Ultrahigh dimensional variable selection: Beyond the linear model

Jianqing Fan

Princeton University

With **Richard Samworth and Yichao Wu**; **Rui Song**

http://www.princeton.edu/∼jqfan

May 16, 2009

Outline

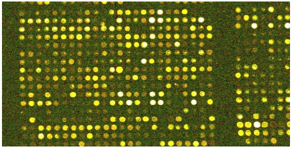
1. Introduction
2. Large-scale screening
3. Moderate-scale Selection
4. Iterative feature selection
5. Numerical Studies

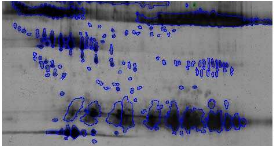
Introduction

Introduction

High-dim variable selection characterizes many contemporary statistical problems.

- Bioinformatic: disease classification using microarray, proteomics, fMRI data.



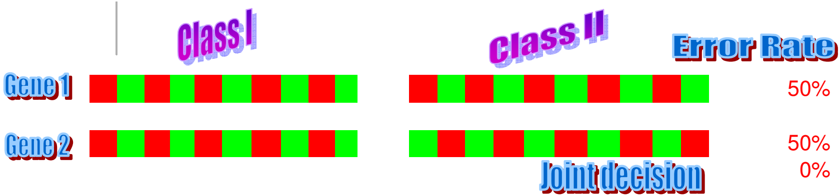


- Document or text classification: E-mail spam.
- Association studies between phenotypes and SNPs.

Growth of Dimensionality

■ Dimensionality grows rapidly with interactions

Portfolio selection and network modeling: 2,000 stocks involves over 2m unknown parameters in the covariance matrix.



Gene-gene inteaction: interactions of 5000 genes result in 12.5m features.

Aims of High-dimensional Regression and Classification

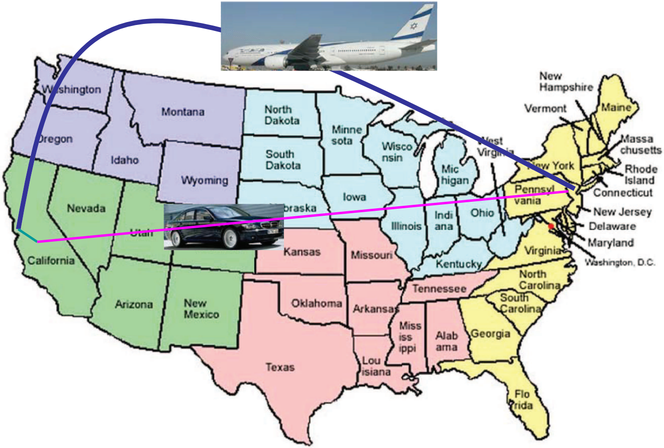
- To construct as effective a method as possible to predict future observations.

- To gain insight into the relationship between features and response for scientific purposes, as well as, hopefully, to construct an improved prediction method.

Bickel (2008) discussion of the SIS paper (JRSS-B).

Challenges with Ultrahigh Dimensionality

- Computational cost
- Estimation accuracy.
- Stability



Key idea: **Large-scale** screening and **moderate-scale** searching.

Large-scale sreening

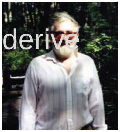
Independence learning

Regression: Feature ranking by **correlation learning** (Fan and Lv, 2008, JRSS-B). When $Y = \pm 1$, this implies

Classification: Feature ranking by two-sample t-tests or other tests (Tibshirani, et al, 03; Fan and Fan, 2008).

SIS: By an appropriate thresholding (e.g., n variables), **relevant features are in the selected set** (Fan and Lv, 08), relying on joint-normality assumption.

Other independent learning: Hall, Titterington and Xue (2009) derive such a method from empirical likelihood point of view.



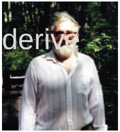
Independence learning

Regression: Feature ranking by **correlation learning** (Fan and Lv, 2008, JRSS-B). When $Y = \pm 1$, this implies

Classification: Feature ranking by two-sample t-tests or other tests (Tibshirani, et al, 03; Fan and Fan, 2008).

SIS: By an appropriate thresholding (e.g., n variables), **relevant features are in the selected set** (Fan and Lv, 08), relying on joint-normality assumption.

Other independent learning: Hall, Titterington and Xue (2009) derive such a method from empirical likelihood point of view.



Independence learning

Regression: Feature ranking by **correlation learning** (Fan and Lv, 2008, JRSS-B). When $Y = \pm 1$, this implies

Classification: Feature ranking by two-sample t-tests or other tests (Tibshirani, et al, 03; Fan and Fan, 2008).

SIS: By an appropriate thresholding (e.g., n variables), **relevant features are in the selected set** (Fan and Lv, 08), relying on joint-normality assumption.

Other independent learning: Hall, Titterington and Xue (2009) derive such a method from empirical likelihood point of view.

Model setting

GLIM: $f_Y(y|X=x;\theta) = \exp\{(y\theta - b(\theta))/\phi + c(y,\phi)\}$ with

$$\textbf{canonial link}: \qquad b'^{-1}(\mu) = \theta = \mathbf{x}^T\beta.$$

Objective: Find sparse β to minimize $Q(\beta) = \sum_{i=1}^n L(Y_i, \mathbf{x}_i^T\beta)$.

- **GLIM**: $L(Y_i, \mathbf{x}_i^T\beta) = b(\mathbf{x}_i^T\beta) - Y_i\mathbf{x}_i^T\beta$.
- **Classification**: $Y = \pm 1$.
  - ★SVM $L(Y_i, \mathbf{x}_i^T\beta) = (1 - Y_i\mathbf{x}_i^T\beta)_+$.
  - ★AdaBoost $L(Y_i, \mathbf{x}_i^T\beta) = \exp(-Y_i\mathbf{x}_i^T\beta)$.
- **Robustness**: $L(Y_i, \mathbf{x}_i^T\beta) = |Y_i - \mathbf{x}_i^T\beta|$.

Model setting

GLIM: $f_Y(y|X=x;\theta) = \exp\{(y\theta - b(\theta))/\phi + c(y,\phi)\}$ with

$$\textbf{canonial link}: \qquad b'^{-1}(\mu) = \theta = \mathbf{x}^T\beta.$$

Objective: Find sparse β to minimize $Q(\beta) = \sum_{i=1}^n L(Y_i, \mathbf{x}_i^T\beta)$.

