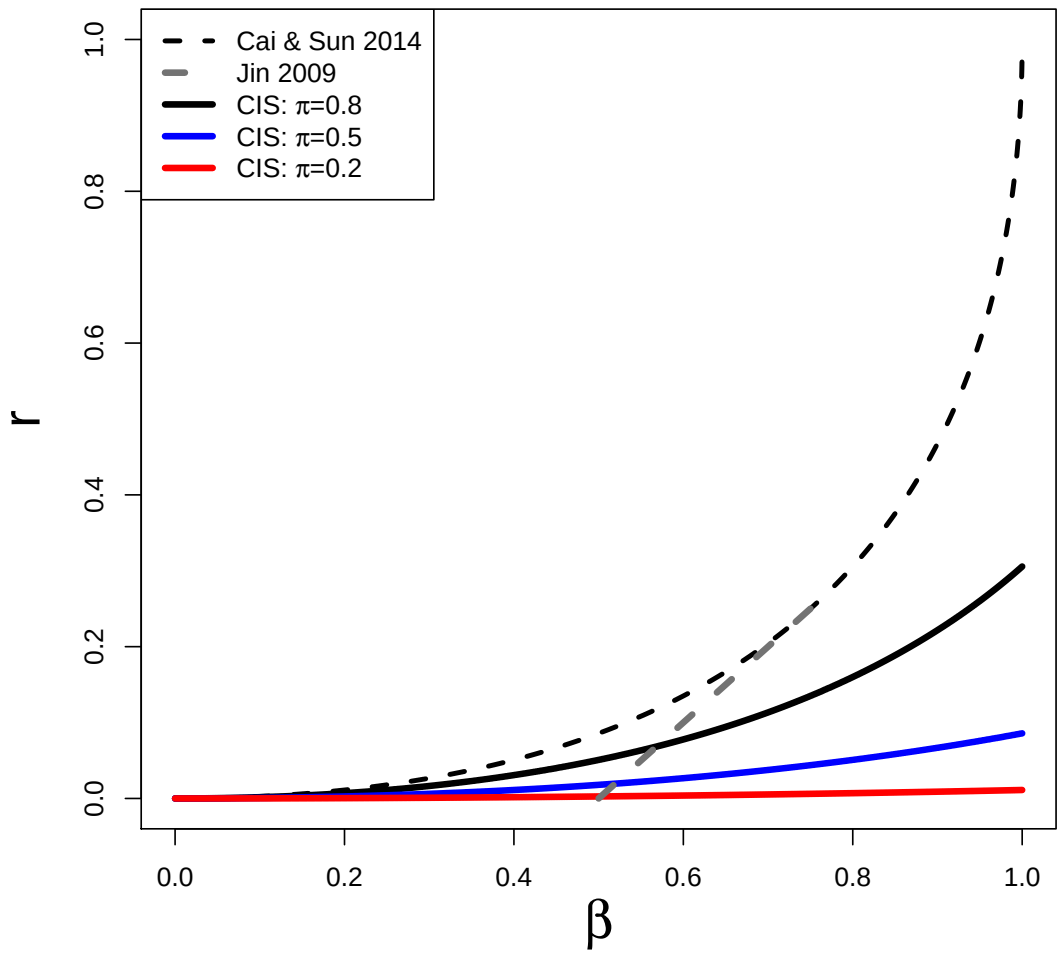
Let $\pi = \min\left\{\frac{\gamma - \log_p g(\alpha_{\min})}{\beta}, 1\right\}$, *where* $\alpha_{\min} = \min_{(j,j')\in\mathcal{E}} |\Sigma_{jj'}|$ *and* g *is an increasing function. Then the discovery boundary of the CIS classification rule* $\mathbf{1}(\hat{\psi}^{CIS}(\tau, \alpha, \nu) \geq 0)$ *is always below that of the independence rule* $\mathbf{1}(\hat{\psi}^{ind}(\tau) \geq 0)$

- $\pi\beta$ can be then interpreted as the effective sparsity of a marginal feature after taking into consideration its correlated MUJI features

