High-dimensional Statistical Learning and Inference

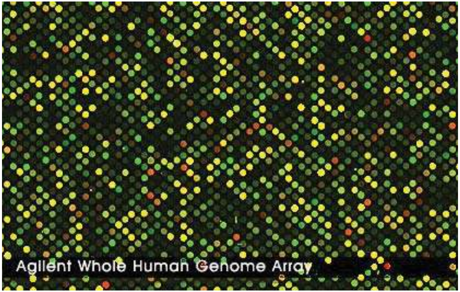
Jianqing Fan

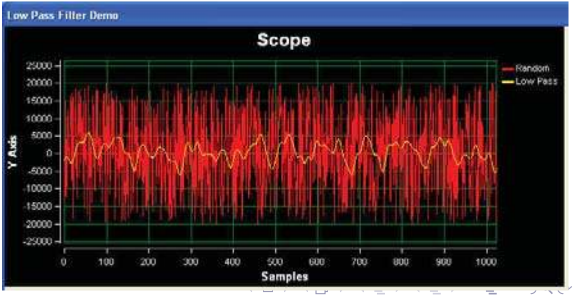
Princeton University

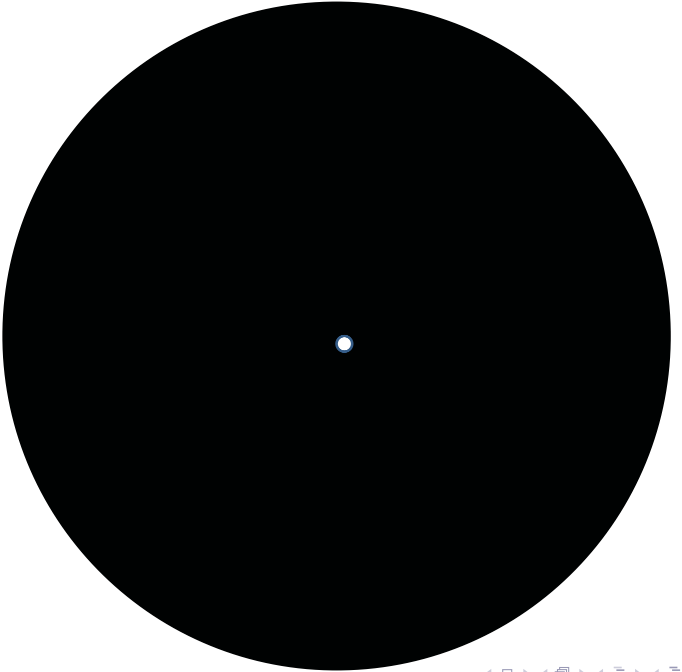
http://www.princeton.edu/∼jqfan

March 25, 2011

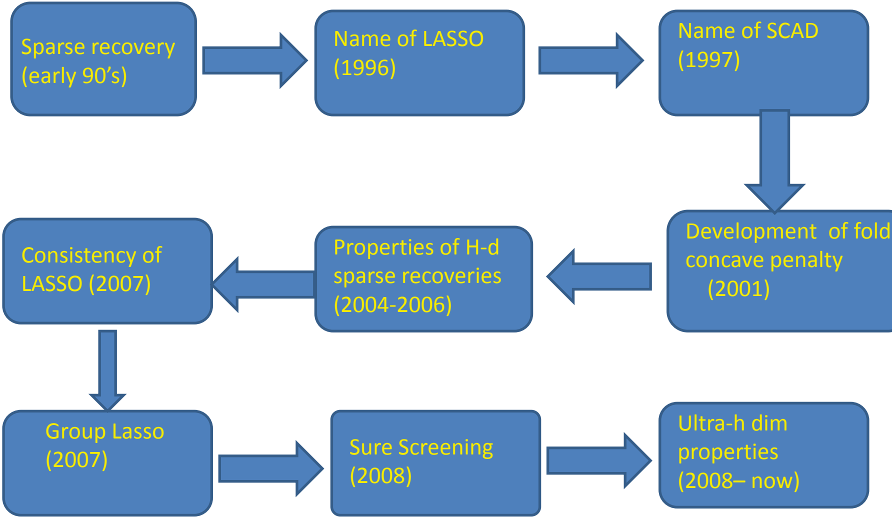
Evolution of Dimensions



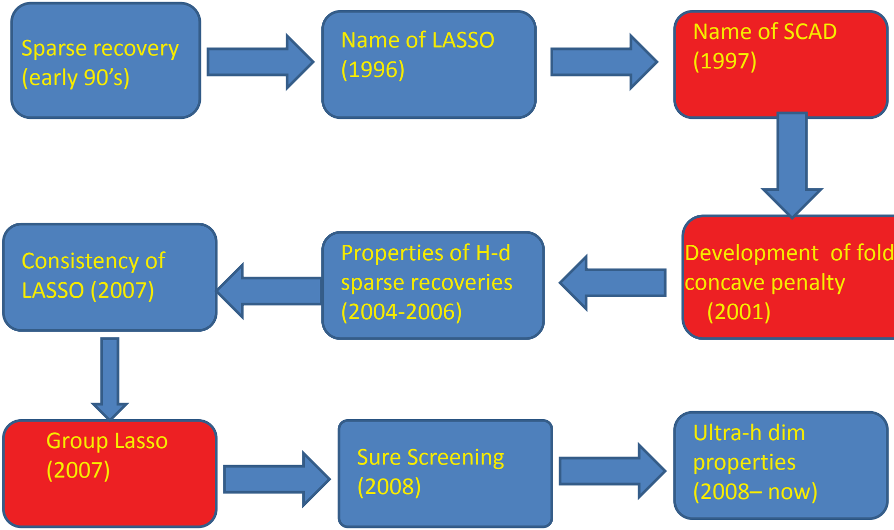




Development of High-dimensional Statistical learning

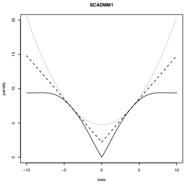


Contribution of Antoniadis



Specific Contributions of Antoniadis

- **Introduce the hard-thresholding penalty, now generalized to MCP.**



Regularization of Wavelet Approximations

Anestis ANTONIADIS and Jianqing FAN

In this paper, we introduce nonlinear regularized wavelet estimators for estimating nonparametric regression functions when sampling points are not uniformly spaced. The approach can apply readily to many other statistical contexts. Various new penalty functions are proposed. The hard-thresholding and soft-thresholding estimators of Donoho and Johnstone are specific members of nonlinear regularized wavelet estimators. They correspond to the lower and upper envelopes of a class of the penalized least squares estimators. Necessary conditions for penalty functions are given for regularized estimators to possess thresholding properties. Oracle inequalities and universal thresholding parameters are obtained for a large class of penalty functions. The sampling properties of nonlinear regularized wavelet estimators are established and are shown to be adaptively minimax. To efficiently solve penalized least squares problems, nonlinear regularized Sobolev interpolators (NRSI) are proposed as initial estimators, which are shown to have good sampling properties. The NRSI is further ameliorated by regularized one-step estimators, which are the one-step estimators of the penalized least squares problems using the NRSI as initial estimators. The graduated nonconvexity algorithm is also introduced to handle penalized least squares problems. The newly introduced approaches are illustrated by a few numerical examples.