- **GLIM**: $L(Y_i, \mathbf{x}_i^T\beta) = b(\mathbf{x}_i^T\beta) - Y_i\mathbf{x}_i^T\beta$.
- **Classification**: $Y = \pm 1$.
  - ★SVM $L(Y_i, \mathbf{x}_i^T\beta) = (1 - Y_i\mathbf{x}_i^T\beta)_+$.
  - ★AdaBoost $L(Y_i, \mathbf{x}_i^T\beta) = \exp(-Y_i\mathbf{x}_i^T\beta)$.
- **Robustness**: $L(Y_i, \mathbf{x}_i^T\beta) = |Y_i - \mathbf{x}_i^T\beta|$.

Questions

1. How to screen **discrete** variables (Genome-wide association)?
2. Do they have **sure screening** property?
3. What is the size of selected model in order to have SIS?

The arguments in Fan and Lv (2008) can not be applied here.

Questions

1. How to screen **discrete** variables (Genome-wide association)?
2. Do they have **sure screening** property?
3. What is the size of selected model in order to have SIS?

The arguments in Fan and Lv (2008) can not be applied here.

Questions

1. How to screen **discrete** variables (Genome-wide association)?
2. Do they have **sure screening** property?
3. What is the size of selected model in order to have SIS?

The arguments in Fan and Lv (2008) can not be applied here.

Independence learning

Marginal utility: Letting $\hat{L}_0 = \min_{\beta_0} n^{-1} \sum_{i=1}^n L(Y_i, \beta_0)$, define

$$\hat{L}_j = \hat{L}_0 - \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^n L(Y_i, \beta_0 + X_{ij}\beta_j) \qquad \textbf{Wilks}.$$

or $\hat{\beta}_j^M$ (**Wald**), assuming $EX_j^2 = 1$.

Feature ranking: Select features w/ **largest marginal utilities**:

$$\widehat{\mathcal{M}}_{\nu_n} = \{j : \hat{L}_j \geq \nu_n\}, \qquad \widehat{\mathcal{M}}_{\gamma_n}^W = \{j : \hat{\beta}_j^M \geq \gamma_n\}$$

Dim. reduction: From $p_n = O(\exp(n^a))$ to $O(n^b)$:

Independence learning

Marginal utility: Letting $\hat{L}_0 = \min_{\beta_0} n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0)$, define

$$\hat{L}_j = \hat{L}_0 - \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + X_{ij}\beta_j) \qquad \textbf{Wilks}.$$

or $\hat{\beta}_j^M$ (**Wald**), assuming $EX_j^2 = 1$.

Feature ranking: Select features w/ **largest marginal utilities**:

$$\widehat{\mathcal{M}}_{\nu_n} = \{j : \hat{L}_j \geq \nu_n\}, \qquad \widehat{\mathcal{M}}_{\gamma_n}^{w} = \{j : \hat{\beta}_j^M \geq \gamma_n\}$$

Dim. reduction: From $p_n = O(\exp(n^a))$ to $O(n^b)$:

Theoretical Basis – Population Aspect I

Marginal utility: $L_j^\star = E\ell(Y, \beta_0^M) - \min E\ell(Y, \beta_0 + \beta_j X_j)$.

Likelihood ratio (Fan and Song, 09)

Theorem 1: $L_j^\star = 0 \Longleftrightarrow \mathrm{cov}(Y, X_j) = \mathrm{cov}(b'(\mathbf{X}^T \beta^\star), X_j) = 0$
$\Longleftrightarrow \beta_j^M = 0$.

For Gaussian covariates, conclusion holds if $|\mathrm{cov}(\mathbf{X}^T \beta^\star, X_j)| = 0$, i.e. independence.

True model: $\mathcal{M}_\star = \{j : \beta_j^\star \neq 0\}$, where $\beta^\star = \mathrm{argmin} EL(Y, \mathbf{X}^T\beta)$.

Theorem 2: If $|\mathrm{cov}(b'(\mathbf{X}^T\beta^\star), X_j)| \geq c_1 n^{-\kappa}$ for $j \in \mathcal{M}_\star$, then

$$\min_{\mathbf{j}\in\mathcal{M}_\star} |\beta_{\mathbf{j}}^{\mathbf{M}}| \geq \mathbf{c_1 n}^{-\kappa}, \qquad \min_{\mathbf{j}\in\mathcal{M}_\star} |\mathbf{L}_{\mathbf{j}}^\star| \geq \mathbf{c_2 n}^{-2\kappa}.$$

If $\{X_j,\ j \notin \mathcal{M}_\star\}$ is independent of $\{X_i,\ i \in \mathcal{M}_\star\}$, then $L_j^\star = 0$.

For Gaussian covariates, conclusion holds if

$$|\mathrm{cov}(\mathbf{X}^{\mathbf{T}}\beta^\star, \mathbf{X_j})| \geq \mathbf{c_1 n}^{-\kappa},$$ min condition even for LS.

Theoretical Basis – Population Aspect II

True model: $\mathcal{M}_\star = \{j : \beta_j^\star \neq 0\}$, where $\beta^\star = \text{argmin} EL(Y, \mathbf{X}^T\beta)$.

Theorem 2: If $|\mathrm{cov}(b'(\mathbf{X}^T\beta^\star), X_j)| \geq c_1 n^{-\kappa}$ for $j \in \mathcal{M}_\star$, then

$$\min_{\mathbf{j}\in\mathcal{M}_\star} |\boldsymbol{\beta}_\mathbf{j}^\mathbf{M}| \geq \mathbf{c_1 n}^{-\kappa}, \qquad \min_{\mathbf{j}\in\mathcal{M}_\star} |\mathbf{L}_\mathbf{j}^\star| \geq \mathbf{c_2 n}^{-2\kappa}.$$

If $\{X_j,\ j \notin \mathcal{M}_\star\}$ is independent of $\{X_i,\ i \in \mathcal{M}_\star\}$, then $L_j^\star = 0$.

For Gaussian covariates, conclusion holds if

$|\mathrm{cov}(\mathbf{X}^\mathbf{T}\beta^\star, \mathbf{X_j})| \geq \mathbf{c_1 n}^{-\kappa}$, min condition even for LS.

