

ISIS: A vehicle for the universe of sparsity

Jianqing Fan

Princeton University

<http://www.princeton.edu/~jqfan>

May 4, 2010

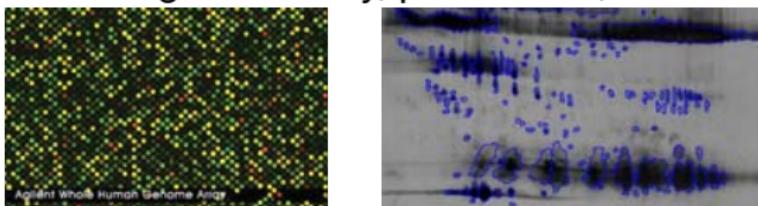
Outline

- ➊ Introduction
- ➋ Impact of Dimensionality
- ➌ The ISIS Method
- ➍ Numerical Studies
- ➎ Properties of Independence Screening

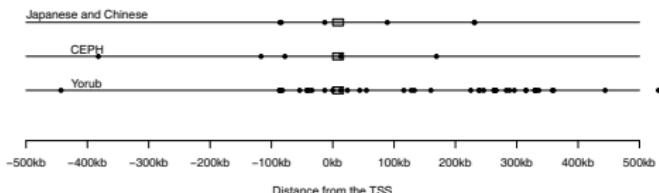
Examples: Biological Sciences and Engineering

High-dim variable selection characterizes many contemporary statistical problems.

- Bioinformatic: disease classification / predicting clinical outcomes using microarray, proteomics, fMRI data;



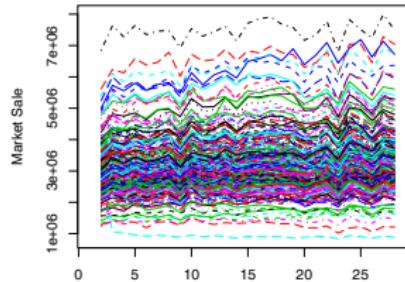
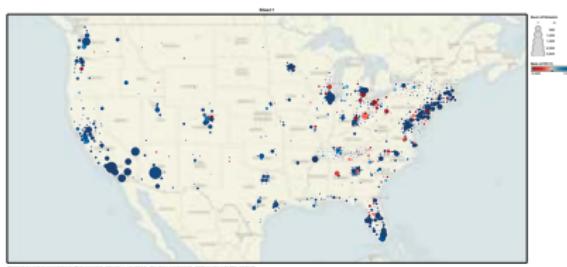
- Association between phenotypes and SNPs or eQTL



- Document or text classification: E-mail spam

Example: Economics, Finance, Marketing

- HPIs / drug sales collected in many regions
- Neighborhood depen. makes dimen. growths quickly
 - 1000 neighborhoods requires 1 m parameters.



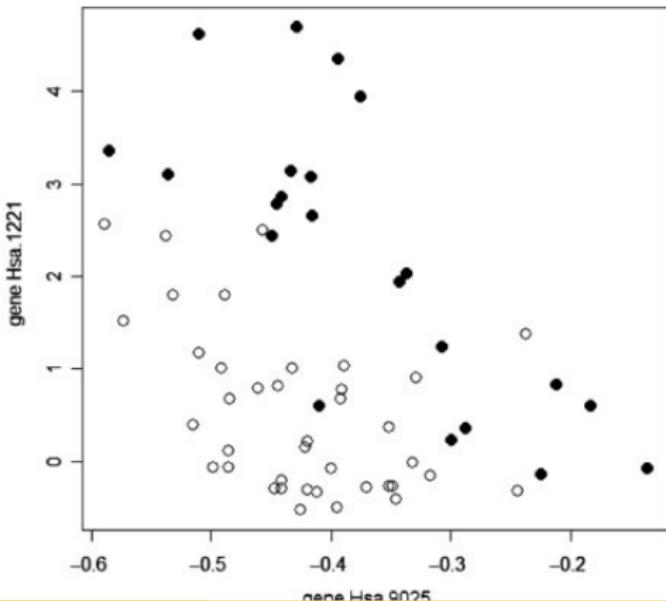
- Managing 2K stocks involves $2m$ elements in covariance.
- Spatial-temporal in Meteorology, Earth Sciences & Ecology

Growth of Dimensionality

Dimen. grows rapidly w/ interactions: 5000 → 12.5m.

Synergy of Two Genes: colon cancer in Hanczar et al (2007).

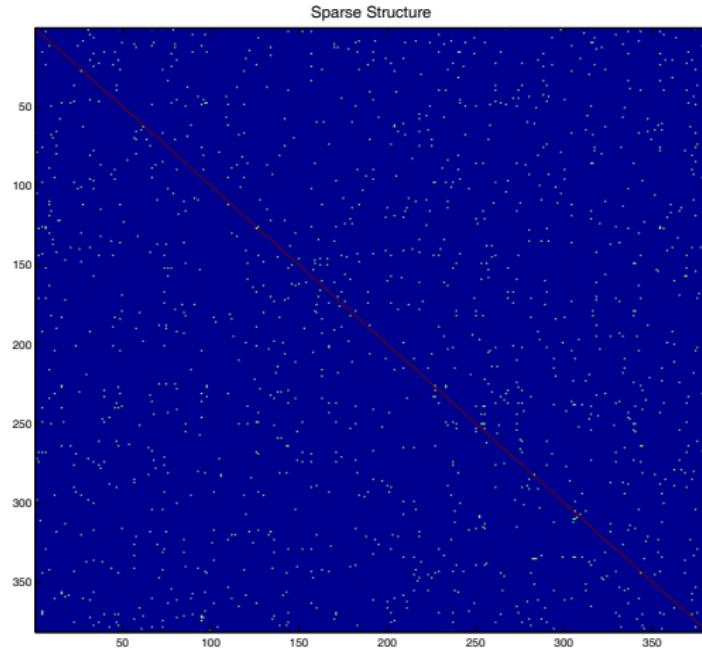
e.g., $Y = I(X_1 + X_2 > 3)$ and $Y \perp X_1$.



Sparsity

Dimensionality: $\log p = O(n^a)$ (**NP**-dimensionality)

Intrinsic dimensionality: $s \ll n$. (Sparsity)

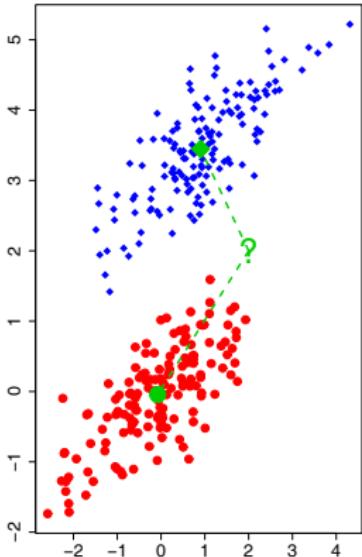


Impact of High-dimensionality

1. Noise accumulation

Regression:

- Not directly implementable if $p > n$.
- Prediction error is $(1 + \frac{p}{n})\sigma^2$, if $p \leq n$.