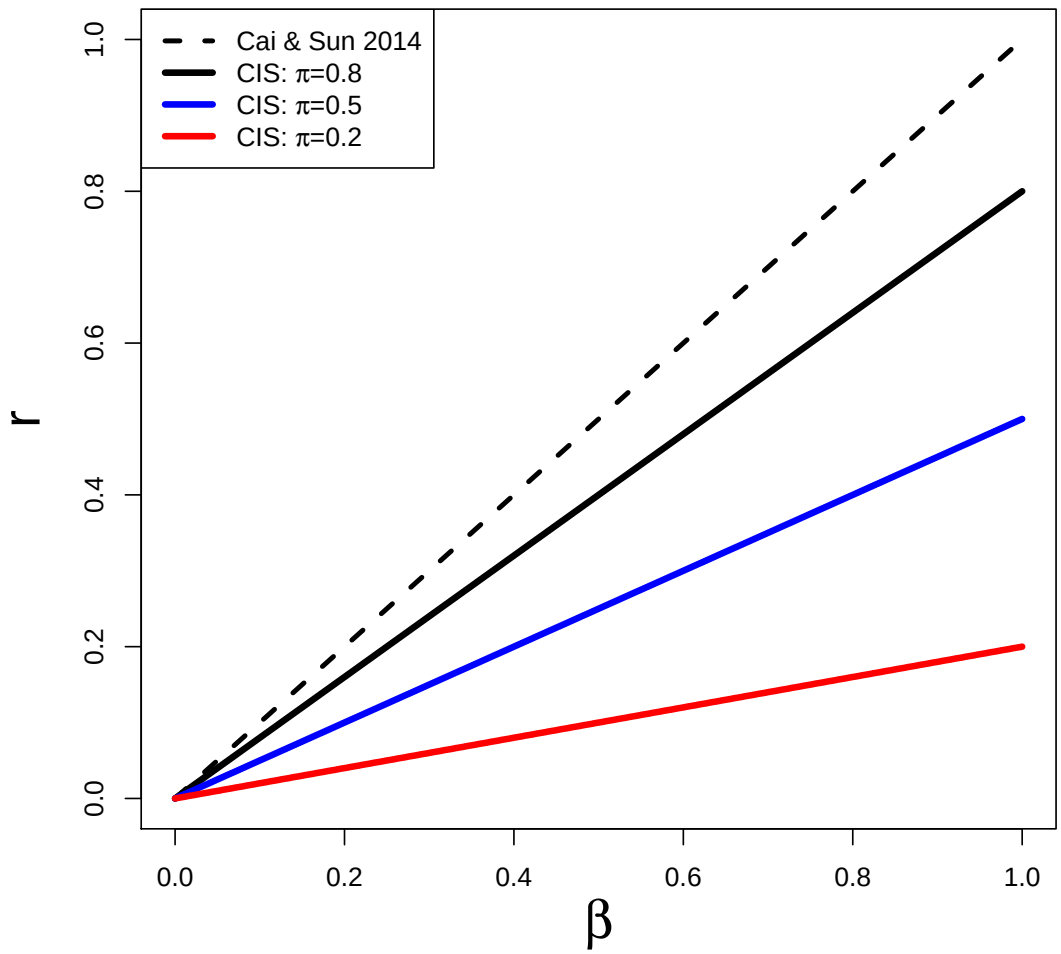
Proposition (**Classification Boundary**)

The classification boundary of the CIS classification rule $\mathbf{1}(\hat{\psi}^{CIS}(\tau, \alpha, \nu) \geq 0)$ *is also always below that of the independence rule* $\mathbf{1}(\hat{\psi}^{ind}(\tau) \geq 0)$

Discovery Boundaries



Classification Boundaries



Outline

1. Background
2. Motivation and Challenges
3. Methods
4. Simulation
5. A Data Example
6. Conclusion

3-Class Classification

- $10,000$ variables and 100 patients in each class
- The first 20 variables have mean structures between classes as in the table below
- The remaining $9,980$ variables are independently and identically distributed from $N(0,1)$

Mean of Variables

Variables	Class 1	Class 2	Class 3
$X_1 - X_4, X_{11} - X_{14}$	0	0	-2.5
X_5, X_{15}	-0.5	2	-2.5
$X_6 - X_{10}, X_{16} - X_{20}$	1.5	-1.5	-1.5

: Marginal informative feature : MUJI feature

Simulation Results

When marginally informative and MUJI signals are correlated with a compound symmetry (CS) structure with a correlation $= 0.5$