Theorem 3: If $\nu_n = cn^{-2\kappa}$ for $\kappa < 1/2$, and $\log s_n = o(n^{1-2\kappa})$, then

$$P\left(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}\right) \to 1 \qquad \text{exponentially fast}$$

No conditions on covariance matrix!

- This is a SIS property w/ size controlled.
- Note that $\hat{L}_j - L_j^\star = O(\log p/\mathbf{n^{1/2}})$ and minimum signal $O(n^{-2\kappa})$. **How to deal with it**? —Appeal to the ranking invariance under monotonic transform.
- Screening using **Wald stat** $\hat{\beta}_j^M$ has SIS property.

Sampling Aspect: Sure independence screening

Theorem 3: If $\nu_n = cn^{-2\kappa}$ for $\kappa < 1/2$, and $\log s_n = o(n^{1-2\kappa})$, then

$$P\left(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}\right) \to 1 \qquad \text{exponentially fast}$$

No conditions on covariance matrix!

- This is a SIS property w/ size controlled.
- Note that $\hat{L}_j - L_j^\star = O(\log p/\mathbf{n^{1/2}})$ and minimum signal $O(n^{-2\kappa})$.
  How to deal with it? —Appeal to the ranking invariance under monotonic transform.
- Screening using **Wald stat** $\hat{\beta}_j^M$ has SIS property.

Sampling Aspect: Sure independence screening

Theorem 3: If $\nu_n = cn^{-2\kappa}$ for $\kappa < 1/2$, and $\log s_n = o(n^{1-2\kappa})$, then

$$P\left(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\nu_n}\right) \to 1 \qquad \text{exponentially fast}$$

No conditions on covariance matrix!

- This is a SIS property w/ size controlled.
- Note that $\hat{L}_j - L_j^\star = O(\log p/\mathbf{n^{1/2}})$ and minimum signal $O(n^{-2\kappa})$. **How to deal with it**? —Appeal to the ranking invariance under monotonic transform.
- Screening using **Wald stat** $\hat{\beta}_j^M$ has SIS property.

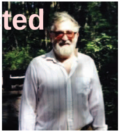
Screening by MMLE

Let $\widehat{\mathcal{M}}_{\gamma_n}^{w} = \{|\hat{\beta}_j^M| \geq \gamma_n\}$.

1. $P(\max_j |\hat{\beta}_j^M - \hat{\beta}_j^M| > c_3 n^{-\kappa}) = o(1)$, if $\log p_n = o(n^{1-2\kappa})$.
2. $P(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\gamma_n}^{w}) \to 1$, if $\gamma_n = c_0 n^{-\kappa}$, $c_0 < c_1/2$.
3. What is the selected model size? We establish

$$\|\beta^{\mathbf{M}}\|^{\mathbf{2}} = \mathbf{O}(\|\Sigma\beta^\star\|^{\mathbf{2}}) = O\{\lambda_{max}(\Sigma)\ \beta^{\star T}\Sigma\beta^\star\} = O(\lambda_{max}(\Sigma)).$$

4. The $\#\{|\beta_j^M| \geq \gamma_n\}$ is $O_P\{\gamma_n^{-2}\lambda_{max}(\Sigma)\}$, and so is the **selected model size**.



Screening by MMLE

Let $\widehat{\mathcal{M}}^w_{\gamma_n} = \{|\hat\beta^M_j| \ge \gamma_n\}$.

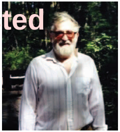
1. $P(\max_j |\hat\beta^M_j - \beta^M_j| > c_3 n^{-\kappa}) = o(1)$, if $\log p_n = o(n^{1-2\kappa})$.

2. $P(\mathcal{M}_\star \subset \widehat{\mathcal{M}}^w_{\gamma_n}) \to 1$, if $\gamma_n = c_0 n^{-\kappa}$, $c_0 < c_1/2$.

3. What is the selected model size? We establish

$$\|\beta^M\|^2 = O(\|\Sigma\beta^\star\|^2) = O\{\lambda_{max}(\Sigma)\ \beta^{\star T}\Sigma\beta^\star\} = O(\lambda_{max}(\Sigma)).$$

4. The $\#\{|\beta^M_j| \ge \gamma_n\}$ is $O_P\{\gamma_n^{-2}\lambda_{max}(\Sigma)\}$, and so is the selected model size.



Screening by MMLE

Let $\widehat{\mathcal{M}}_{\gamma_n}^{w} = \{|\hat{\beta}_j^M| \geq \gamma_n\}$.

1. $P(\max_j |\hat{\beta}_j^M - \hat{\beta}_j^M| > c_3 n^{-\kappa}) = o(1)$, if $\log p_n = o(n^{1-2\kappa})$.

2. $P(\mathcal{M}_\star \subset \widehat{\mathcal{M}}_{\gamma_n}^{w}) \to 1$, if $\gamma_n = c_0 n^{-\kappa}$, $c_0 < c_1/2$.

3. What is the selected model size? We establish

$$\|\boldsymbol{\beta}^{\mathbf{M}}\|^{\mathbf{2}} = \mathbf{O}(\|\Sigma\beta^\star\|^{\mathbf{2}}) = O\{\lambda_{max}(\Sigma)\ \beta^{\star T}\Sigma\beta^\star\} = O(\lambda_{max}(\Sigma)).$$

4. The $\#\{|\beta_j^M| \geq \gamma_n\}$ is $O_P\{\gamma_n^{-2}\lambda_{max}(\Sigma)\}$, and so is the **selected model size**.

Sampling Aspect: Controlling number of features

Theorem 4: If $\log p_n = o(n^{1-2\kappa})$,

$$\mathbf{P}[|\widehat{\mathcal{M}}_{\nu_n}| \le \mathbf{O}\{\mathbf{n}^{2\kappa}\lambda_{\max}(\Sigma)\}] \to \mathbf{1}.$$

■ Establish $\|\mathbf{L}^\star\|^2 = O(\|\beta^M\|^2) = O(\|\Sigma\beta^\star\|^2)$.