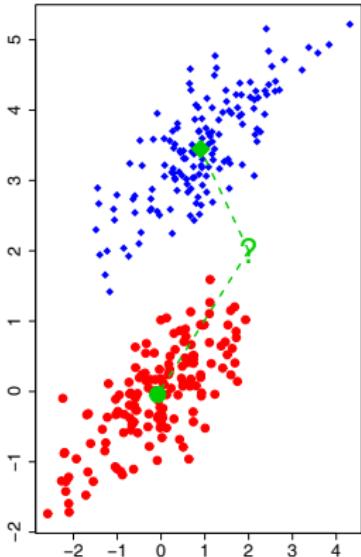
Classification: No implementation problems, but **error rates**

- depend on C_p^2 / \sqrt{p} (Fan & Fan 08), C_p is **distance**.
- **perfectly classifiable** if $C_p^2 / \sqrt{p} \rightarrow \infty$ (Hall, Pittelkow & Ghosh, 08).

1. Noise accumulation

Regression:

- Not directly implementable if $p > n$.
- Prediction error is $(1 + \frac{p}{n})\sigma^2$, if $p \leq n$.



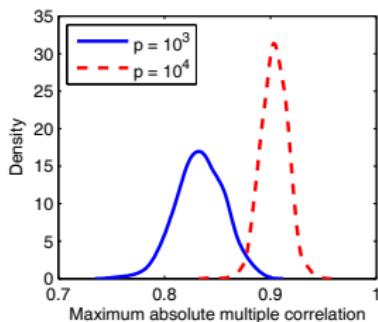
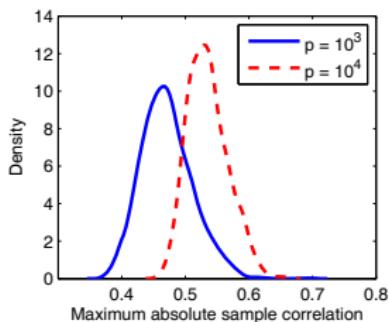
Classification: No implementation problems, but **error rates**

- depend on C_p^2 / \sqrt{p} (*Fan & Fan 08*), C_p is **distance**.
- **perfectly classifiable** if $C_p^2 / \sqrt{p} \rightarrow \infty$ (*Hall, Pittelkow & Ghosh, 08*).

2. Spurious Relation

An experiment: Generate $n = 50$ $Z_1, \dots, Z_p \sim i.i.d. N(0, 1)$;

■ compute $r = \max_{2 \leq j \leq p} |\text{corr}(Z_1, Z_j)|$.



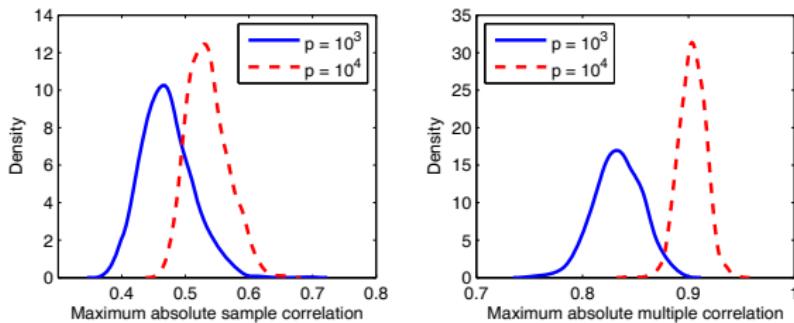
■ compute maximum multiple correlation:

$$R = \max_{|S|=5} |\text{corr}(Z_1, Z_S)|.$$

2. Spurious Relation

An experiment: Generate $n = 50$ $Z_1, \dots, Z_p \sim i.i.d. N(0, 1)$;

■ compute $r = \max_{2 \leq j \leq p} |\text{corr}(Z_1, Z_j)|$.



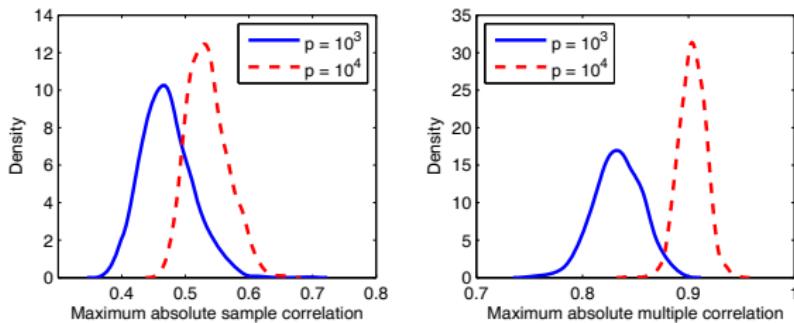
■ compute maximum multiple correlation:

$$R = \max_{|S|=5} |\text{corr}(Z_1, Z_S)|.$$

2. Spurious Relation

An experiment: Generate $n = 50$ $Z_1, \dots, Z_p \sim i.i.d. N(0, 1)$;

■ compute $r = \max_{2 \leq j \leq p} |\text{corr}(Z_1, Z_j)|$.



■ compute maximum multiple correlation:

$$R = \max_{|S|=5} |\text{corr}(Z_1, Z_S)|.$$

2a. False Outcomes

Scientific implication: If Z_1 is responsible for breast cancer, but we can also discover other 5 genes, **indep of outcome!**

False statistical inferences: If $Y = Z_1$ and fit

$$Y = \mathbf{Z}_{\hat{S}}^T \boldsymbol{\beta} + \varepsilon,$$

the residual variance

$$\hat{\sigma}^2 = \frac{RSS}{n - |\hat{S}|} \approx (1 - 0.9^2) \times \frac{49}{45} \approx 0.2,$$

★ more variables to be called “statistically significant”.