KEY WORDS: Asymptotic minimax; Irregular designs; Nonquadratic penalty functions; Oracle inequalities; Penalized least-squares; ROSE; Wavelets.

- Introduce folded concave penalties

sion of our work is to minimize, in the wavelet coefficients domain, the following penalized least squares:

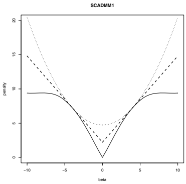
$$\sum_{(b)} \|z_{(b)} - \theta_{(b)}\|^2 + \sum_{(b)} p_\lambda(\|\theta_{(b)}\|), \qquad (2)$$

where $p_\lambda(\cdot)$ is a penalty function given in Theorem 1. Similar to equation (3) of Professor Moulin's contribution, the flexibility can be further enhanced by introducing a weight $\lambda_{(b)}$ in the penalty part of (2) or more generally by using

- Introduce group penalty and/or group LASSO

Specific Contributions of Antoniadis

- **Introduce the hard-thresholding penalty, now generalized to MCP.**



Regularization of Wavelet Approximations

Anestis ANTONIADIS and Jianqing FAN

In this paper, we introduce nonlinear regularized wavelet estimators for estimating nonparametric regression functions when sampling points are not uniformly spaced. The approach can apply readily to many other statistical contexts. Various new penalty functions are proposed. The hard-thresholding and soft-thresholding estimators of Donoho and Johnstone are specific members of nonlinear regularized wavelet estimators. They correspond to the lower and upper envelopes of a class of the penalized least squares estimators. Necessary conditions for penalty functions are given for regularized estimators to possess thresholding properties. Oracle inequalities and universal thresholding parameters are obtained for a large class of penalty functions. The sampling properties of nonlinear regularized wavelet estimators are established and are shown to be adaptively minimax. To efficiently solve penalized least squares problems, nonlinear regularized Sobolev interpolators (NRSI) are proposed as initial estimators, which are shown to have good sampling properties. The NRSI is further ameliorated by regularized one-step estimators, which are the one-step estimators of the penalized least squares problems using the NRSI as initial estimators. The graduated nonconvexity algorithm is also introduced to handle penalized least squares problems. The newly introduced approaches are illustrated by a few numerical examples.

KEY WORDS: Asymptotic minimax; Irregular designs; Nonquadratic penality functions; Oracle inequalities; Penalized least-squares; ROSE; Wavelets.

- **Introduce folded concave penalties**

sion of our work is to minimize, in the wavelet coefficients domain, the following penalized least squares:

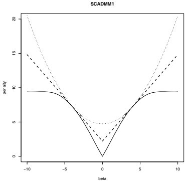
$$\sum_{(b)} \|z_{(b)} - \theta_{(b)}\|^2 + \sum_{(b)} p_\lambda(\|\theta_{(b)}\|), \tag{2}$$

where $p_\lambda(\cdot)$ is a penalty function given in Theorem 1. Similar to equation (3) of Professor Moulin's contribution, the flexibility can be further enhanced by introducing a weight $\lambda_{(b)}$ in the penalty part of (2) or more generally by using

- **Introduce group penalty and/or group LASSO**

Specific Contributions of Antoniadis

- **Introduce the hard-thresholding penalty, now generalized to MCP.**



Regularization of Wavelet Approximations

Anestis ANTONIADIS and Jianqing FAN

In this paper, we introduce nonlinear regularized wavelet estimators for estimating nonparametric regression functions when sampling points are not uniformly spaced. The approach can apply readily to many other statistical contexts. Various new penalty functions are proposed. The hard-thresholding and soft-thresholding estimators of Donoho and Johnstone are specific members of nonlinear regularized wavelet estimators. They correspond to the lower and upper envelopes of a class of the penalized least squares estimators. Necessary conditions for penalty functions are given for regularized estimators to possess thresholding properties. Oracle inequalities and universal thresholding parameters are obtained for a large class of penalty functions. The sampling properties of nonlinear regularized wavelet estimators are established and are shown to be adaptively minimax. To efficiently solve penalized least squares problems, nonlinear regularized Sobolev interpolators (NRSI) are proposed as initial estimators, which are shown to have good sampling properties. The NRSI is further ameliorated by regularized one-step estimators, which are the one-step estimators of the penalized least squares problems using the NRSI as initial estimators. The graduated nonconvexity algorithm is also introduced to handle penalized least squares problems. The newly introduced approaches are illustrated by a few numerical examples.

KEY WORDS: Asymptotic minimax; Irregular designs; Nonquadratic penality functions; Oracle inequalities; Penalized least-squares; ROSE; Wavelets.

- **Introduce folded concave penalties**

sion of our work is to minimize, in the wavelet coefficients domain, the following penalized least squares:

$$\sum_{(b)} \|z_{(b)} - \theta_{(b)}\|^2 + \sum_{(b)} p_\lambda(\|\theta_{(b)}\|), \qquad (2)$$

where $p_\lambda(\cdot)$ is a penalty function given in Theorem 1. Similar to equation (3) of Professor Moulin's contribution, the flexibility can be further enhanced by introducing a weight $\lambda_{(b)}$ in the penalty part of (2) or more generally by using

- **Introduce group penalty and/or group LASSO**

Outline

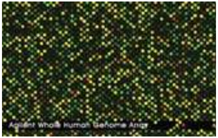
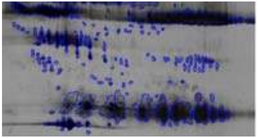
1. Introduction
2. Impact of Dimensionality
3. A two-scale approach
4. Numerical Studies
5. Sure independence screening
6. Properties of penalized likelihood

Introduction

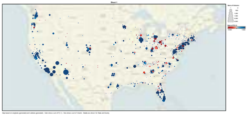
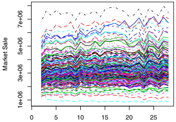
Rise of high-dimensionality

High-dim characterizes many statistical problems:

- **Biological science**: disease classification / predicting clinical outcomes using high-throughput data; association studies;

- **Engineering**: Doc or text classification, computer vision.
- **Economics, Finance, Marketing**: sale data collected in many regions.