■ The number of selected covariates depends on the population covariance. It is actually bounded by

$$\mathbf{O}(\gamma_n^{-2}\|\Sigma\beta^\star\|^2) = \mathbf{O}\{\mathbf{n}^{2\kappa}\lambda_{\max}(\Sigma)\}.$$

Sampling Aspect: Controlling number of features

Theorem 4: If $\log p_n = o(n^{1-2\kappa})$,

$$\mathbf{P}[|\widehat{\mathcal{M}}_{\nu_n}| \le \mathbf{O}\{\mathbf{n}^{2\kappa}\lambda_{\max}(\Sigma)\}] \to \mathbf{1}.$$

- Establish $\|\mathbf{L}^\star\|^2 = O(\|\beta^M\|^2) = O(\|\Sigma\beta^\star\|^2)$.

- The number of selected covariates depends on the population covariance. It is actually bounded by

$$\mathbf{O}(\gamma_{\mathbf{n}}^{-2}\|\Sigma\beta^\star\|^{\mathbf{2}}) = \mathbf{O}\{\mathbf{n}^{2\kappa}\lambda_{\max}(\Sigma)\}.$$

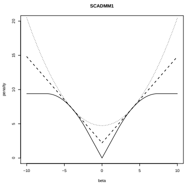
Moderate-scale selection

Moderate-scale of Model Selectors

Penalized lik.: $n^{-1}\sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{i,d}^T \beta) + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$.

Simultaneously estimate coefs and choose variables.

- Lasso (Tibshirani, 96), LARS (Efron *et al.*, 04), Adaptive Lasso(Zou, 06), Approx sparse (Huang and Zhang, 06).
- SCAD (Fan & Li, 01, 06; Fan & Peng, 04) LQA (Fan & Li, 01), MM (Hunter & Li, 05), LA (Li and Zou, 07), and PLUS (Zhang, 07).



Dantzig selector (Candes & Tao, 07)

$$\min_{\beta \in \mathbf{R}^{p_n}} \|\beta\|_1 \quad \text{subject to} \quad \|\mathbf{X}^T\mathbf{r}\|_\infty \leq \lambda_{p_n}\sigma$$

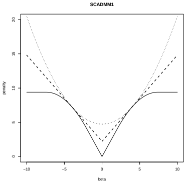
with $\lambda_{p_n} > 0$, $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$ and σ noise level. $\approx$ **Lasso** (Bickel, et al, 2008)

Moderate-scale of Model Selectors

Penalized lik.: $n^{-1}\sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{i,d}^T \beta) + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$.

Simultaneously estimate coefs and choose variables.

- Lasso (Tibshirani, 96), LARS (Efron *et al.*, 04),
  Adaptive Lasso(Zou, 06), Approx sparse (Huang and Zhang, 06).
- SCAD (Fan & Li, 01, 06; Fan & Peng, 04)
  LQA (Fan & Li, 01), MM (Hunter & Li, 05),
  LA (Li and Zou, 07), and PLUS (Zhang, 07).



Dantzig selector (Candes & Tao, 07)

$$\min_{\beta \in \mathbf{R}^{p_n}} \|\beta\|_1 \quad \text{subject to} \quad \|\mathbf{X}^T \mathbf{r}\|_\infty \le \lambda_{p_n} \sigma$$

with $\lambda_{p_n} > 0$, $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$ and σ noise level. $\approx$ **Lasso** (Bickel, et al, 2008)

Moderate-scale of Model Selectors

Penalized lik.: $n^{-1}\sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{i,d}^T \beta) + \sum_{j=1}^{d} p_\lambda(|\beta_j|)$.

Simultaneously estimate coefs and choose variables.

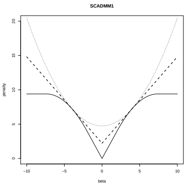
- Lasso (Tibshirani, 96), LARS (Efron *et al.*, 04),

  Adaptive Lasso(Zou, 06), Approx sparse (Huang and Zhang, 06).

- SCAD (Fan & Li, 01, 06; Fan & Peng, 04)

  LQA (Fan & Li, 01), MM (Hunter & Li, 05),

  LA (Li and Zou, 07), and PLUS (Zhang, 07).

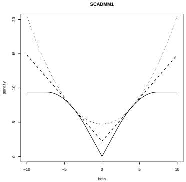


Dantzig selector (Candes & Tao, 07)

$$\min_{\beta \in \mathbf{R}^{p_n}} \|\beta\|_1 \quad \text{subject to} \quad \left\|\mathbf{X}^T\mathbf{r}\right\|_\infty \leq \lambda_{p_n}\sigma$$

with $\lambda_{p_n} > 0$, $\mathbf{r} = \mathbf{y} - \mathbf{X}\beta$ and σ noise level. $\approx$ **Lasso** (Bickel, et al, 2008)

Connections among penalized least-squares



PLS: $\|\mathbf{y}-\mathbf{X}\beta\|^2+\sum_{i=1}^{p_n} p_\lambda(|\beta_i|)$.

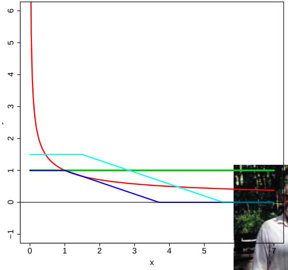
LLA: with initial value β_0 (Zou & Li, 08),

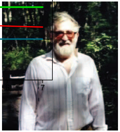
$$\|\mathbf{y}-\mathbf{X}\beta\|^2+\sum_{i=1}^{p_n}\{p_\lambda(|\beta_{i,0}|)+p_\lambda(|\beta_{i,0}|)'(|\beta_i|-|\beta_{i,0}|)\}.$$

Weighted L_1: $\|\mathbf{y}-\mathbf{X}\beta\|^2+\sum_{i=1}^{p_n} \mathbf{w}(|\beta_{\mathbf{i,0}}|)|\beta_i|$.

Fan and Li (01) stressed the unbiasedness.

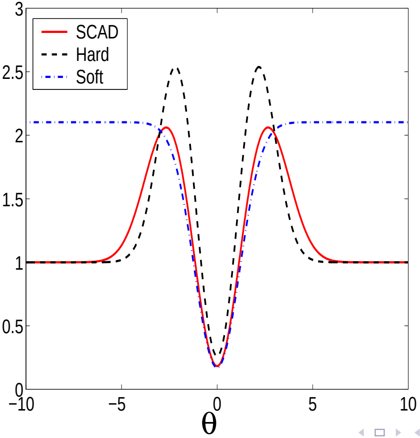
Convergence: Objective function decreasing.