2a. False Outcomes

Scientific implication: If Z_1 is responsible for breast cancer, but we can also discover other 5 genes, **indep of outcome!**

False statistical inferences: If $Y = Z_1$ and fit

$$Y = \mathbf{Z}_{\hat{S}}^T \boldsymbol{\beta} + \varepsilon,$$

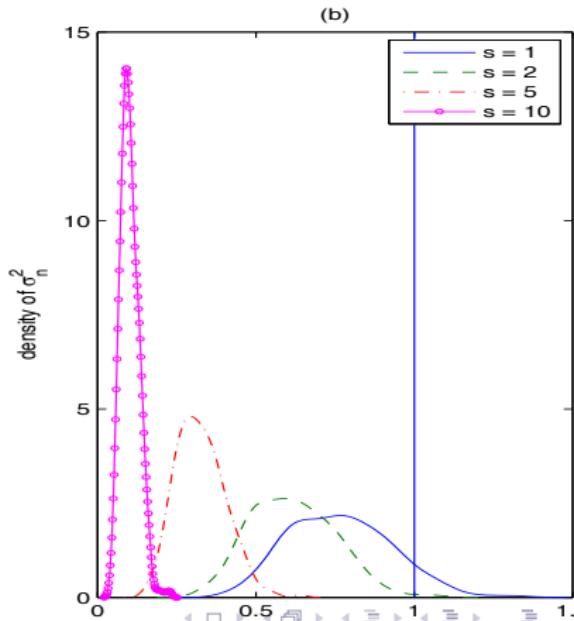
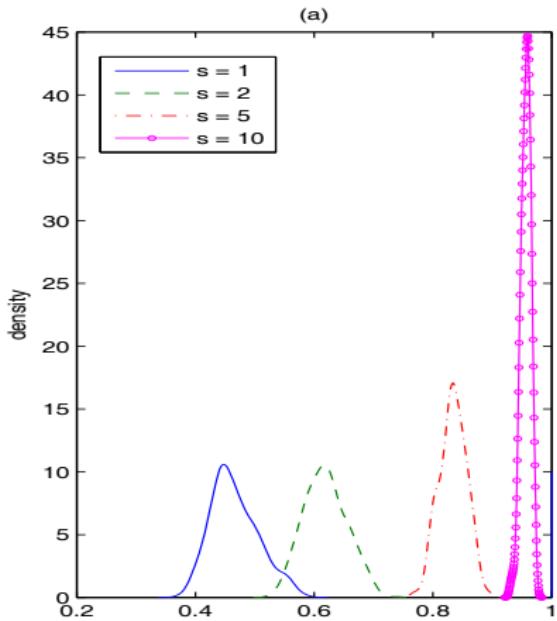
the residual variance

$$\hat{\sigma}^2 = \frac{RSS}{n - |\hat{S}|} \approx (1 - \mathbf{0.9}^2) \times \frac{49}{45} \approx 0.2,$$

★ more variables to be called “statistically significant”.

2b. False statistical inference

Spurious variables: recruited to predict realized noise,
seriously underestimate σ , e.g. $Y = 2X_1 + 0.3X_2 + \varepsilon$.



Curse of Ultrahigh Dimensionality

- Computational cost ■ Stability
- Estimation accuracy:
 - ★ noise accumulation
 - ★ spurious corr



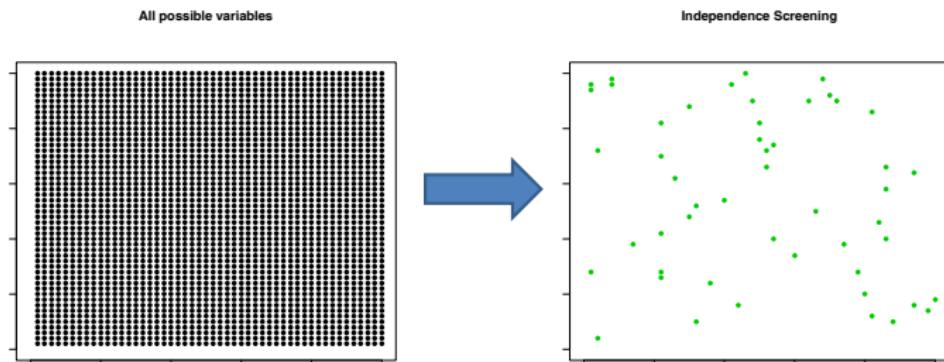
Key Idea: Large-scale screening + moderate-scale searching.

The ISIS Method

a two-scale framework

Hydrogen Atom: Large scale-screening

Indep learning: Feature ranking by **Marginal** correlation (*Fan & Lv, 08*) or generalized correlation (*Hall & Miller, 09*);



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

Extensions & Questions

Other methods: ★**Marginal LR** (*Fan, Samworth & Wu, 09*);
★**MMLE** (*Fan and Song, 09*); ★**MPLE** (*Zhao & Li, 09*);
★**Nonparametric learning** (*Fan, Feng, Song, 09*)
★**Data-tilting**; (*Hall, Titterington & Xue, 09*).

- ➊ Model selection consistency? (*Geneve, Jin, Wasserman, 10*)
- ➋ Sure screening property? In what capacity? (*Fan & Lv, 08*)
- ➌ How to choose a thresholding parameter? (*Zhao & Li, 10*)
- ➍ How to reduce FDR? (*Fan, Samworth, Wu, 09*)
- ➎ What are the possible drawbacks?

Extensions & Questions

Other methods: ★**Marginal LR** (*Fan, Samworth & Wu, 09*);
★**MMLE** (*Fan and Song, 09*); ★**MPLE** (*Zhao & Li, 09*);
★**Nonparametric learning** (*Fan, Feng, Song, 09*)
★**Data-tilting**; (*Hall, Titterington & Xue, 09*).

- ➊ Model selection consistency? (*Geneve, Jin, Wasserman, 10*)
- ➋ Sure screening property? In what capacity? (*Fan & Lv, 08*)
- ➌ How to choose a thresholding parameter? (*Zhao & Li, 10*)
- ➍ How to reduce FDR? (*Fan, Samworth, Wu, 09*)
- ➎ What are the possible drawbacks?

Extensions & Questions

Other methods: ★**Marginal LR** (*Fan, Samworth & Wu, 09*);
★**MMLE** (*Fan and Song, 09*); ★**MPLE** (*Zhao & Li, 09*);
★**Nonparametric learning** (*Fan, Feng, Song, 09*)
★**Data-tilting**; (*Hall, Titterington & Xue, 09*).

- ➊ Model selection consistency? (*Geneve, Jin, Wasserman, 10*)
- ➋ Sure screening property? In what capacity? (*Fan & Lv, 08*)
- ➌ How to choose a thresholding parameter? (*Zhao & Li, 10*)
- ➍ How to reduce FDR? (*Fan, Samworth, Wu, 09*)
- ➎ What are the possible drawbacks?