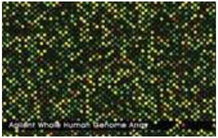



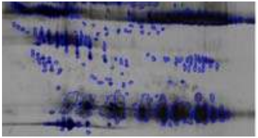
- **Spatial-temporal**: Meteorology; Earth Sciences; Ecology

Rise of high-dimensionality

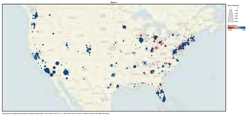
High-dim characterizes many statistical problems:

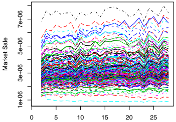
- **Biological science**: disease classification / predicting clinical outcomes using high-throughput data; association studies;





- **Engineering**: Doc or text classification, computer vision.
- **Economics, Finance, Marketing**: sale data collected in many regions.





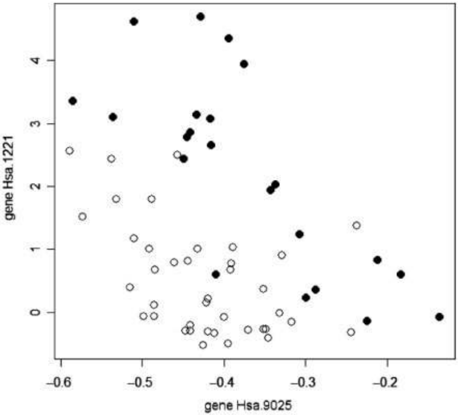
- **Spatial-temporal**: Meteorology; Earth Sciences; Ecology

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** $Y \perp X_1$.



Aims of High-dimensional Regression and Classification

Bickel (2008) discussion of the SIS paper published in JRSS-B (*Fan & Lv, 08*).

- 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.

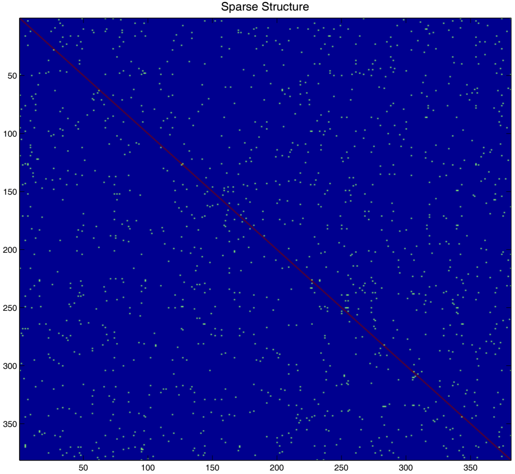
Aims of High-dimensional Regression and Classification

Bickel (2008) discussion of the SIS paper published in JRSS-B (*Fan & Lv, 08*).

- 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.

Popular Assumption: Sparsity

Dimen: $\log p = O(n^a)$ **Intrinsic dim**: $s \ll n$. (Sparsity)



■much easier to get **sure screening** than selection consistency.

Impact of Dimensionality

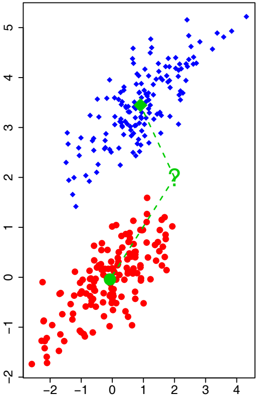
Impact of Dimensionality

■ Computational cost ■ Stability

■ Estimation accuracy: ★noise accumulation ★spurious corr

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

1. Noise accumulation



Regression:

■**Not** directly implementable if $p > n$.

■Prediction error is $(1 + \frac{p}{n})\sigma^2$, if $p \le n$.

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

—depend on $C_p^2/\sqrt{p}$ (*Fan & Fan 08*), C_p is **distance**.

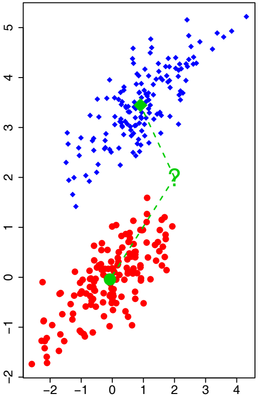
—**perfectly classifiable** if $C_p^2/\sqrt{p} \longrightarrow \infty$ (*Hall, Pittelkow & Ghosh,08*).

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**

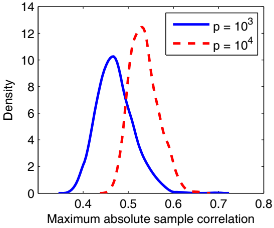
—depend on $C_p^2/\sqrt{p}$ (*Fan & Fan 08*), C_p is **distance**.

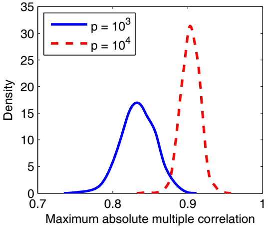
—**perfectly classifiable** if $C_p^2/\sqrt{p} \to \infty$ (*Hall, Pittelkow & Ghosh,08*).

2. Spurious Correlations

An experiment: Generate $n = 50$ $Z_1, \cdots, Z_p \sim_{i.i.d.} N(0, 1)$;

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



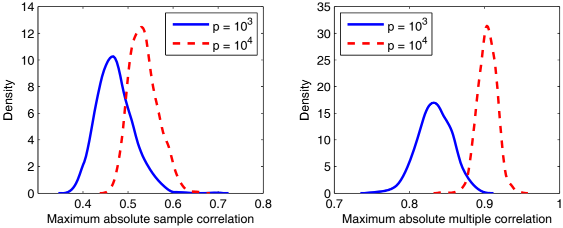


■compute maximum multiple correlation:

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

An experiment: Generate $n = 50$ $Z_1, \cdots, Z_p \sim_{i.i.d.} N(0,1)$;

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



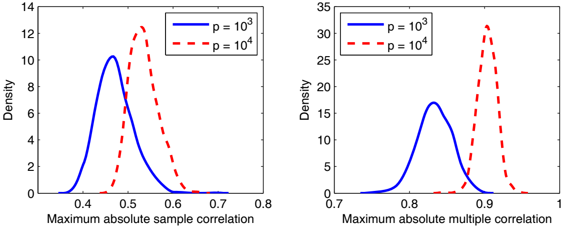
■compute maximum multiple correlation:

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

2. Spurious Correlations

An experiment: Generate $n = 50$ $Z_1, \cdots, Z_p \sim_{i.i.d.} N(0,1)$;

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



■ compute maximum multiple correlation:

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

False Statistical Inference

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

$$Y = \mathbf{X}_{\hat{M}}^T \beta + \varepsilon,$$

the residual variance

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_{\hat{M}})\mathbf{y}}{n - \hat{s}} = (1 - \hat{\gamma}_n^2)\frac{\|\varepsilon\|^2}{n - \hat{s}}.$$

Fraction of bias: $\gamma_n^2 = \varepsilon^T \mathbf{P}_{\hat{M}} \varepsilon / \|\varepsilon\|^2 = O_P(\hat{s} \log p / n)$.

Naive two-stage: Use the **selected** model and refit the data.

Seriously underestimate the variance.