Method	class pair	FP#	FN#	se	sp	MMS	ER
$\alpha = 0.2$							
CIS	1-2	1.1	***0.1***	0.994	0.999	***24***	1.4
	2-3	9.3	0	1	0.999	19	3.1
MS	1-2	16.4	***9.1***	0.54	0.998	***8685***	3.1
	2-3	3.0	0	1	0.999	10	3.4
$\alpha = 0.5$							
CIS	1-2	11.8	***2.5***	0.88	0.999	***48***	2.7
	2-3	5.6	0	1	0.999	10	2.9
MS	1-2	18.0	**9.0**	0.55	0.998	**8634**	3.0
	2-3	3.5	0	1	0.999	10	3.4
$\alpha = 0.9$							
CIS	1-2	12.2	***7.8***	0.61	0.999	***8698***	1.6
	2-3	2.7	0	1	0.999	10	3.4
MS	1-2	17.3	***9.1***	0.54	0.998	***8876***	3.1
	2-3	2.7	0	1	0.999	10	3.4

Outline

1. Background
2. Motivation and Challenges
3. Methods
4. Simulation
5. A Data Example
6. Conclusion

Genomic Classification of Kidney Post-transplant Rejection Types

- 62 kidney tissue samples
- Four rejection types:

C	– 17 normal donor kidneys
TX	– 19 well-functioning transplants without rejections
AR	– 13 acute rejection
ADNR	– 13 acute dysfunction without rejection

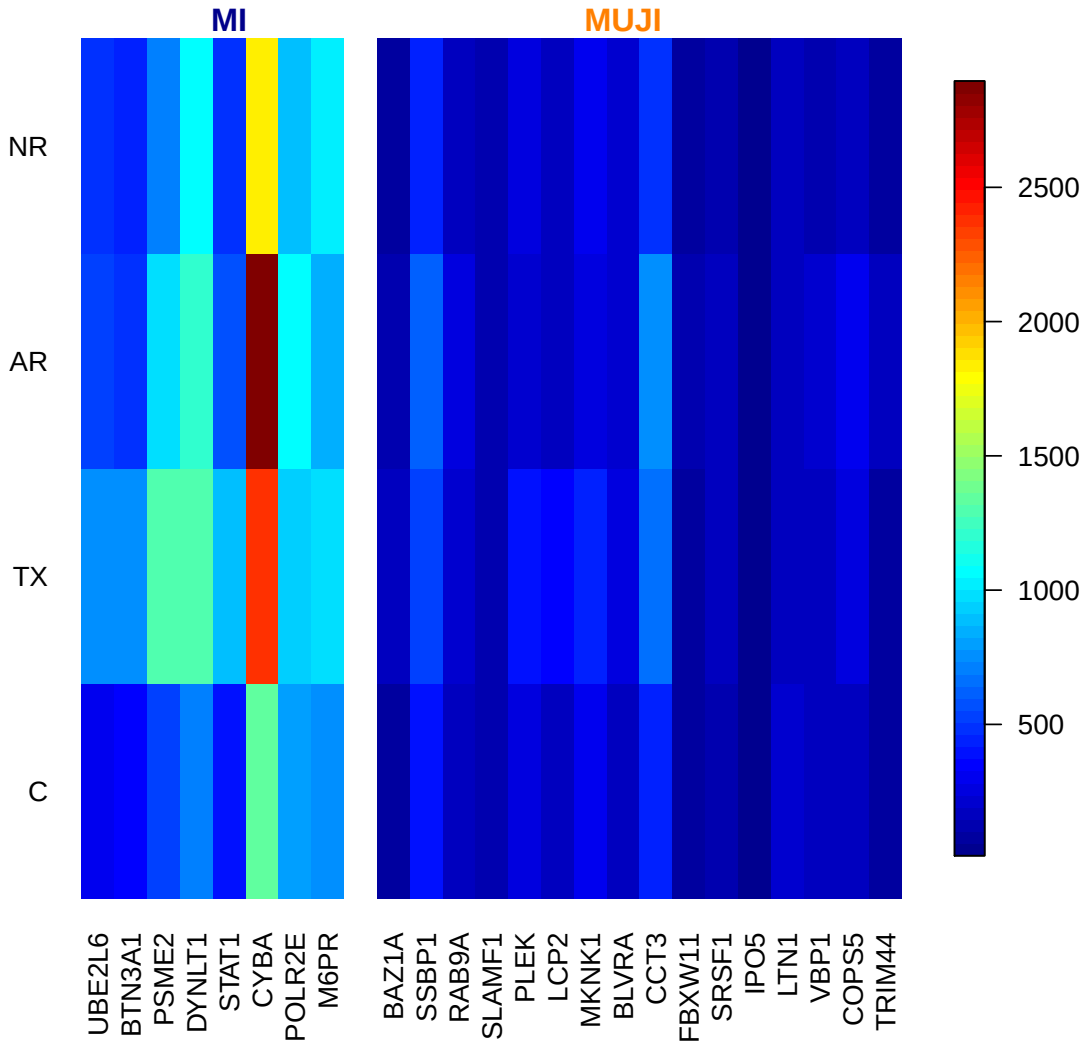
- $12,625$ gene expressions from patient kidney biopsies before transplant