Potential Drawbacks

- ◆ **False Negative:** What if X_j marginally uncorrelated with Y , but jointly correlated with Y ?

$$Y = X_1 + X_2 + X_3 + \beta_4 X_4 + \varepsilon \quad \text{s.t.} \quad \text{cov}(Y, X_4) = 0.$$

- ◆ **False Positive:** What if X_2, \dots, X_{99} highly correlated with an important X_1 , but weakly correlated with Y conditionally?

$$Y = X_1 + 0.2 X_{100} + \varepsilon$$

Potential Drawbacks

- ◆ **False Negative:** What if X_j marginally uncorrelated with Y , but jointly correlated with Y ?

$$Y = X_1 + X_2 + X_3 + \beta_4 X_4 + \varepsilon \quad \text{s.t.} \quad \text{cov}(Y, X_4) = 0.$$

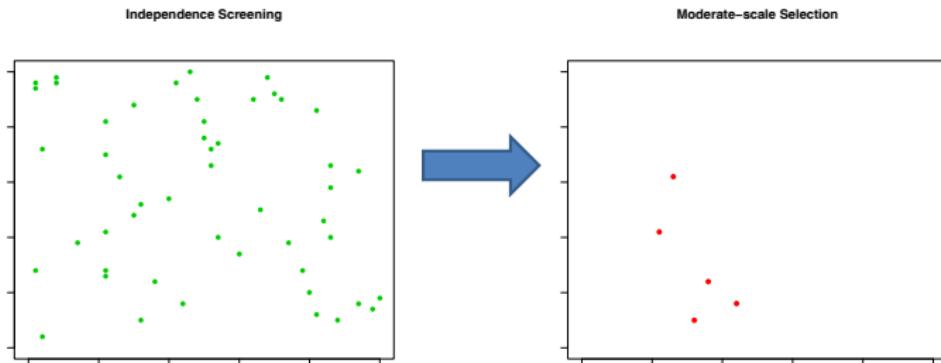
- ◆ **False Positive:** What if X_2, \dots, X_{99} highly correlated with an important X_1 , but weakly correlated with Y conditionally?

$$Y = X_1 + 0.2 X_{100} + \varepsilon$$

Oxygen Atom: Penalized likelihood estimation

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

■ Simultaneously estimate coeffs and choose variables.

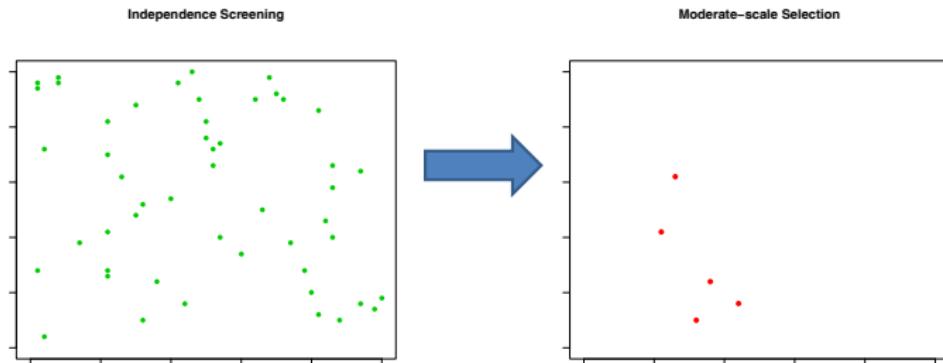


- How high dimensionality can such methods handle?
- What is the role of penalty functions?
- Does it possess an oracle property? How to compute?

Oxygen Atom: Penalized likelihood estimation

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

■ Simultaneously estimate coeffs and choose variables.



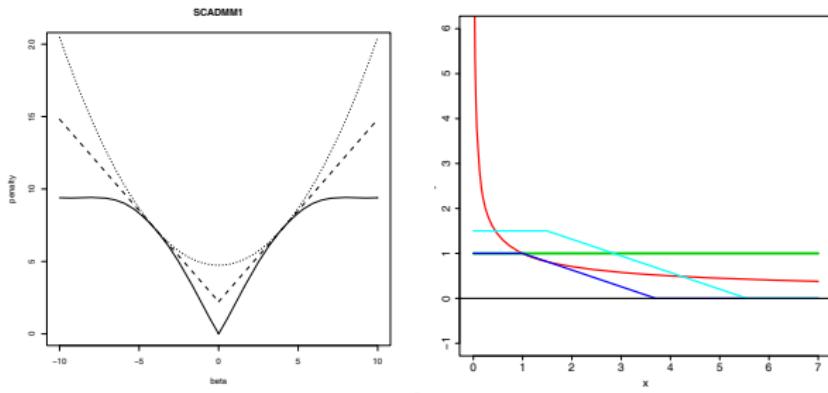
- How high dimensionality can such methods handle?
- What is the role of penalty functions?
- Does it possess an oracle property? How to compute?

Iterated reweighted L_1 -estimator

Penalty: Popular choice L_1 . Preferred choice: SCAD or MCP.

$$Q(\beta) \approx n^{-1} \sum_{i=1}^n L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^p \left\{ p_\lambda(|\beta_j^{(k)}|) + \mathbf{p}'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|) \right\}.$$

$$p_\lambda(|\beta_j|)$$



$$Q^{\text{app}}(\beta) = n^{-1} \sum_{i=1}^n L(Y_i, \mathbf{x}_{i,d}^T \beta) + \sum_{j=1}^d w_j |\beta_j|, \quad w_j = p'_\lambda(|\beta_j^{(k)}|)$$

■ $\beta^{(0)} = 0 \implies$ LASSO.

■ Iteration reduces the bias

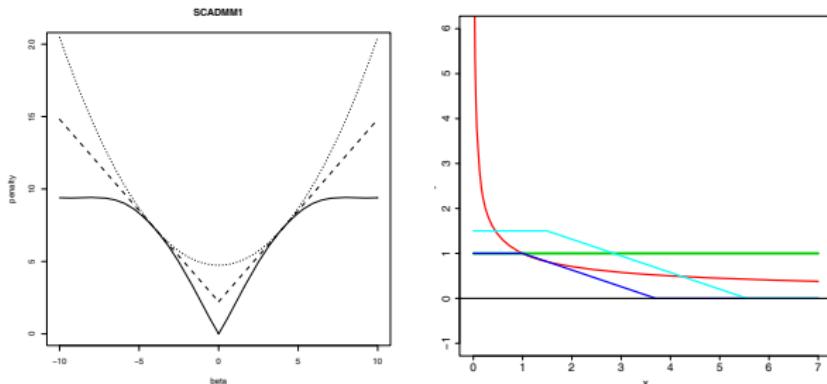
■ Zero is a non-absorbing state (comparing $w_j = 1/|\beta_j^{(k)}|^\gamma$)

Iterated reweighted L_1 -estimator

Penalty: Popular choice L_1 . Preferred choice: SCAD or MCP.

$$Q(\beta) \approx n^{-1} \sum_{i=1}^n L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^p \left\{ p_\lambda(|\beta_j^{(k)}|) + \mathbf{p}'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|) \right\}.$$

$$p_\lambda(|\beta_j|)$$



$$Q^{\text{app}}(\beta) = n^{-1} \sum_{i=1}^n L(Y_i, \mathbf{x}_{i,d}^T \beta) + \sum_{j=1}^d w_j |\beta_j|, \quad w_j = p'_\lambda(|\beta_j^{(k)}|)$$

■ $\beta^{(0)} = 0 \implies$ LASSO.