False Statistical Inference

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

$$Y = \mathbf{X}_{\hat{M}}^T \beta + \varepsilon,$$

the residual variance

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T(\mathbf{I}_n - \mathbf{P}_{\hat{M}})\mathbf{y}}{n - \hat{s}} = (1 - \hat{\gamma}_n^2)\frac{\|\varepsilon\|^2}{n - \hat{s}}.$$

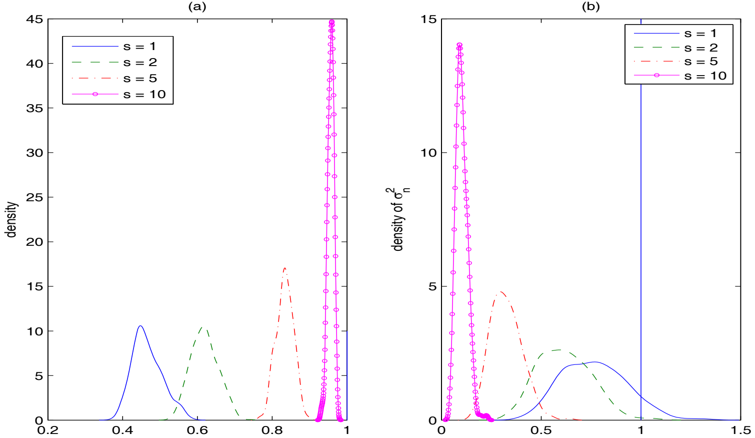
Fraction of bias: $\gamma_n^2 = \varepsilon^T \mathbf{P}_{\hat{M}} \varepsilon / \|\varepsilon\|^2 = O_P(\hat{s} \log p / n)$.

Naive two-stage: Use the **selected** model and refit the data.

Seriously underestimate the variance.

Impact of spurious correlation on variance est

- $p = 1000, n = 50$ with various spurious variables $\hat{s}$.



- **Spurious variables** are selected to predict noises:

$Y = 2X_1 + 0.3X_2 + \varepsilon$

Penalized likelihood estimation

Fan and Lv (2011, IEEE-Information Theory)

Penalized likelihood estimation

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

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

Penalized likelihood:

$$n^{-1}\sum_{i=1}^{n}\{y_i\mathbf{x}_i^T\beta - b(\mathbf{x}_i^T\beta)\} - \sum_{j=1}^{p} p_\lambda(|\beta_j|)$$

$$= \mathbf{n}^{-1}\left[\mathbf{y}^T\mathbf{X}\beta - \mathbf{1}^T\mathbf{b}(\mathbf{X}\beta)\right] - \sum_{\mathbf{j}=1}^{\mathbf{p}} \mathbf{p}_\lambda(|\beta_\mathbf{j}|).$$

Sparsity: $p'_\lambda(0+) > 0$, **singularity at origin** (Antoniadis & Fan, 01).

Penalized likelihood estimation

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

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

Penalized likelihood:

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}\{y_i\mathbf{x}_i^T\beta - b(\mathbf{x}_i^T\beta)\} - \sum_{j=1}^{p} p_\lambda(|\beta_j|) \\ = \quad & \mathbf{n}^{-1}\left[\mathbf{y}^T\mathbf{X}\beta - \mathbf{1}^T\mathbf{b}(\mathbf{X}\beta)\right] - \sum_{\mathbf{j}=1}^{\mathbf{p}} \mathbf{p}_\lambda(|\beta_\mathbf{j}|). \end{aligned}$$

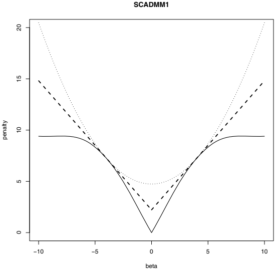
Sparsity: $p_\lambda'(0+) > 0$, **singularity at origin** (Antoniadis & Fan, 01).

Iterated reweighted L_1-estimator

Penalty: Popular choice L_1. Preferred: SCAD (*Fan & Li, 01*).

$$Q(\beta) = n^{-1} \sum_{i=1}^{n} L(Y_i, \mathbf{x}_i^T \beta) + \sum_{j=1}^{p} \mathbf{p}_\lambda(|\beta_\mathbf{j}|).$$

$$p_\lambda(|\beta_j^{(k)}|) + \mathbf{p}'_\lambda(|\beta_\mathbf{j}^{(\mathbf{k})}|)(|\beta_\mathbf{j}| - |\beta_j^{(k)}|)$$



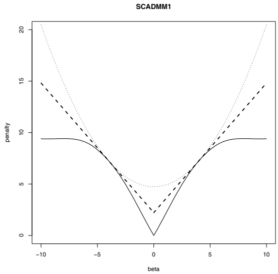
- $\beta^{(0)} = 0 \Longrightarrow$ LASSO.
- Iteration reduces the bias
- Zero is a non-absorbing state (comparing $w_j = 1/|\beta_j^{(k)}|^\gamma$).

Iterated reweighted L_1-estimator

Penalty: Popular choice L_1. Preferred: SCAD (*Fan & Li, 01*).

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

$$p_\lambda(|\beta_j^{(k)}|) + \mathbf{p}'_\lambda(|\beta_j^{(k)}|)(|\beta_j| - |\beta_j^{(k)}|)$$



■ $\beta^{(0)} = 0 \Longrightarrow$ 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:

$$Q(\beta^{(k+1)}) \le Q^{app}(\beta^{(k+1)}) \le Q^{app}(\beta^{(k)}) = Q(\beta^{(k)}).$$

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:

$$Q(\beta^{(k+1)}) \le Q^{app}(\beta^{(k+1)}) \le Q^{app}(\beta^{(k)}) = Q(\beta^{(k)}).$$

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.

Computing algorithms for SCAD

1. LQA algorithm (Fan and Li, 01).
2. LLA: Iterated reweighted LASSO (Zou and Li, 08).
3. PLUS: an extension of LARS Zhang (2009)
4. Coordinate optimization algorithm. (Fu and Jiang, 99, Li, Bühman, Hastie, Tibshirani, Fan, Lv)

■ L_1-penalty does not have much **computation advantages**.

Computing algorithms for SCAD

1. LQA algorithm (Fan and Li, 01).
2. LLA: Iterated reweighted LASSO (Zou and Li, 08).
3. PLUS: an extension of LARS Zhang (2009)
4. Coordinate optimization algorithm. (Fu and Jiang, 99, Li, Bühman, Hastie, Tibshirani, Fan, Lv)

■ L_1-penalty does not have much **computation advantages**.

Computing algorithms for SCAD

1. LQA algorithm (Fan and Li, 01).
2. LLA: Iterated reweighted LASSO (Zou and Li, 08).
3. PLUS: an extension of LARS Zhang (2009)
4. Coordinate optimization algorithm. (Fu and Jiang, 99, Li, Bühman, Hastie, Tibshirani, Fan, Lv)

■ L_1-penalty does not have much **computation advantages**.