Data Analysis

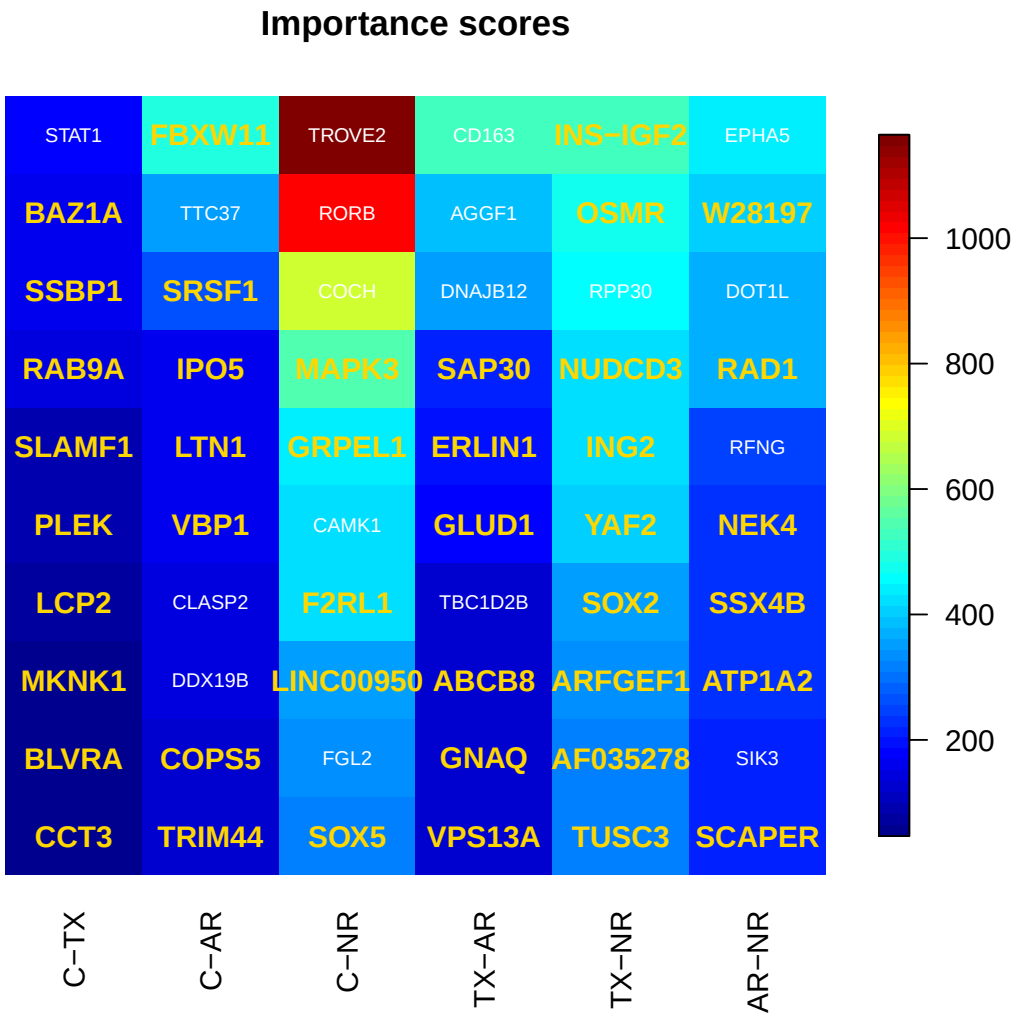
- The same data were analyzed in Xu et al. 2014
- Xu et al. 2014 method can not handle ultrahigh-dimensional data. They used an ad-hoc way to pre-select a subset of genes, and only involved 200 genes with the largest variations
- This could miss many MUJI genes
- We applied the CIS procedure for the purposes of MUJI gene identification and post-transplantation rejection type classification