■ Iteration reduces the bias

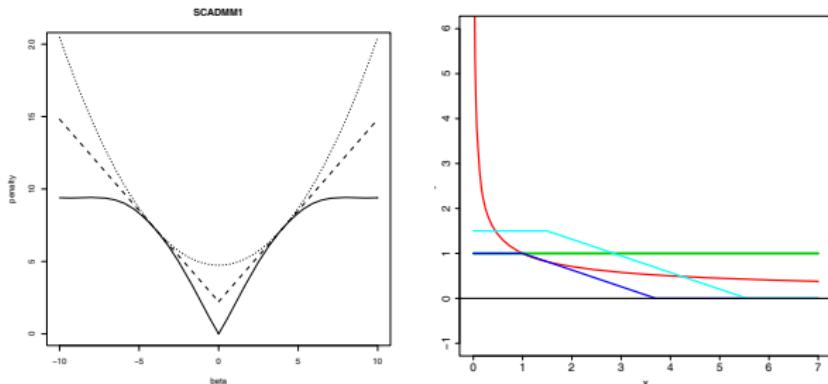
■ Zero is a non-absorbing state (comparing $w_j = 1/|\beta_j^{(k)}|^\gamma$)

Iterated reweighted L_1 -estimator

Penalty: Popular choice L_1 . Preferred choice: SCAD or MCP.

$$Q(\beta) \approx n^{-1} \sum_{i=1}^n L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^p \left\{ p_\lambda(|\beta_j^{(k)}|) + \mathbf{p}'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|) \right\}.$$

$$p_\lambda(|\beta_j|)$$



,

$$Q^{\text{app}}(\beta) = n^{-1} \sum_{i=1}^n L(Y_i, \mathbf{x}_{i,d}^T \beta) + \sum_{j=1}^d w_j |\beta_j|, \quad w_j = p'_\lambda(|\beta_j^{(k)}|)$$

■ $\beta^{(0)} = 0 \implies$ LASSO.

■ Iteration reduces the bias

■ Zero is a non-absorbing state (comparing $w_j = 1/|\beta_j^{(k)}|^\gamma$).

Remarks

Convergence: A Majorization-Minimization (MM) algorithm.

Other algorithms: **LQA** (*Fan & Li, 01*); **LLA** (*Zou & Li, 08*);

PLUS (*Zhang, 09*); **Coordinate optimization** (*Fu & Jiang, 99*).

Capacity: handle NP-dimensionality with wider capacity.

■ possesses an oracle property (*Fan & Lv, 09*),
reducing the bias of LASSO.

Remarks

Convergence: A Majorization-Minimization (MM) algorithm.

Other algorithms: **LQA** (*Fan & Li, 01*); **LLA** (*Zou & Li, 08*);

PLUS (*Zhang, 09*); **Coordinate optimization** (*Fu & Jiang, 99*).

Capacity: handle NP-dimensionality with wider capacity.

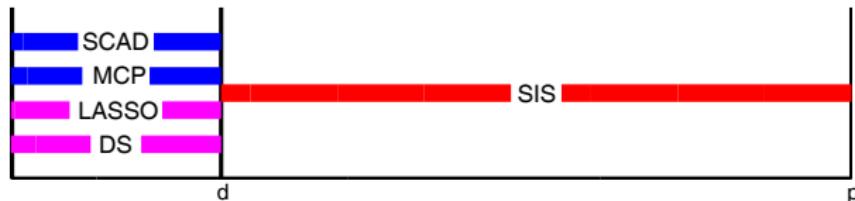
■ possesses an oracle property (*Fan & Lv, 09*),
reducing the bias of LASSO.

Carbon Atom

Iterative application of

large-scale **screening** and

moderate-scale **selection.**



■ ISIS ((Fan & Lv, 08; Fan, Samworth & Wu, 09)), **available in R.**

Iterative feature selection

- ➊ ■(screening): Apply SIS to pick a set \mathcal{A}_1 ;
■(selection): Employ a penalized likelihood to select a subset \mathcal{M}_1 of these indices.

- ➋ (conditional screening): Rank features according to the additional contribution:

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

resulting in \mathcal{A}_2 .

—Improvement of residual approach (FL 08): $\beta_{\mathcal{M}_1} = \hat{\beta}_{\mathcal{M}_1}$.

Iterative feature selection

- ➊ ■(screening): Apply SIS to pick a set \mathcal{A}_1 ;
■(selection): Employ a penalized likelihood to select a subset \mathcal{M}_1 of these indices.
- ➋ (conditional screening): Rank features according to the additional contribution:

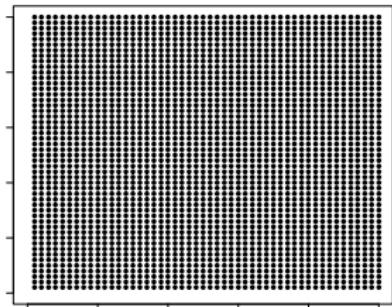
$$L_j^{(2)} = \min_{\beta_0, \beta_{\mathcal{M}_1}, \beta_j} n^{-1} \sum_{i=1}^n L(Y_i, \beta_0 + \mathbf{x}_{i,\mathcal{M}_1}^\top \boldsymbol{\beta}_{\mathcal{M}_1} + X_{ij}\beta_j),$$

resulting in \mathcal{A}_2 .

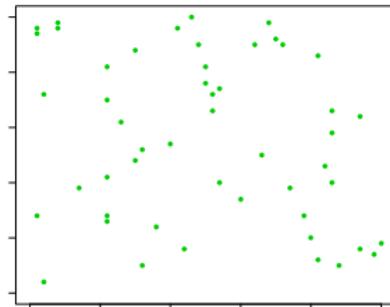
—Improvement of residual approach (FL 08): $\beta_{\mathcal{M}_1} = \hat{\beta}_{\mathcal{M}_1}$.