Limited Capacity of L_1-penalty

■Consistent condition for LASSO is limited(Zhao and Yu, 06):
$\|(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_{2,j}\|_1 < 1$, **relaxed to** $\min(\frac{\mathbf{p}'_\lambda(\mathbf{0}+)}{\mathbf{p}'_\lambda(\mathbf{d_n})}, O(n^{\alpha_1}))$

■The capacity is about the same or weaker than for SIS.
[Fan and Lv (08), Fan and Song (10), Zhang (2010), Geneve, Jin, Wasserman (11)].

An executive summary

1. Give conditions under which FCPMLE is a **global** maximizer or **restricted global** maximizer.
2. FCPMLE possesses an oracle property up to **NP**-dimensionality: selection consistency + uniform rates + asymp. normality.
3. The result is applicable to L_1, but the condition for L_1 is much more restrictive than SCAD.
4. L_1 penalty does not possess the oracle property. The dimensionality and convergence rates need to compromise.

An executive summary

1. Give conditions under which FCPMLE is a **global** maximizer or **restricted global** maximizer.
2. FCPMLE possesses an oracle property up to **NP**-dimensionality:
   selection consistency $+$ uniform rates $+$ asymp. normality.
3. The result is applicable to L_1, but the condition for L_1 is much more restrictive than SCAD.
4. L_1 penalty does not possess the oracle property. The dimensionality and convergence rates need to compromise.

An executive summary

1. Give conditions under which FCPMLE is a **global** maximizer or **restricted global** maximizer.
2. FCPMLE possesses an oracle property up to **NP**-dimensionality: selection consistency + uniform rates + asymp. normality.
3. The result is applicable to L_1, but the condition for L_1 is much more restrictive than SCAD.
4. L_1 penalty does not possess the oracle property. The dimensionality and convergence rates need to compromise.

An executive summary

1. Give conditions under which FCPMLE is a **global** maximizer or **restricted global** maximizer.
2. FCPMLE possesses an oracle property up to **NP**-dimensionality: selection consistency $+$ uniform rates $+$ asymp. normality.
3. The result is applicable to L_1, but the condition for L_1 is much more restrictive than SCAD.
4. L_1 penalty does not possess the oracle property. The dimensionality and convergence rates need to compromise.

■ $\mathbf{X}$ full column rank and let β_* of $\ell_n(\beta)$.

■ $\mathcal{L}_c = \{\beta \in \mathbf{R}^p : \ell_n(\beta) \in [c, \ell_n(\beta_*)]\}$ for some $c < \ell_n(\mathbf{0})$.

Theorem 1: FCPMLE $\widehat{\beta}$ is a **global maximizer**, if

$$\min_{\beta \in \mathcal{L}_c} \lambda_{\min}\left[n^{-1}\mathbf{X}^T b''(\mathbf{X}\beta)\mathbf{X}\right] \geq \gamma(p_\lambda),$$

the maximum concavity.

Global optimality ($p \leq n$)

■ $\mathbf{X}$ full column rank and let β_* of $\ell_n(\beta)$.

■ $\mathcal{L}_c = \{\beta \in \mathbf{R}^p : \ell_n(\beta) \in [c, \ell_n(\beta_*)]\}$ for some $c < \ell_n(\mathbf{0})$.

Theorem 1: FCPMLE $\widehat{\beta}$ is a **global maximizer**, if

$$\min_{\beta \in \mathcal{L}_c} \lambda_{\min} \left[n^{-1} \mathbf{X}^T b''(\mathbf{X}\beta) \mathbf{X} \right] \geq \gamma(p_\lambda),$$

the maximum concavity.

Restricted Global optimality ($p > n$)

■The true model $\mathrm{supp}(\beta_0) = \{1, \cdots, s\}$

■$\mathbb{S}_s$: Union of all s-dimensional coordinate subspaces of $\mathbf{R}^p$.

Theorem 1': If the conditions 1 of Theorem 1 hold for each $n \times (2s)$ submatrix of $\mathbf{X}$, then the FCPMLE $\widehat{\beta}$ is a global maximizer on $\mathbb{S}_s$.

Restricted Global optimality ($p > n$)

■The true model $\mathrm{supp}(\beta_0) = \{1, \cdots, s\}$

■$\mathbb{S}_s$: Union of all s-dimensional coordinate subspaces of $\mathbf{R}^p$.

Theorem 1': If the conditions 1 of Theorem 1 hold for each $n \times (2s)$ submatrix of $\mathbf{X}$, then the FCPMLE $\widehat{\beta}$ is a global maximizer on $\mathbb{S}_s$.

Technical conditions

- min signal: $d_n = \min\{|\beta_{0,j}| : \beta_{0,j} \neq 0\}/2 \gg \mathbf{n}^{-\kappa} \log \mathbf{n}$.
- The design matrix $\mathbf{X}$ satisfies (for some $C < 1$)

$$\left\| \left[n^{-1} \mathbf{X}_1^T b''(\theta_0) \mathbf{X}_1 \right]^{-1} \right\|_\infty = O(b_s), \qquad b_s \to \infty; \qquad \theta_0 = \mathbf{X}\beta_0$$

$$\left\| \mathbf{X}_2^T b''(\theta_0) \mathbf{X}_1 \left[\mathbf{X}_1^T b''(\theta_0) \mathbf{X}_1 \right]^{-1} \right\|_\infty \leq \min\left(\mathbf{C} \frac{\mathbf{p}'_\lambda(\mathbf{0}+)}{\mathbf{p}'_\lambda(\mathbf{d_n})}, O(n^{\alpha_1}) \right).$$

For least squares, $b''(\cdot) = 1$, it reduces to **irrepresentable condition**.

♣For Lasso, RHS is bounded by C (**almost iff condition**).

♣For SCAD, LHS = $O(n^{\alpha_1})$, **much weaker**.

Technical conditions

- min signal: $d_n = \min\{|\beta_{0,j}| : \beta_{0,j} \neq 0\}/2 \gg \mathbf{n}^{-\kappa} \log \mathbf{n}$.
- The design matrix $\mathbf{X}$ satisfies (for some $C < 1$)

$$\left\| \left[n^{-1} \mathbf{X}_1^T b''(\theta_0) \mathbf{X}_1 \right]^{-1} \right\|_\infty = O(b_s), \qquad b_s \to \infty; \qquad \theta_0 = \mathbf{X}\beta_0$$

$$\left\| \mathbf{X}_2^T b''(\theta_0) \mathbf{X}_1 \left[\mathbf{X}_1^T b''(\theta_0) \mathbf{X}_1 \right]^{-1} \right\|_\infty \leq \min\left(\mathbf{C} \frac{\mathbf{p}'_\lambda(\mathbf{0}+)}{\mathbf{p}'_\lambda(\mathbf{d_n})}, O(n^{\alpha_1})\right).$$

For least squares, $b''(\cdot) = 1$, it reduces to

irrepresentable condition.