Results

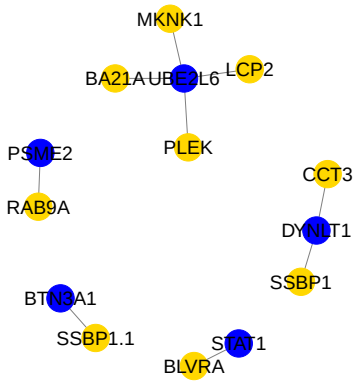
Mean expression scores for top marginally informative and MUJI genes



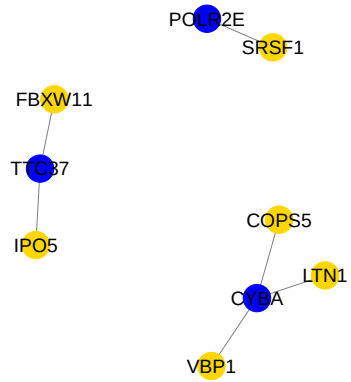
Results(Cont.)



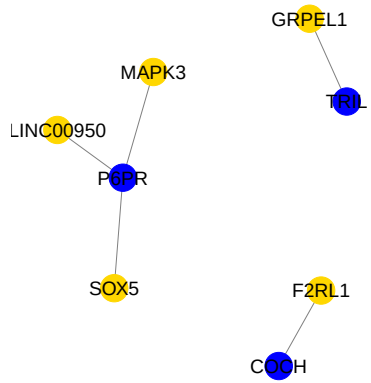
Results (Cont.)