Illustration of ISIS

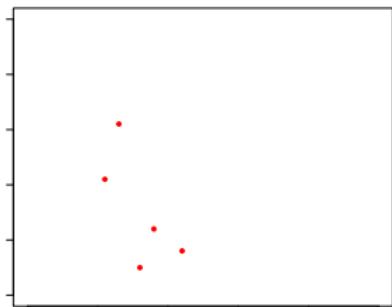
All possible variables



Independence Screening



Moderate-scale Selection



All candidates

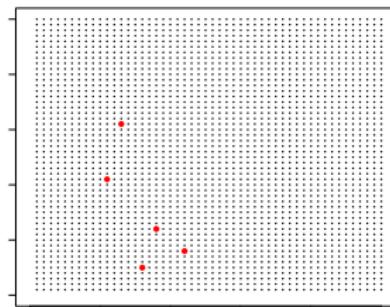
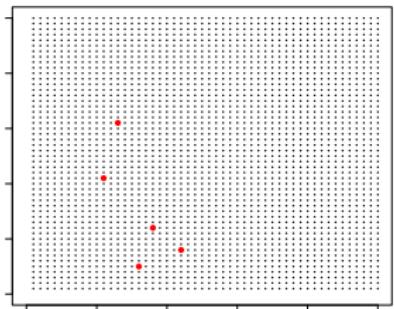
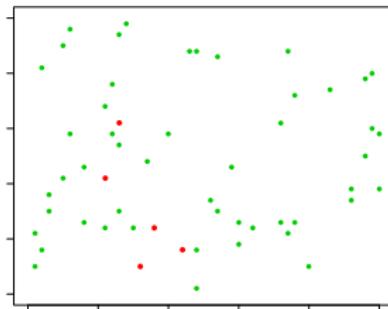


Illustration of ISIS

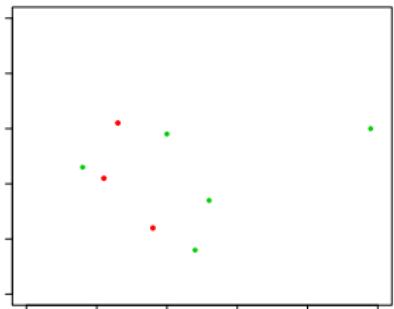
All candidates



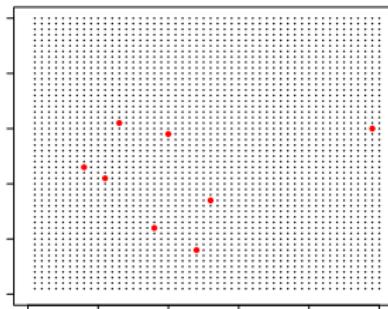
Conditional Screening



Moderate-scale selection



All candidates



Iterative feature selection (II)

- ③ (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.

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



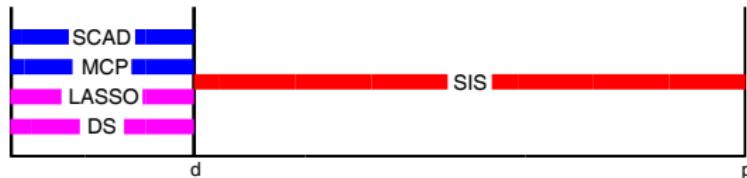
Iterative feature selection (II)

- ③ (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.

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



Applicability of ISIS idea

The idea of ISIS is widely applicable. It can be applied to

- Classification (*Fan, Samworth, & Wu, 09*).
- Survival analysis (*Fan, Feng, & Wu, 09; Zhao & Li, 09*).
- Nonparametric learning (*Fan, Feng, & Song, 09*).
- Robust and quantile regression (*Bradic, Fan, & Wang, 09*)

Applicability of ISIS idea

The idea of ISIS is widely applicable. It can be applied to

- Classification (*Fan, Samworth, & Wu, 09*).
- Survival analysis (*Fan, Feng, & Wu, 09; Zhao & Li, 09*).
- Nonparametric learning (*Fan, Feng, & Song, 09*).
- Robust and quantile regression (*Bradic, Fan, & Wang, 09*)

Applicability of ISIS idea

The idea of ISIS is widely applicable. It can be applied to

- Classification (*Fan, Samworth, & Wu, 09*).
- Survival analysis (*Fan, Feng, & Wu, 09; Zhao & Li, 09*).
- Nonparametric learning (*Fan, Feng, & Song, 09*).
- Robust and quantile regression (*Bradic, Fan, & Wang, 09*)

Applicability of ISIS idea

The idea of ISIS is widely applicable. It can be applied to

- Classification (*Fan, Samworth, & Wu, 09*).
- Survival analysis (*Fan, Feng, & Wu, 09; Zhao & Li, 09*).
- Nonparametric learning (*Fan, Feng, & Song, 09*).
- Robust and quantile regression (*Bradic, Fan, & Wang, 09*)

Applicability of ISIS idea

The idea of ISIS is widely applicable. It can be applied to

- Classification (*Fan, Samworth, & Wu, 09*).
- Survival analysis (*Fan, Feng, & Wu, 09; Zhao & Li, 09*).
- Nonparametric learning (*Fan, Feng, & Song, 09*).
- Robust and quantile regression (*Bradic, Fan, & Wang, 09*)

Numerical Studies

Logistic regression, a very difficult case

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

Bayes error: 0.1040.

$n = 400, p = 1000, N_{sim} = 100$

	Van-SIS	ISIS	Var2-ISIS	LASSO	NSC
med($\ \beta - \hat{\beta}\ _1$)	20.6	2.69	3.24	23.2	N/A
med($\ \beta - \hat{\beta}\ _2^2$)	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
$2Q(\hat{\beta}_0, \hat{\beta})$ (training)	269	188	188	109	N/A
AIC	289	198	199	311	N/A
BIC	337	218	219	714	N/A
$2Q(\hat{\beta}_0, \hat{\beta})$ (test)	361	225	226	276	N/A
0-1 test error	.193	.112	.112	.146	.387

Neuroblastoma Data (MAQC-II)

- ➊ 251 patients of the German Neuroblastoma Trials
NB90-NB2004, diagnosed between 1989 and 2004, aged from 0 to 296 months (median 15 months).
- ➋ 251 customized oligonucleotide microarray with $p = 10,707$.
- ➌ focus on “3-year Event Free Survival”, ($n = 239$ w/ 49 “+” and 190 “-”).
- ➍ **Aims:** To study which genes are responsible for neuroblastoma and its risk association.

Neuroblastoma Data (MAQC-II)

- ➊ 251 patients of the German Neuroblastoma Trials
NB90-NB2004, diagnosed between 1989 and 2004, aged from 0 to 296 months (median 15 months).
- ➋ 251 customized oligonucleotide microarray with $p = 10,707$.
- ➌ focus on “3-year Event Free Survival”, ($n = 239$ w/ 49 “+” and 190 “-”).
- ➍ **Aims:** To study which genes are responsible for neuroblastoma and its risk association.

Results

Training set and endpoints:

- ① “**3-y EFS**”: Sample $n = 125$ subjs (25 “+” and 100 “-”).
- ② “**Gender**”: Sample 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**”: Sample $n = 125$ subjs (25 “+” and 100 “-”).
- ② “**Gender**”: Sample 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

Sure Independence Screening

Model setting

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

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

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

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

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**: $L(Y_i, \mathbf{x}_i^T \beta) = |Y_i - \mathbf{x}_i^T \beta|.$

Independence learning

M-Utility: **Wilks**: $\hat{L}_j = \hat{L}_0 - \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^n L(Y_i, \beta_0 + X_{ij}\beta_j)$

Wald: $|\hat{\beta}_j^M|$, assuming $E X_j^2 = 1$.