♣ For Lasso, RHS is bounded by C (**almost iff condition**).

♣ For SCAD, LHS = $O(n^{\alpha_1})$, **much weaker**.

Model selection consistency and rate of convergence

■**Choice of λ**: Letting $\alpha = \min(\frac{1}{2}, 2\kappa) - \alpha_1$,

$$p'_{\lambda_n}(d_n) = o(b_s^{-1} n^{-\kappa} \log n) \qquad \lambda_n \gg n^{-\alpha} (\log n)^2.$$

■**Capacity**: $s = o(n)$, $\qquad \log p = O(n^{1-2\alpha})$

Theorem 2: With probability $\geq 1 - 2[sn^{-1} + (p-s)e^{-n^{1-2\alpha}\log n}]$, there exists an estimator, satisfying:

- **Sparsistency**: $\widehat{\beta}_2 = \mathbf{0}$;
- **Uniform rate of convergence**: $\|\widehat{\beta}_1 - \beta_1\|_\infty = O(n^{-\kappa}\log n)$.

Model selection consistency and rate of convergence

Choice of λ: Letting $\alpha = \min(\frac{1}{2}, 2\kappa) - \alpha_1$,

$$p'_{\lambda_n}(d_n) = o(b_s^{-1} n^{-\kappa} \log n) \qquad \lambda_n \gg n^{-\alpha} (\log n)^2.$$

Capacity: $s = o(n)$, $\qquad \log p = O(n^{1-2\alpha})$

Theorem 2: With probability $\geq 1 - 2[sn^{-1} + (p-s)e^{-n^{1-2\alpha} \log n}]$, there exists an estimator, satisfying:

- **Sparsistency**: $\widehat{\beta}_2 = \mathbf{0}$;
- **Uniform rate of convergence**: $\|\widehat{\beta}_1 - \beta_1\|_\infty = O(n^{-\kappa} \log n)$.

Oracle Property

Theorem 3: With probability tending to one, there exists a local maximizer such that $\widehat{\beta}_2 = \mathbf{0}$ and $\|\widehat{\beta} - \beta_0\|_2 = O_P(\sqrt{s}n^{-1/2})$ with the following asymptotic normality:

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{D} N(\mathbf{0}, \phi\left[n^{-1}\mathbf{X}_1^T b''(\theta_0)\mathbf{X}_1\right]^{-1}).$$

Fisher information bound of an oracle estimator

For any $\mathbf{A}_n$ such that $\mathbf{A}_n\mathbf{A}_n^T \to \mathbf{G}$,

$$\mathbf{A}_n\left[\mathbf{X}_1^T b''(\theta_0)\mathbf{X}_1\right]^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{D} N(\mathbf{0}, \phi\mathbf{G}).$$

Oracle Property

Theorem 3: With probability tending to one, there exists a local maximizer such that $\widehat{\beta}_2 = \mathbf{0}$ and $\|\widehat{\beta} - \beta_0\|_2 = O_P(\sqrt{s}n^{-1/2})$ with the following asymptotic normality:

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{D} N(\mathbf{0}, \phi\left[n^{-1}\mathbf{X}_1^T b''(\theta_0)\mathbf{X}_1\right]^{-1}).$$

Fisher information bound of an oracle estimator

For any $\mathbf{A}_n$ such that $\mathbf{A}_n\mathbf{A}_n^T \to \mathbf{G}$,

$$\mathbf{A}_n\left[\mathbf{X}_1^T b''(\theta_0)\mathbf{X}_1\right]^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{D} N(\mathbf{0}, \phi\mathbf{G}).$$

Oracle Property

Theorem 3: With probability tending to one, there exists a local maximizer such that $\widehat{\beta}_2 = \mathbf{0}$ and $\|\widehat{\beta} - \beta_0\|_2 = O_P(\sqrt{s}n^{-1/2})$ with the following asymptotic normality:

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{D} N(\mathbf{0}, \phi\left[n^{-1}\mathbf{X}_1^T b''(\theta_0)\mathbf{X}_1\right]^{-1}).$$

Fisher information bound of an oracle estimator

For any $\mathbf{A}_n$ such that $\mathbf{A}_n\mathbf{A}_n^T \to \mathbf{G}$,

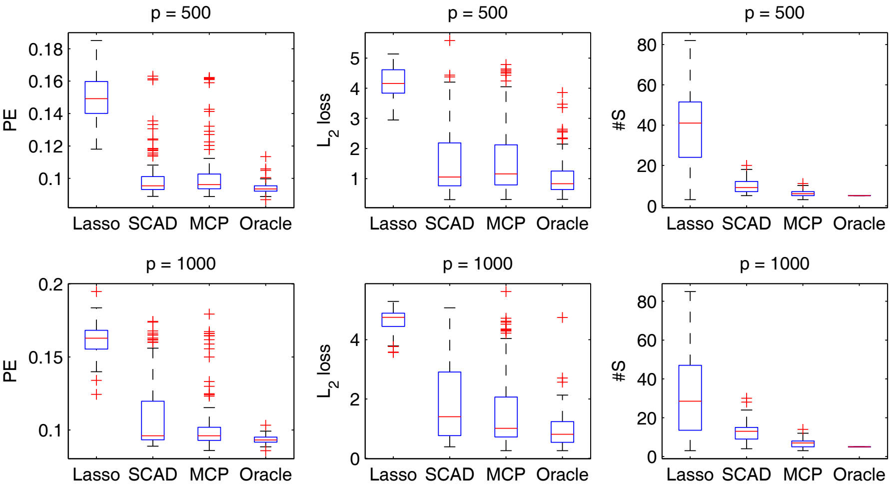
$$\mathbf{A}_n\left[\mathbf{X}_1^T b''(\theta_0)\mathbf{X}_1\right]^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{D} N(\mathbf{0}, \phi\mathbf{G}).$$

Logistic regression — small p

- Covariate $\mathbf{x} \sim N(0, \Sigma)$ with $\Sigma = (0.5^{|i-j|})$.
- $\beta_1 = (2.5, -1.9, 2.8, -2.2, 3)^T$, $n = 200$, $p = 25$.