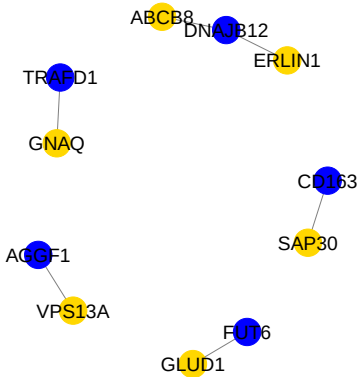
C-TX



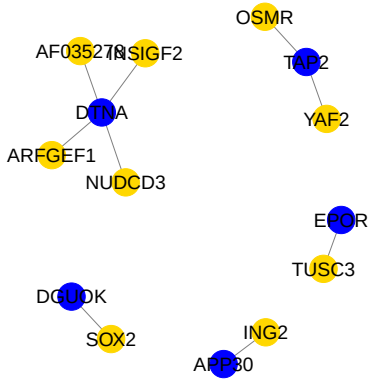
C-AR



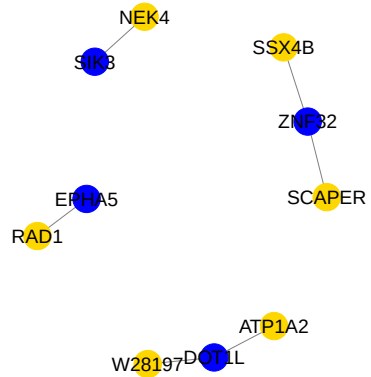
C-NR



TX-AR

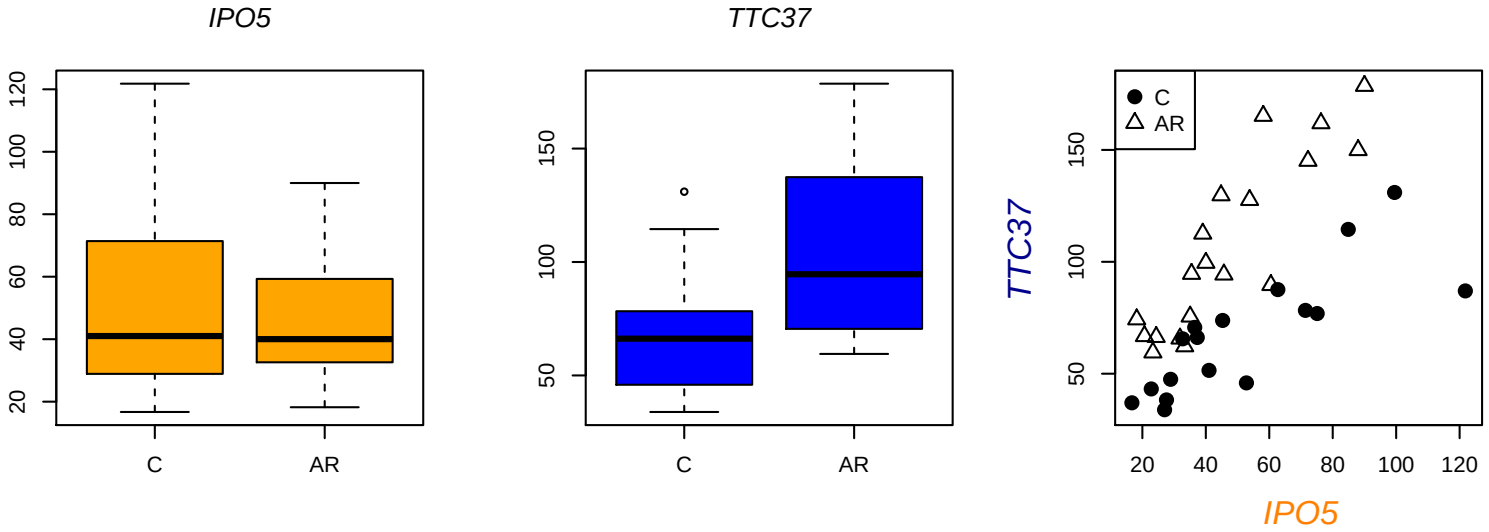


TX-NR



AR-NR

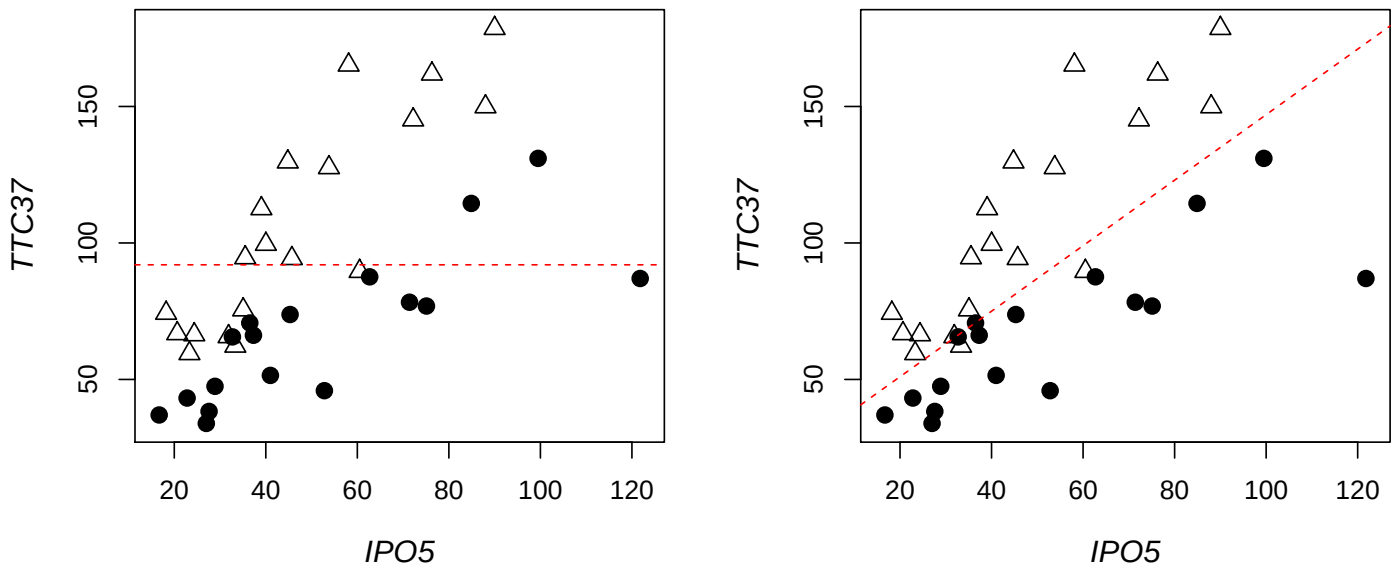
gene *IPO5*



gene j	$\hat{\delta}_j^*$ of gene	rank of gene	correlated gene	cor-relation	rank of IS_j
IPO5	0.20	10279th	*TTC37*	0.7	4th

* $\hat{\delta}_j$ = standardized marginal mean difference of gene j.