F-Ranking: $\widehat{\mathcal{M}}_{v_n} = \{j : \hat{L}_j \geq v_n\}$, $\widehat{\mathcal{M}}_{\gamma_n}^{wald} = \{j : |\hat{\beta}_j^M| \geq \gamma_n\}$.

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

Theorem 1: $L_j^* = 0 \iff \text{cov}(Y, X_j) = 0 \iff \beta_j^M = 0$.

Independence learning

M-Utility: **Wilks**: $\hat{L}_j = \hat{L}_0 - \min_{\beta_0, \beta_j} n^{-1} \sum_{i=1}^n L(Y_i, \beta_0 + X_{ij}\beta_j)$

Wald: $|\hat{\beta}_j^M|$, assuming $E X_j^2 = 1$.

F-Ranking: $\widehat{\mathcal{M}}_{v_n} = \{j : \hat{L}_j \geq v_n\}$, $\widehat{\mathcal{M}}_{\gamma_n}^{wald} = \{j : |\hat{\beta}_j^M| \geq \gamma_n\}$.

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

Theorem 1: $L_j^* = 0 \iff \text{cov}(Y, X_j) = 0 \iff \beta_j^M = 0$.

Theoretical Basis

True model: $\mathcal{M}_* = \{j : \beta_j^* \neq 0\}$.

Theorem 2: If $|\text{cov}(Y, X_j)| \geq c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$, then

$$\min_{j \in \mathcal{M}_*} |\beta_j^M| \geq c_1 n^{-\kappa}, \quad \min_{j \in \mathcal{M}_*} |L_j^*| \geq c_2 n^{-2\kappa}.$$

■ If **active** indep of **inactive**, then $L_j^* = 0, j \notin \mathcal{M}_*$
⇒ model selection consistency.

Theoretical Basis

True model: $\mathcal{M}_* = \{j : \beta_j^* \neq 0\}$.

Theorem 2: If $|\text{cov}(Y, X_j)| \geq c_1 n^{-\kappa}$ for $j \in \mathcal{M}_*$, then

$$\min_{j \in \mathcal{M}_*} |\beta_j^M| \geq c_1 n^{-\kappa}, \quad \min_{j \in \mathcal{M}_*} |L_j^*| \geq c_2 n^{-2\kappa}.$$

■ If **active** indep of **inactive**, then $L_j^* = 0, j \notin \mathcal{M}_*$
⇒ model selection consistency.

Sampling Aspect: Sure independence screening

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

$$P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{v_n}\right) \rightarrow 1 \quad \text{exponentially fast}$$

No conditions on covariance matrix!

- Note that $\hat{L}_j - L_j^* = O(\log p / n^{1/2})$ and minimum signal $O(n^{-2\kappa})$. How to deal with it?
 - ★Appeal to rank invariance under monotonic transform.
- Screening using Wald stat $\hat{\beta}_j^M$ has also SS property.

Sampling Aspect: Sure independence screening

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

$$P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{v_n}\right) \rightarrow 1 \quad \text{exponentially fast}$$

No conditions on covariance matrix!

- Note that $\hat{L}_j - L_j^* = O(\log p / n^{1/2})$ and minimum signal $O(n^{-2\kappa})$. **How to deal with it?**
 - ★ Appeal to rank invariance under monotonic transform.
- Screening using **Wald stat** $\hat{\beta}_j^M$ has also SS property.

Sampling Aspect: Sure independence screening

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

$$P\left(\mathcal{M}_* \subset \widehat{\mathcal{M}}_{v_n}\right) \rightarrow 1 \quad \text{exponentially fast}$$

No conditions on covariance matrix!

- Note that $\hat{L}_j - L_j^* = O(\log p / n^{1/2})$ and minimum signal $O(n^{-2\kappa})$. **How to deal with it?**
 - ★ Appeal to rank invariance under monotonic transform.
- Screening using **Wald stat** $\hat{\beta}_j^M$ has also SS property.

Sampling Aspect: Controlling number of features

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

$$\mathbf{P}[\widehat{\mathcal{M}}_{v_n} \leq \mathbf{O}\{n^{2\kappa}\lambda_{\max}(\Sigma)\}] \rightarrow 1.$$

When $\lambda_{\max}(\Sigma) = O(n^\tau)$, model size = $O(n^{2\kappa+\tau})$ (Fan and Lv, 08).

■ More precise bound for $|\widehat{\mathcal{M}}_{v_n}|$ is

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

Screening by MMLE

Result holds for MMLE screening.

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

② $P(\min_{j \in \mathcal{M}_*} |\hat{\beta}_j^M| \geq \gamma_n) \rightarrow 1$, if $\gamma_n = c_0 n^{-\kappa}$, $c_0 < c_1/2$.

③ What is the selected model size? We establish

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

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

Result holds for MMLE screening.

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

② $P(\min_{j \in \mathcal{M}_*} |\hat{\beta}_j^M| \geq \gamma_n) \rightarrow 1$, if $\gamma_n = c_0 n^{-\kappa}$, $c_0 < c_1/2$.

③ What is the selected model size? We establish

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

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

Result holds for MMLE screening.

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

② $P(\min_{j \in \mathcal{M}_*} |\hat{\beta}_j^M| \geq \gamma_n) \rightarrow 1$, if $\gamma_n = c_0 n^{-\kappa}$, $c_0 < c_1/2$.

③ What is the selected model size? We establish

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

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

Performance of Independence Screening

- compare **minimum model size** for sure screening w/ LASSO.
- The capacity is about the same for SIS.
Fan and Song (09), Zhang (09), Jin, Geneve, Wasserman (09).
- Consistent condition for LASSO is stringent (Zhao and Yu, 06):
 $\|(\mathbf{x}_1^T \mathbf{x}_1)^{-1} \mathbf{x}_1^T \mathbf{x}_{2,j}\|_1 < 1$.

Design 1: $\{X_j = \frac{\varepsilon_j + a_j \varepsilon}{\sqrt{1+a_j^2}}\}_{j=1}^q$, rest indep.

Performance of Independence Screening

- compare **minimum model size** for sure screening w/ LASSO.
- The capacity is about the same for SIS.
Fan and Song (09), Zhang (09), Jin, Geneve, Wasserman (09).
- Consistent condition for LASSO is stringent (Zhao and Yu, 06):
 $\|(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_{2,j}\|_1 < 1$.

Design 1: $\{X_j = \frac{\varepsilon_j + a_j \varepsilon}{\sqrt{1+a_j^2}}\}_{j=1}^q$, rest indep.