Risk Comparisons of popularized least-sqaures

- Penalized least-squares: $(Z-\theta)^2 + p_\lambda(|\theta|)$
- $R(\hat\theta,\theta) = E_\theta(\hat\theta-\theta)^2$ with $Z \sim N(\theta,1)$
- $\lambda = 2$ for hard thresholding



Iterative feature selection

Drawback of Independence Screening

False negative: The features such that $\mathrm{cov}(X_j, \mathbf{X}^T\beta^\star) = 0$ can not be selected, but this can be a **signature variable**.

Example: If $\{X_j\}_{j=1}^{J}$ has common correlation ρ, then

$$\mathrm{cov}(\mathbf{X_{J+1}}, X_1 + \cdots + X_J - \mathbf{J\rho X_{J+1}}) = 0.$$

False positive: Rank too high predictors **jointly unimportant** but marginally important:

$$\mathrm{cov}(\mathbf{X_{J+1}}, X_1 + \cdots + X_J - 0.2X_{p+1}) = J\rho.$$

False negative: The features such that $\mathrm{cov}(X_j, \mathbf{X}^T\beta^\star) = 0$ can not be selected, but this can be a **signature variable**.

Example: If $\{X_j\}_{j=1}^J$ has common correlation ρ, then

$$\mathrm{cov}(\mathbf{X_{J+1}}, X_1 + \cdots + X_J - \mathbf{J\rho X_{J+1}}) = 0.$$

False positive: Rank too high predictors **jointly unimportant** but marginally important:

$$\mathrm{cov}(\mathbf{X_{J+1}}, X_1 + \cdots + X_J - 0.2X_{p+1}) = J\rho.$$

Drawback of Independence Screening

False negative: The features such that $\text{cov}(X_j, \mathbf{X}^T\beta^\star) = 0$ can not be selected, but this can be a **signature variable**.

Example: If $\{X_j\}_{j=1}^J$ has common correlation ρ, then

$$\text{cov}(\mathbf{X_{J+1}}, X_1 + \cdots + X_J - \mathbf{J\rho X_{J+1}}) = 0.$$

False positive: Rank too high predictors **jointly unimportant** but marginally important:

$$\text{cov}(\mathbf{X_{J+1}}, X_1 + \cdots + X_J - 0.2X_{p+1}) = J\rho.$$

Iterative feature selection

1. **(Large-scale screening)**: Apply SIS to pick a set $\mathcal{A}_1$;
   (Moderate-scale selection): Employ a penalized likelihood to select a subset $\mathcal{M}_1$ of these indices.

2. **(Large-scale screening)**: Rank features according to the additional (**conditional**) contribution:

$$L_j^{(2)} = \min_{\beta_0, \beta_{\mathcal{M}_1}, \beta_j} n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{\mathbf{i},\mathcal{M}_1}^{\mathbf{T}} \beta_{\mathcal{M}_1} + X_{ij}\beta_j).$$

   —Resulting in new feature sets $\mathcal{A}_2$.
   —An improvement over Fan and Lv (08) who set $\beta_{\mathcal{M}_1} = \hat{\beta}_{\mathcal{M}_1}$ from previous fit.

Iterative feature selection

1. **(Large-scale screening)**: Apply SIS to pick a set $\mathcal{A}_1$;
   (Moderate-scale selection): Employ a penalized likelihood to select a subset $\mathcal{M}_1$ of these indices.

2. **(Large-scale screening)**: Rank features according to the additional (**conditional**) contribution:

$$L_j^{(2)} = \min_{\beta_0, \beta_{\mathcal{M}_1}, \beta_j} n^{-1} \sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{\mathbf{i},\mathcal{M}_1}^{\mathbf{T}} \beta_{\mathcal{M}_1} + X_{ij}\beta_j).$$

—Resulting in new feature sets $\mathcal{A}_2$.

—An improvement over Fan and Lv (08) who set $\beta_{\mathcal{M}_1} = \hat{\beta}_{\mathcal{M}_1}$ from previous fit.

Iterative feature selection (II)

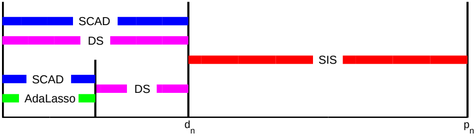
3. **(Moderate-scale selection)**: Minimize wrt $\beta_{\mathcal{M}_1}$, $\beta_{\mathcal{A}_2}$

$$\sum_{i=1}^{n} L(Y_i, \beta_0 + \mathbf{x}_{i,\mathcal{M}_1}^T \beta_{\mathcal{M}_1} + \mathbf{x}_{i,\mathcal{A}_2}^T \beta_{\mathcal{A}_2}) + \sum_{j \in \mathcal{M}_1 \cup \mathcal{A}_2} p_\lambda(|\beta_j|).$$

—Resulting in $\mathcal{M}_2$

—**Allow deletion**, improvement over ISIS (Fan and Lv, 08).

4. Repeat Steps 1–3 until $|\mathcal{M}_\ell| = d$ (prescribed) or $\mathcal{M}_\ell = \mathcal{M}_{\ell-1}$.



Reduction of false selection rates

Variant 1: ■Randomly split samples to obtain $\widehat{\mathscr{A}}^{(1)}$ and $\widehat{\mathscr{A}}^{(2)}$.
■Take $\widehat{\mathscr{A}} = \widehat{\mathscr{A}}^{(1)} \cap \widehat{\mathscr{A}}^{(2)}$.

Intuition: If both have SIS property, so does $\widehat{\mathscr{A}}$ with lower FSR.

Theorem 1: With prescribed d,

$$P(|\widehat{\mathscr{A}} \cap \mathcal{M}_\star^c| \geq r) \leq \frac{\binom{d}{r}^2}{\binom{p-|\mathcal{M}_\star|}{r}} \leq \frac{1}{r!}\left(\frac{d^2}{p-|\mathcal{M}_\star|}\right)^r,$$

—Blessing of dimensionality!

Variant 2: Recruit as many variables into equal-sized sets $\widetilde{\mathscr{A}}^{(1)}$ and $\widetilde{\mathscr{A}}^{(2)}$ as required such that $|\widehat{\mathscr{A}}| = d$ (prescribed).

Reduction of false selection rates

Variant 1: ■Randomly split samples to obtain $\widehat{\mathcal{A}}^{(1)}$ and $\widehat{\mathcal{A}}^{(2)}$.

■Take $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}^{(1)} \cap \widehat{\mathcal{A}}^{(2)}$.