Measures	Lasso	SCAD	MCP	Oracle
PE	**0.11**(0.01)	**0.10**(0.01)	0.10(0.01)	0.09(0.00)
L_2 loss	**3.06**(0.66)	**0.94**(0.55)	0.94(0.55)	0.88(0.34)
L_1 loss	**7.25**(1.10)	**1.87**(1.46)	1.87(1.46)	1.73(0.77)
Deviance	**129.4**(19.2)	**111.8**(15.8)	111.82(15.80)	113.12(16.0
#S	**9(**2.97)	**5**(0.74)	5(0.74)	5(0)
FN	0(0)	0(0)	0(0)	0(0)

Logistic regression — large p



Poisson regression

■ $n = 200$, $p = 1000$, $\beta_1 = (1.25, -0.95, 0.9, -1.1, 0.6)^T$

	Lasso	SCAD	MCP	Oracle
PE	**33.07**(14.09)	**5.52**(2.03)	5.14(1.81)	3.68(0.77)
L_2 loss	**0.97**(0.21)	**0.21**(0.09)	0.19(0.09)	0.108(0.047)
L_1 loss	**2.99**(0.69)	**0.49**(0.23)	0.443(0.20)	0.20(0.09)
Deviance	**200.0**(22.9)	**180.3**(13.1)	181.2(15.3)	187.98(17.22
#S	**34**(7.41)	**11.5**(4.08)	9(2.22)	5(0)
FN	0(0)	0(0)	0(0)	0(0)

■ **Bias of LASSO** forces selecting more var. and increase PE.

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. 251 customized oligonucleotide microarray with $p = 10,707$.
3. focus on "3-year Event Free Survival", ($n = 239$ w/ 49 "$+$" and 190 "$-$").
4. Aims: To study which genes are responsible for neuroblastoma and their 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. 251 customized oligonucleotide microarray with $p = 10,707$.
3. focus on "3-year Event Free Survival", ($n = 239$ w/ 49 "$+$" and 190 "$-$").
4. Aims: To study which genes are responsible for neuroblastoma and their risk association.

Results

Training set and endpoints:

1. **"3-y EFS"**: Random 25 "$+$" and 100 "$-$".
2. **"Gender"**: Random 120 males and 50 females. Total: 246.

Table: Classification errors in the neuroblastoma data set

	3-year EFS		Gender	
Method	# of genes	Test error	# of genes	Test error
Lasso	**56**	**23/114**	**4**	**5/126**
SCAD	**10**	**18/114**	**2**	**4/126**
MCP	7	23/114	1	12/126
SIS	5	19/114	6	4/126
ISIS	23	22/114	2	4/126

Results

Training set and endpoints:

1. **"3-y EFS"**: Random 25 "$+$" and 100 "$-$".
2. **"Gender"**: Random 120 males and 50 females. Total: 246.

Table: Classification errors in the neuroblastoma data set

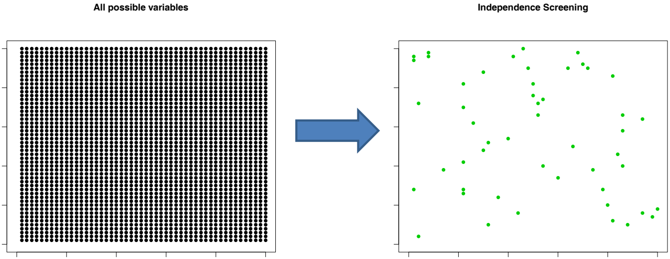
	3-year EFS		Gender	
Method	# of genes	Test error	# of genes	Test error
Lasso	**56**	**23/114**	**4**	**5/126**
SCAD	**10**	**18/114**	**2**	**4/126**
MCP	7	23/114	1	12/126
SIS	5	19/114	6	4/126
ISIS	23	22/114	2	4/126

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, 11*); ★**Nonparametric learning** (*Fan, Feng, Song, 09*) ★**Data-tilting**; (*Hall, Titterington & Xue, 09*).

1. Sure screening property? In what capacity? (*Fan & Lv, 08*)
2. Model selection consistency? (*Geneve, Jin, Wasserman, 11*)
3. How to choose a thresholding parameter? (*Zhao & Li, 11*)
4. How to reduce FDR? (*Fan, Samworth, Wu, 09*)
5. What are the possible drawbacks?

Extensions & Questions

Other methods: ★**Marginal LR** (*Fan, Samworth & Wu, 09*); ★**MMLE** (*Fan and Song, 09*); ★**MPLE** (*Zhao & Li, 11*); ★**Nonparametric learning** (*Fan, Feng, Song, 09*) ★**Data-tilting**; (*Hall, Titterington & Xue, 09*).

1. Sure screening property? In what capacity? (*Fan & Lv, 08*)
2. Model selection consistency? (*Geneve, Jin, Wasserman, 11*)
3. How to choose a thresholding parameter? (*Zhao & Li, 11*)
4. How to reduce FDR? (*Fan, Samworth, Wu, 09*)
5. What are the possible drawbacks?

Extensions & Questions

Other methods: ★**Marginal LR** (*Fan, Samworth & Wu, 09*);
★**MMLE** (*Fan and Song, 09*); ★**MPLE** (*Zhao & Li, 11*);
★**Nonparametric learning** (*Fan, Feng, Song, 09*)
★**Data-tilting**; (*Hall, Titterington & Xue, 09*).

1. Sure screening property? In what capacity? (*Fan & Lv, 08*)
2. Model selection consistency? (*Geneve, Jin, Wasserman, 11*)
3. How to choose a thresholding parameter? (*Zhao & Li, 11*)
4. How to reduce FDR? (*Fan, Samworth, Wu, 09*)
5. What are the possible drawbacks?

Potential Drawbacks

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

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

- **False Positive**: What if $X_2, \cdots, X_{99}$ 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?

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

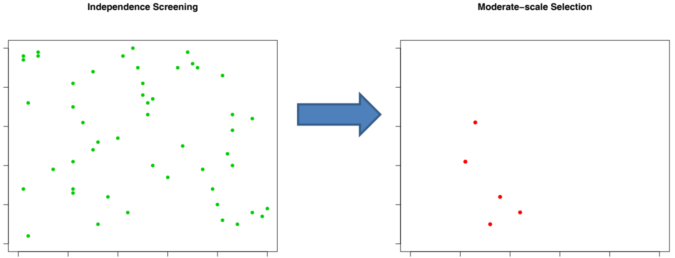
- **False Positive**: What if $X_2, \cdots, X_{99}$ highly correlated with an important X_1, but weakly correlated with Y conditionally?

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

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 coefs and choose variables.

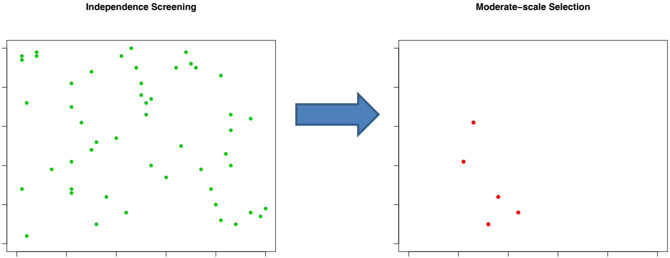


- How high dimensionality can such methods handle?
- What is the role of penalty functions?
- Does it possess an oracle property? How to compute?

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 coefs 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?