Linear regression, $p = 40,000$, $q = 15$

ρ	n	SIS-MLR	SIS-MMLE	n	SIS-MLR	SIS-MMLE
		$s = 3, \beta^* = (1, 1.3, 1)^T$				$s = 6, \beta^* = (1, 1, 3, 1, \dots)^T$
0	80	12(18)	12(18)	150	42(157)	42(157)
0.2	80	3(0)	3(0)	150	6(0)	6(0)
0.4	80	3(0)	3(0)	150	6.5(1)	6.5(1)
0.6	80	3(0)	3(0)	150	6(1)	6(1)
0.8	80	3(0)	3(0)	150	7(1)	7(1)
		$s = 12, \beta^* = (1, 1.3, \dots)^T$				$s = 15, \beta^* = (1, 1.3, \dots)^T$
0	300	143(282)	143(282)	400	135.5(167)	135.5(167)
0.2	200	13(1)	13(1)	200	15(0)	15(0)
0.4	200	13(1)	13(1)	200	15(0)	15(0)
0.6	200	13(1)	13(1)	200	15(0)	15(0)
0.8	200	13(1)	13(1)	200	15(0)	15(0)

Logistic regression, $p = 2,000$, $q = 50$

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
$s = 6, \beta^* = (3, -3, 3, -3, 3, -3)^T$					
0.4	200	51(77)	64.5(76)	20(10)	16.5(6)
0.6	300	77.5(139)	77.5(132)	20(13)	19(9)
0.8	400	306.5(347)	313(336)	86(40)	70.5(35)
$s = 12, \beta^* = (3, 4, \dots)^T$					
0.4	600	32(10)	30(10)	18(3)	17(4)
0.6	600	38(9)	38(10)	22(3)	22(4)
0.8	600	38(7)	39(8)	1071(6)	1042(34)
$s = 24, \beta^* = (3, 4, \dots)^T$					
0.4	600	46(3)	47(2)	1099(17)	1093(1456)
0.6	600	48(2)	48(2)	1078(5)	1065(23)
0.8	600	48(1)	48(1)	1072(4)	1067(13)

Logistic regression, $p = 5,000$, $q = 15$

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
$s = 6, \beta^* = (1, 1.3, 1, 1.3, 1, 1.3)^T$					
0.4	200	51(77)	64.5(76)	20(10)	16.5(6)
0.6	300	77.5(139)	77.5(132)	20(13)	19(9)
0.8	400	306.5(347)	313(336)	86(40)	70.5(35)
$s = 12, \beta^* = (1, 1.3, \dots)^T$					
0.4	300	14(1)	14(1)	14(1861)	13(1865)
0.6	300	14(1)	14(1)	2552(85)	12(3721)
0.8	300	14(1)	14(1)	2556(10)	12(3722)
$s = 15, \beta^* = (3, 4, \dots)^T$					
0.4	300	15(0)	15(0)	38(3719)	15(3720)
0.6	300	15(0)	15(0)	2555(87)	15(1472)
0.8	300	15(0)	15(0)	2552(8)	15(1322)

Logistic regression, $p = 2000$, $n = 600$

Design 2: $\{X_k\}_{k=1}^{p-50} \sim_{i.i.d.} N(0, 1)$.

$$X_k = \sum_{j=1}^s X_j (-1)^{j+1}/5 + \sqrt{25-s}/5 \varepsilon_k, \quad k \geq p-49$$

Regression Coefs: $\beta^* = (1, -1, 1, -1, \dots)^T$

s	M- λ_{\max} (RSD)	SIS-MLR	SIS-MMLE	LASSO	SCAD
3	8.47(0.17)	3(0)	3(0)	3(1)	3(0)
6	10.36(0.26)	56(0)	56(0)	1227(7)	1142(64)
12	14.69(0.39)	63(6)	63(6)	1148(8)	1093(59)
24	23.70(0.14)	214.5(93)	208.5(82)	1120(5)	1087(24)

Linear regression, $p = 2000$, $n = 600$

True coef: $\beta^* = (1, -1, \dots)^T$

S	M- λ_{\max} (RSD)	SIS-MLR	SIS-MMLE	LASSO	SCAD
3	8.47(0.17)	3(0)	3(0)	3(0)	3(0)
6	10.36(0.26)	56(0)	56(0)	47(4)	45(3)
12	14.69(0.39)	62(0)	62(0)	1610(10)	1304(2)
24	23.70(0.14)	81(19)	81(23)	1637(14)	1303(1)

Conclusion

- Impact of dimensionality: Noise accumulation, spurious correlation, computation.
- Spurious relations arises **easily** in NP-dimensionality and have adverse effect on statistical inference.
- ISIS is effective in high-dimensional regression and classification.
- Fold-concave penalized MLE can handle NP-dimensionality.
- It reduces significantly the biases of L_1 -penalty and requires much less condition for selection consistency.

Conclusion

- Impact of dimensionality: Noise accumulation, spurious correlation, computation.
- Spurious relations arises **easily** in NP-dimensionality and have adverse effect on statistical inference.
- ISIS is effective in high-dimensional regression and classification.
- Fold-concave penalized MLE can handle NP-dimensionality.
- It reduces significantly the biases of L_1 -penalty and requires much less condition for selection consistency.

Conclusion

- Impact of dimensionality: Noise accumulation, spurious correlation, computation.
- Spurious relations arises **easily** in NP-dimensionality and have adverse effect on statistical inference.
- ISIS is effective in high-dimensional regression and classification.
- Fold-concave penalized MLE can handle NP-dimensionality.
- It reduces significantly the biases of L_1 -penalty and requires much less condition for selection consistency.

Conclusion

- Impact of dimensionality: Noise accumulation, spurious correlation, computation.
- Spurious relations arises **easily** in NP-dimensionality and have adverse effect on statistical inference.
- ISIS is effective in high-dimensional regression and classification.
- Fold-concave penalized MLE can handle NP-dimensionality.
- It reduces significantly the biases of L_1 -penalty and requires much less condition for selection consistency.

Conclusion

- Impact of dimensionality: Noise accumulation, spurious correlation, computation.
- Spurious relations arises **easily** in NP-dimensionality and have adverse effect on statistical inference.
- ISIS is effective in high-dimensional regression and classification.
- Fold-concave penalized MLE can handle NP-dimensionality.
- It reduces significantly the biases of L_1 -penalty and requires much less condition for selection consistency.

Acknowledgement

Thank



You

In collaboration with

- ★ Jinchi Lv (*University of Southern California; Fan & Lv; 2008*)
- ★ Richard Samworth (*Cambridge University; FSW, 2009*).
- ★ Rui Song (*Colorado State University, Fan & Song, 2009*).
- ★ Yichao Wu (*North Carolina State University, FSW, 2009*).