Intuition: If both have SIS property, so does $\widehat{\mathcal{A}}$ with lower FSR.

Theorem 1: With prescribed d,

$$P(|\widehat{\mathcal{A}} \cap \mathcal{M}_\star^c| \geq r) \leq \frac{\binom{d}{r}^2}{\binom{p-|\mathcal{M}_\star|}{r}} \leq \frac{1}{r!}\left(\frac{d^2}{p-|\mathcal{M}_\star|}\right)^r,$$

—Blessing of dimensionality!

Variant 2: Recruit as many variables into equal-sized sets $\widetilde{\mathcal{A}}^{(1)}$ and $\widetilde{\mathcal{A}}^{(2)}$ as required such that $|\widehat{\mathcal{A}}| = d$ (prescribed).

Reduction of false selection rates

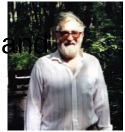
Variant 1: ■Randomly split samples to obtain $\widehat{\mathcal{A}}^{(1)}$ and $\widehat{\mathcal{A}}^{(2)}$.
■Take $\widehat{\mathcal{A}} = \widehat{\mathcal{A}}^{(1)} \cap \widehat{\mathcal{A}}^{(2)}$.
Intuition: If both have SIS property, so does $\widehat{\mathcal{A}}$ with lower FSR.

Theorem 1: With prescribed d,

$$P(|\widehat{\mathcal{A}} \cap \mathcal{M}_\star^c| \ge r) \le \frac{\binom{d}{r}^2}{\binom{p-|\mathcal{M}_\star|}{r}} \le \frac{1}{r!}\left(\frac{d^2}{p-|\mathcal{M}_\star|}\right)^r,$$

—Blessing of dimensionality!

Variant 2: Recruit as many variables into equal-sized sets $\widetilde{\mathcal{A}}^{(1)}$ and $\widetilde{\mathcal{A}}^{(2)}$ as required such that $|\widehat{\mathcal{A}}| = d$ (prescribed).



Numerical Studies

Design of Simulations

Contexts: ★Logistic ★Poission ★L_1-reg; ★Multiclass SVM

Covariates: $p = 1000, \qquad X_i \sim N(0,1)$.

1. $X_1, \ldots, X_p \sim_{i.i.d.} N(0,1)$
2. $\text{corr}(X_i, X_4) = 1/\sqrt{2}$ and otherwise $\text{corr}(X_i, X_j) = 1/2$.
3. The same except $\text{corr}(X_i, X_{p+1}) = 0$.

Logistic regression, independent covariate

$\beta_1 = 1.24$, $\beta_2 = -1.34$, $\beta_3 = -1.35$, $\beta_4 = -1.80$, $\beta_5 = -1.58$, $\beta_6 = -1.60$.

Bayes test error: 0.1368. $n = 400$, $N_{\text{sim}} = 100$.

	SIS	ISIS	Var2-SIS	LASSO	NSC
$\text{med}(\|\beta - \widehat{\beta}\|_1)$	**1.11**	1.25	1.21	**8.48**	N/A
$\text{med}(\|\beta - \widehat{\beta}\|_2^2)$	0.49	0.52	0.52	1.70	N/A
True positive	**0.99**	0.84	0.91	**1.00**	**0.34**
Med. model size	**6**	6	6	**94**	**3**
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	237	247	243	164	N/A
AIC	250	260	256	353	N/A
BIC	278	285	282	725	N/A
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	272	273	273	319	N/A
0-1 test error	**0.14**	0.14	0.14	**0.17**	**0.36**

Logistic regression, difficult case — false negative

$\beta_1 = 4$, $\beta_2 = 4$, $\beta_3 = 4$, $\beta_4 = -6\sqrt{2}$, $\mathrm{cov}(X_4, \mathbf{x}^T\beta^\star) = 0$.

Signature variable: Bayes error: **0.107** and **.344** w/ and w/o X_4.

	Van-SIS	ISIS	Var2-ISIS	LASSO	NSC
$\mathrm{med}(\|\beta - \widehat{\beta}\|_1)$	**20.1**	**1.94**	1.85	**21.6**	N/A
$\mathrm{med}(\|\beta - \widehat{\beta}\|_2^2)$	9.41	1.05	0.98	9.11	N/A
True positive	**0.00**	**1.00**	1.00	**0.00**	**0.21**
Med. model size	**16**	**4**	4	**91**	**16.5**
$2Q(\hat{\beta}_0, \widehat{\beta})$(training)	307	187	187	127	N/A
AIC	334	196	195	311	N/A
BIC	386	212	212	672	N/A
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	344	204	204	259	N/A
0-1 test error	**.193**	**.109**	.109	**0.141**	**0.377**

Logistic, the most difficult case

$\beta_1 = 4$, $\beta_2 = 4$, $\beta_3 = 4$, $\beta_4 = -6\sqrt{2}$, $\beta_{p+1} = 4/3$, $\mathrm{cov}(X_4, \mathbf{X}^T \beta^\star) = 0$.

Bayes error: 0.1040.

	Van-SIS	ISIS	Var2-ISIS	LASSO	NSC
$\mathrm{med}(\|\beta - \widehat{\beta}\|_1)$	**20.6**	**2.69**	3.24	**23.2**	N/A
$\mathrm{med}(\|\beta - \widehat{\beta}\|_2^2)$	9.46	1.36	1.59	9.11	N/A
True Positive	**0.00**	**0.90**	0.98	**0.00**	**0.17**
Med. model size	**16**	**5**	5	**102**	10
$2Q(\hat{\beta}_0, \widehat{\beta})$(training)	269	188	188	109	N/A
AIC	289	198	199	311	N/A
BIC	337	218	219	714	N/A
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	361	225	226	276	N/A
0-1 test error	**.193**	**.112**	.112	**.146**	**.387**

Possion, independent covariates

$\beta_0 = 5$, $\beta_1 = -0.54$, $\beta_2 = 0.53$, $\beta_3 = -0.50$, $\beta_4 = -0.49$, $\beta_5 = -0.41$, $\beta_6 = 0.52$, $n = 200$, $N_{\text{sim}} = 100$.