Results (Cont.)



- Left panel: based only on *TTC37*, the minimal misclassification number is **10**
- Right panel: based on *TTC37* and its correlated MUJI gene *IPO5*, the minimal misclassification number is **2**

Results (Cont.)

- Use the leave-one-out procedure to predict each individual's class
  - Leave *i*th subject out at a time and use the rest as the training set
  - Apply the classification rule $\mathbf{1}(\hat{\delta}^{\mathrm{CIS}}(\mathbf{X}_i; \rho, \tau, m) \geq 0)$ to predict the label Y_i between each pair of classes
  - Take a majority voting on the final predicted class for subject *i*
- In total 6 misclassifications out of 62 subjects

Outline

1. Background
2. Motivation and Challenges
3. Methods
4. Simulation
5. A Data Example
6. Conclusion

Summary

- Leverages inter-feature correlations but circumvent estimation of the full precision matrix
- Works well with sparse and weak signals. Can effectively select both marginally informative and MUJI signals
- With nice theoretical properties, including optimal post-screening misclassification rates and improved discovery and classification boundaries
- Can we go beyond LDA?


WHAT WE KNOW...
THE REST...
BIG DATA

Thank You!

Sub-Gaussian Distributions

Definition

A random variable X is sub-Gaussian with variance proxy σ^2 if its moment generating function satisfies

$$E[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right)$$

for any $s \in \mathbb{R}$.

A random vector $\mathbf{X} = (X_1, \cdots, X_p)$ is said to be sub-Gaussian with variance proxy σ^2 if for any $\mathbf{a} \in \mathbb{R}^p$,

$$E[\exp(\mathbf{a}'\mathbf{X})] \leq \exp\left(\frac{\sigma^2 \|\mathbf{a}\|_2^2}{2}\right).$$

Some Notations Used

- $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{A})$ the graph induced by a $p \times p$ symmetric matrix $\mathbf{A}$, where the node set $\mathcal{V} = \{1, \cdots, p\}$ corresponds to the row indices of $\mathbf{A}$ and the edge set $\mathcal{E}$ contains all the edges. An edge between nodes j and j' exists if and only if $A_{jj'} \neq 0$

- Denote $\mathcal{G}(\mathcal{V}_l, \mathcal{E}_l, \mathbf{A}_l)$ the connected components of $\mathcal{G}(\mathcal{V}, \mathcal{E}; \mathbf{A})$

Additional Simulation Results

When marginally informative and MUJI signals are correlated with a AR(1) structure and a correlation coefficient $= 0.5$

Method	class pair	FP#	FN#	se	sp	MMS	ER
$\alpha = 0.2$							
CIS	1-2	0.9	***0.05***	0.997	0.999	***31***	0.2
	2-3	9.9	0	1	0.999	17	0.9
MS	1-2	19.9	***9.6***	0.52	0.998	***8774***	2.1
	2-3	4.5	0	1	0.999	10	1.2
$\alpha = 0.5$							
CIS	1-2	9.1	***5.3***	0.74	0.999	***127***	0.6
	2-3	10.9	0	1	0.999	20	1.4
MS	1-2	11.2	***11.0***	0.45	0.999	***9354***	3.3
	2-3	4.6	0	1	0.999	10	1.4
$\alpha = 0.9$							
CIS	1-2	1.3	***9.1***	0.54	0.999	***8247***	0.4
	2-3	0.2	0.008	0.999	0.999	10	0.9
MS	1-2	1.9	***11.7***	0.41	0.999	***7470***	5.6
	2-3	0.2	0.02	0.998	0.999	10	1.0