Carbon Atom: Iteration

Iterative application of

large-scale **screening** and

moderate-scale **selection**.

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

Iterative feature selection

1. **(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.

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

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

   resulting in $\mathcal{A}_2$.

Iterative feature selection

1. ■**(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.

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

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

   resulting in $\mathcal{A}_2$.

Illustration of ISIS

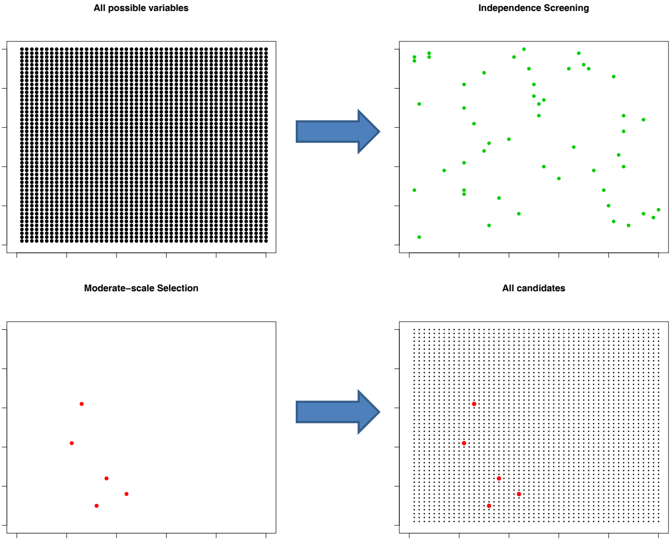
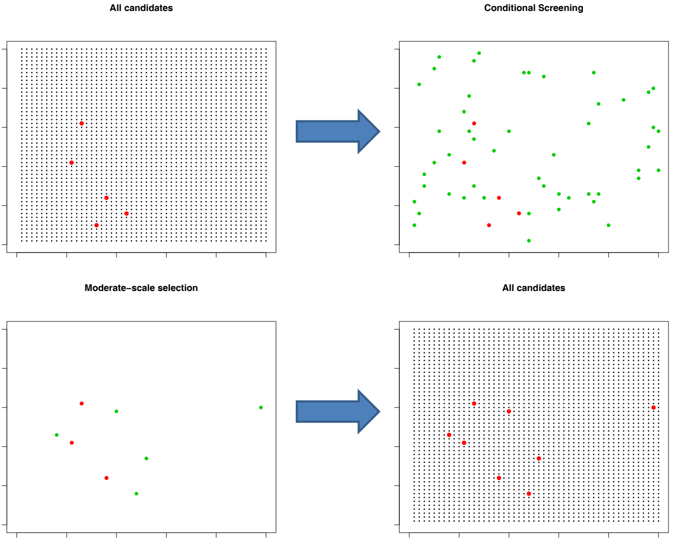


Illustration of ISIS



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 $\mathcal{M}_\ell = \mathcal{M}_{\ell-1}$.

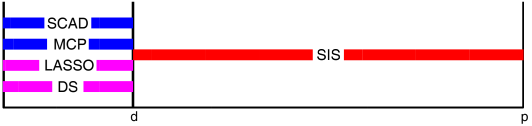
Iterative feature selection (II)

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, 11*)

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, 11*)

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, 11*)

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, 11*)

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, 11*)

Logistic regression, a very difficult case

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

Bayes error: 0.1040. $n = 400, p = 1000, N_{sim} = 100$

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

Sure Independence Screening

Fan and Song (2010, Ann. Statist.)

Model setting

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

- **GLIM**: $L(Y_i, \mathbf{x}_i^T\beta) = b(\mathbf{x}_i^T\beta) - Y_i \mathbf{x}_i^T \beta$, as

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

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

- **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

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

- **GLIM**: $L(Y_i, \mathbf{x}_i^T\beta) = b(\mathbf{x}_i^T\beta) - Y_i\mathbf{x}_i^T\beta$, as

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

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

- **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-Utilility: **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 $EX_j^2 = 1$.

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

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

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

Independence learning

M-Utilility: **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 $EX_j^2 = 1$.

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

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

Theorem 1: $L_j^\star = 0 \Longleftrightarrow \mathrm{cov}(Y, X_j) = 0 \Longleftrightarrow \beta_j^M = 0.$

Theoretical Basis

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

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

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

- If **active** indep of **inactive**, then $L_j^\star = 0, j \notin \mathcal{M}_\star$

$\Longrightarrow$ model selection consistency.

Theoretical Basis

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

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

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

■If **active** indep of **inactive**, then $L_j^\star = 0, j \notin \mathcal{M}_\star$

$\Longrightarrow$ model selection consistency.

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!

- Note that $\hat{L}_j - L_j^\star = O(\mathbf{n}^{-1/2})$ 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 $\nu_n = cn^{-2\kappa}$ 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!

- Note that $\hat{L}_j - L_j^\star = O(\mathbf{n}^{-1/2})$ 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 $\nu_n = cn^{-2\kappa}$ 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!

- Note that $\hat{L}_j - L_j^\star = O(\mathbf{n}^{-1/2})$ 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 $\log p_n = o(n^{1-2\kappa})$,

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

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

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

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

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

3. What is the selected model size? We establish

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

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

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

3. What is the selected model size? We establish

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

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

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

3. What is the selected model size? We establish

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

4. The $\#\{|\beta_j^M| \ge \gamma_n\}$ 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. $P(\min_{j\in\mathcal{M}_\star} |\hat{\beta}_j^M| \geq \gamma_n) \to 1$, if $\gamma_n = c_0 n^{-\kappa}$, $c_0 < c_1/2$.

3. What is the selected model size? We establish

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

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

Performance of Independence Screening

■compare **minimum model size** for sure screening w/ LASSO.

■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.

■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
	$s = 3$, $\beta^\star = (1, 1.3, 1)^T$			$s = 6$, $\beta^\star = (1, 1, 3, 1, \ldots)^T$		
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)
	$s = 12$, $\beta^\star = (1, 1.3, \ldots)^T$			$s = 15$, $\beta^\star = (1, 1.3, \ldots)^T$		
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, $p = 5,000$, $q = 15$

ρ	n	SIS-MLR	SIS-MMLE	LASSO	SCAD
	$s = 6$, $\beta^\star = (1, 1.3, 1, 1.3, 1, 1.3)^T$				
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)
	$s = 12$, $\beta^\star = (1, 1.3, \ldots)^T$				
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)
	$s = 15$, $\beta^\star = (3, 4, \ldots)^T$				
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)

Summary

- 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.

Summary

- 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.

Summary

- 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.

Summary

- 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.

Summary

- 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, 11*)
- ★ Richard Samworth (*Cambridge University; FSW, 2009*).
- ★ Rui Song (*Colorado State University, Fan & Song, 2009*).
- ★ Yichao Wu (*North Carolina State University, FSW, 2009*).