	SIS	ISIS	Var2-ISIS	LASSO
$\text{med}(\|\beta - \widehat{\beta}\|_1)$	.070	.124	.122	.197
$\text{med}(\|\beta - \widehat{\beta}\|_2^2)$	.023	.032	.033	.054
True Positive	.76	1.00	1.00	1.00
Med. model size	12	18	17	27
$2Q(\hat{\beta}_0, \widehat{\beta})$(training)	1561	1502	1510	1534
AIC	1586	1538	1542	1587
BIC	1627	1597	1595	1674
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	1558	1594	1589	1645

Poisson Regression, difficult case

$\beta_0 = 5$, $\beta_1 = 0.6$, $\beta_2 = 0.6$, $\beta_3 = 0.6$, $\beta_4 = -0.9\sqrt{2}$

$\mathrm{cov}(X_4, \mathbf{X}^T\beta^\star) = 0$.

	ISIS	Var2-ISIS	LASSO
$\mathrm{med}(\|\beta - \widehat{\beta}\|_1)$	**.271**	.225	**3.07**
$\mathrm{med}(\|\beta - \widehat{\beta}\|_2^2)$	.072	.068	**1.29**
True positive	**1.00**	.97	**0.00**
Median final model size	**18**	16	**174**
$2Q(\hat{\beta}_0, \widehat{\beta})$(training)	1494	1509	1364
AIC	1531	1541	1718
BIC	1590	1596	2293
$2Q(\hat{\beta}_0, \widehat{\beta})$(test)	**1629**	1615	**2213**

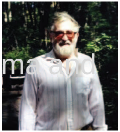
Poisson Regression, the most difficult case

$\beta_0 = 5$, $\beta_1 = 0.6$, $\beta_2 = 0.6$, $\beta_3 = 0.6$, $\beta_4 = -0.9\sqrt{2}$, $\beta_{p+1} = -0.15$
$\mathrm{cov}(X_4, \mathbf{X}^T\beta^\star) = 0$.

	Van-ISIS	Var2-ISIS	LASSO
$\mathrm{med}(\|\beta - \widehat{\beta}\|_1)$	**.254**	.232	**3.09**
$\mathrm{med}(\|\beta - \widehat{\beta}\|_2^2)$	**.068**	.068	**1.29**
True positive	**.97**	.91	**0.00**
Median final model size	**18**	16	**174**
$2Q(\hat{\beta}_0, \widehat{\beta})$ (training)	1500	1516	1367
AIC	1536	1547	1715
BIC	1595	1600	2294
$2Q(\hat{\beta}_0, \widehat{\beta})$ (test)	**1640**	1631	2389

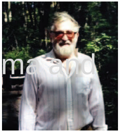
Neuroblastoma Data (MAQC-II)

1. 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004, aged from 0 to 296 months (median 15 months).
2. Neuroblastoma is a common paediatric solid cancer (15%)
3. 251 customized oligonucleotide microarray with $p = 10,707$.
4. focus on "3-year Event Free Survival", —whether each patient survived 3 years after the diagnosis of neuroblastoma ($n = 239$ w/ 49 "$+$" and 190 "$-$").
5. Aims: To study which genes are responsible for neuroblastoma and its risk association.



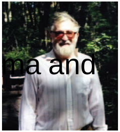
Neuroblastoma Data (MAQC-II)

1. 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004, aged from 0 to 296 months (median 15 months).
2. Neuroblastoma is a common paediatric solid cancer (15%)
3. 251 customized oligonucleotide microarray with $p = 10,707$.
4. focus on "3-year Event Free Survival", —whether each patient survived 3 years after the diagnosis of neuroblastoma ($n = 239$ w/ 49 "$+$" and 190 "$-$").
5. Aims: To study which genes are responsible for neuroblastoma and its risk association.



Neuroblastoma Data (MAQC-II)

1. 251 patients of the German Neuroblastoma Trials NB90-NB2004, diagnosed between 1989 and 2004, aged from 0 to 296 months (median 15 months).
2. Neuroblastoma is a common paediatric solid cancer (15%)
3. 251 customized oligonucleotide microarray with $p = 10,707$.
4. focus on "3-year Event Free Survival", —whether each patient survived 3 years after the diagnosis of neuroblastoma ($n = 239$ w/ 49 "$+$" and 190 "$-$").
5. Aims: To study which genes are responsible for neuroblastoma and its risk association.



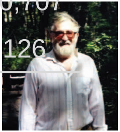
Results

Training set and endpoints:

1. **"3-y EFS"**: Random $n = 125$ subjects (25 "$+$" and 100 "$-$").
2. **"Gender"**: Random 120 males and 50 females. Total: 246.

Testing set: The remainder are used as the testing set.

Object	Method	SIS	ISIS	var2-ISIS	LASSO	NSC	Total
3-y EFS	No. pred.	5	23	12	57	9413	10,707
	Test error	19	22	21	22	24	114
Gender	No. pred.	6	2	2	42	3	10,707
	Test error	4	4	4	5	4	126



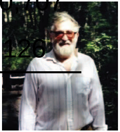
Results

Training set and endpoints:

1. **"3-y EFS"**: Random $n = 125$ subjects (25 "$+$" and 100 "$-$").
2. **"Gender"**: Random 120 males and 50 females. Total: 246.

Testing set: The remainder are used as the testing set.

Object	**Method**	SIS	ISIS	var2-ISIS	LASSO	NSC	Total
3-y EFS	No. pred.	5	23	12	57	9413	10,707
	Test error	19	22	21	22	24	114
Gender	No. pred.	6	2	2	42	3	10,707
	Test error	4	4	4	5	4	126



Multi-category Classification

Linear classifier: $\text{argmax}_k f_k(\mathbf{x})$, where $f_k(\mathbf{x}) \equiv \beta_{0k} + \mathbf{x}^T \beta_k$.

Loss: $L(Y, \mathbf{f}(\mathbf{x}; \mathbf{B})) = \sum_{j \neq Y} [1 + f_j(\mathbf{x})]_+$

Marginal utility of the j-feature (Lee et al, 2004; Liu, et al, 2007):
$L_j = \min_{\mathbf{B}} \sum_{i=1}^n L(Y_i, \mathbf{f}(X_{ij}, \mathbf{B})) + \frac{1}{2} \sum_k \beta_{jk}^2$ (identifiability)

Simulation Experiments

Design: $\tilde{X}_1, \ldots, \tilde{X}_4$ U$[-\sqrt{3}, \sqrt{3}]$, and $\tilde{X}_5, \ldots, \tilde{X}_p \sim N(0,1)$.