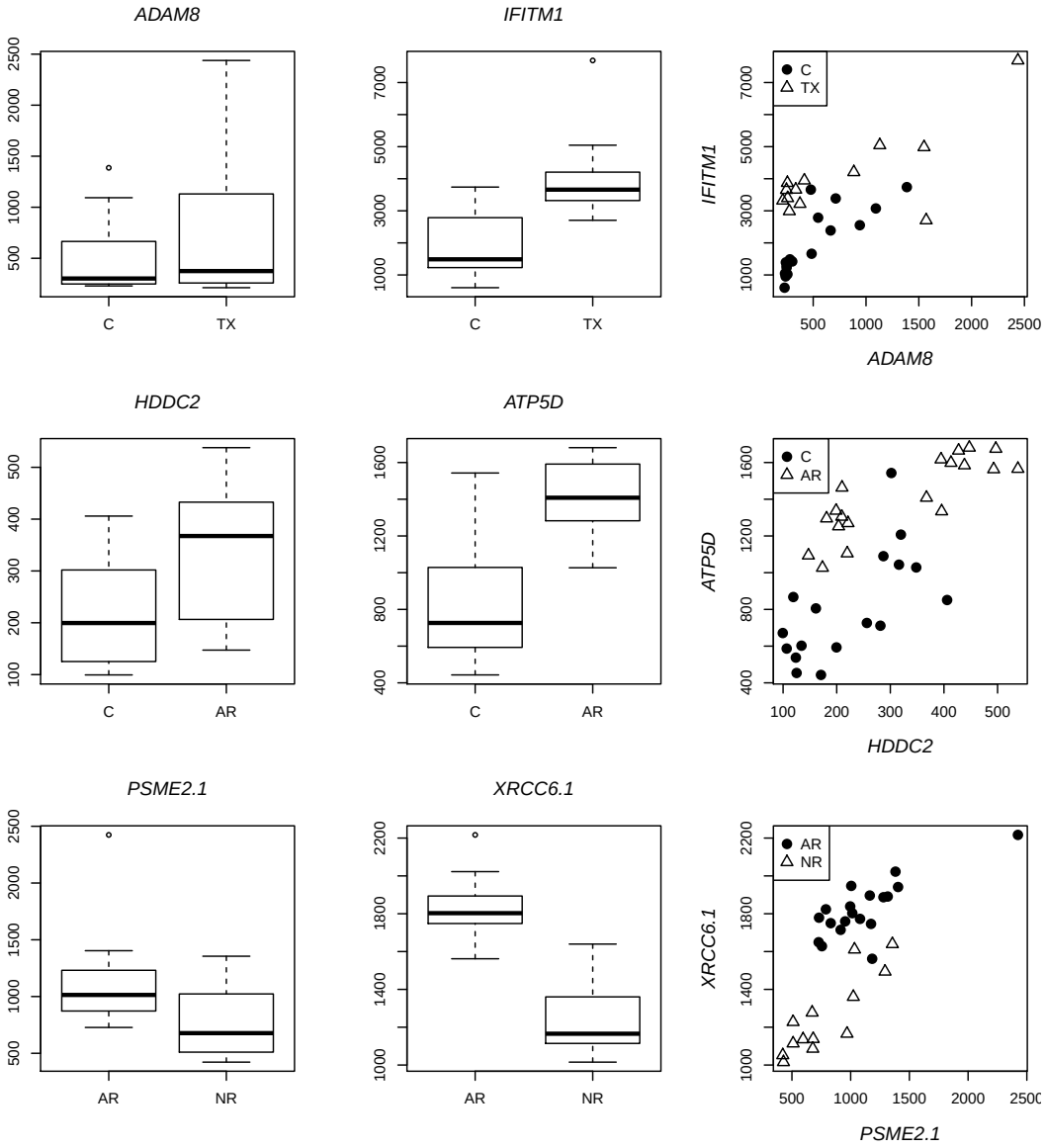
Computational efficiency

Comparisons of computing time (in minutes)

Methods	CIS-classification	ROAD	LPD	CED
$p = 200$	0.4	1.6	2.5	11.5
$p = 1000$	3.0	19.0	36.5	NA
$p = 10000$	4.5	NA	NA	NA
$p = 50000$	12.0	NA	NA	NA

* Computational time longer than 5 hours is reported as NA.

More Examples of MUJI Genes Detected



More Examples of MUJI Genes Detected

Marginal and joint effects of MUJI genes

MUJI gene	$\hat{\delta}_j^*$ of MUJI gene	rank of MUJI gene	correleted MI** gene	$\hat{\delta}_j$ of MI gene	rank of MI gene	$\hat{\rho}$	rank of IS_j
ADAM8	-0.48	5071	*IFITM1*	-1.35	22	0.70	18
HDDC2	-0.82	2497	*ATP5D*	-1.53	51	0.71	77
PSME2.1	0.84	3962	*XRCC6.1*	1.71	22	0.74	22

* $\hat{\delta}_j$ = standardized marginal mean difference of gene j. ** MI= marginally informative