Case 1: $X_j = \tilde{X}_j$ for $j = 1, \ldots, p$

Case 2: $X_1 = \tilde{X}_1 - \sqrt{2}\tilde{X}_5$, $X_2 = \tilde{X}_2 + \sqrt{2}\tilde{X}_5$, $X_3 = \tilde{X}_3 - \sqrt{2}\tilde{X}_5$,

$X_4 = \tilde{X}_4 + \sqrt{2}\tilde{X}_5$,

$X_j = \sqrt{3}\tilde{X}_j$ for $j = 5, \ldots, p$.

Response: 4 categories $P(Y = k | \tilde{\mathbf{X}} = \tilde{\mathbf{x}}) \propto \exp\{f_k(\tilde{\mathbf{x}})\}$,

$f_1(\tilde{\mathbf{x}}) = -a\tilde{x}_1 + a\tilde{x}_4$, $f_2(\tilde{\mathbf{x}}) = a\tilde{x}_1 - a\tilde{x}_2$,

$f_3(\tilde{\mathbf{x}}) = a\tilde{x}_2 - a\tilde{x}_3$ and $f_4(\tilde{\mathbf{x}}) = a\tilde{x}_3 - a\tilde{x}_4$ with $a = 5/\sqrt{3}$.

Simulation Experiments

Design: $\tilde{X}_1, \ldots, \tilde{X}_4$ U$[-\sqrt{3}, \sqrt{3}]$, and $\tilde{X}_5, \ldots, \tilde{X}_p \sim N(0,1)$.

Case 1: $X_j = \tilde{X}_j$ for $j = 1, \ldots, p$

Case 2: $X_1 = \tilde{X}_1 - \sqrt{2}\tilde{X}_5$, $X_2 = \tilde{X}_2 + \sqrt{2}\tilde{X}_5$, $X_3 = \tilde{X}_3 - \sqrt{2}\tilde{X}_5$,
$X_4 = \tilde{X}_4 + \sqrt{2}\tilde{X}_5$,
$X_j = \sqrt{3}\tilde{X}_j$ for $j = 5, \ldots, p$.

Response: 4 categories $P(Y = k | \widetilde{\mathbf{X}} = \tilde{\mathbf{x}}) \propto \exp\{f_k(\tilde{\mathbf{x}})\}$,
$f_1(\tilde{\mathbf{x}}) = -a\tilde{x}_1 + a\tilde{x}_4$, $f_2(\tilde{\mathbf{x}}) = a\tilde{x}_1 - a\tilde{x}_2$,
$f_3(\tilde{\mathbf{x}}) = a\tilde{x}_2 - a\tilde{x}_3$ and $f_4(\tilde{\mathbf{x}}) = a\tilde{x}_3 - a\tilde{x}_4$ with $a = 5/\sqrt{3}$.

Simulation results, $n = 400$

	SIS	ISIS	Var2-ISIS	LASSO	NSC
			Case 1		
True positive	**1.00**	**1.00**	1.00	**0.00**	**0.68**
Median modal size	**2.5**	**4**	5	**19**	**4**
0-1 test error	**0.306**	**.301**	.292	**.330**	**.452**
Standard error	.007	.006	.006	.008	.021
			Case 2		
True positive	**.10**	**1.00**	1.00	**.33**	**.30**
Median modal size	**4**	**11**	9	**54**	**9**
0-1 test error	**.436**	**.304**	.298	**.430**	**.624**
Standard error	.007	.007	.006	.004	.008

Test errors: based on $200n$ cases.

Children Cancer Data

Classification: ★neuroblastoma (NB), ★rhabdomyosarcoma (RMS), ★non-Hodgkin lymphoma (NHL), ★Ewing family of tumors (EWS).

Data: cDNA microarrays with 2308 genes (from 6567).

- Training: 63 (12 NBs, 20 RMSs, 8 NHLs, and 23 EWS)
- Testing: 20 (6 NBs, 5 RMSs, 3 NHLs, and 6 EWS)

Results: All methods have zero testing errors.

Method	ISIS	var2-ISIS	LASSO	NSC
# selected genes	15	14	71	343

Children Cancer Data

Classification: ★neuroblastoma (NB), ★rhabdomyosarcoma (RMS), ★non-Hodgkin lymphoma (NHL), ★Ewing family of tumors (EWS).

Data: cDNA microarrays with 2308 genes (from 6567).

- Training: 63 (12 NBs, 20 RMSs, 8 NHLs, and 23 EWS)
- Testing: 20 (6 NBs, 5 RMSs, 3 NHLs, and 6 EWS)

Results: All methods have zero testing errors.

Method	ISIS	var2-ISIS	LASSO	NSC
# selected genes	15	14	71	343

Summary and Conclusion

1. Propose large scale-screening and moderate-selection
   - Use conditional independence screening.
   - Allow variable deletion in the process.
   - Estimation accuracy, comp expediency, algorithmic stability.
2. Applicable to many contexts: ★GLIM; ★Robust; ★Machine learning
3. Demonstrate its utility via extensive simulation. Handle well the most difficulty case.
4. Provide theoretical foundation to independence learning.

Summary and Conclusion

1. Propose large scale-screening and moderate-selection
   - Use conditional independence screening.
   - Allow variable deletion in the process.
   - Estimation accuracy, comp expediency, algorithmic stability.
2. Applicable to many contexts: ★GLIM; ★Robust; ★Machine learning
3. Demonstrate its utility via extensive simulation. Handle well the most difficulty case.
4. Provide theoretical foundation to independence learning.

Summary and Conclusion

1. Propose large scale-screening and moderate-selection
   - Use conditional independence screening.
   - Allow variable deletion in the process.
   - Estimation accuracy, comp expediency, algorithmic stability.
2. Applicable to many contexts: ★GLIM; ★Robust; ★Machine learning
3. Demonstrate its utility via extensive simulation. Handle well the most difficulty case.
4. Provide theoretical foundation to independence learning.

Summary and Conclusion

1. Propose large scale-screening and moderate-selection
   - Use conditional independence screening.
   - Allow variable deletion in the process.
   - Estimation accuracy, comp expediency, algorithmic stability.
2. Applicable to many contexts: ★GLIM; ★Robust; ★Machine learning
3. Demonstrate its utility via extensive simulation. Handle well the most difficulty case.
4. Provide theoretical foundation to independence learning.

The End

Happy Birthday!