Statistical Foundations of Data Science

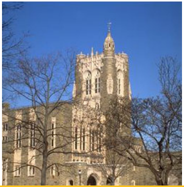
Jianqing Fan

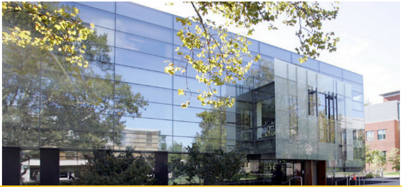
Princeton University

https://fan.princeton.edu

ZOOM ID Lectures: **970 4936 8998** Office Hours: **996 4030 7631**

Annotated Lecture Notes: web view





4. Feature Screening and Selection

4.1. Independence Screening (§8.1)

4.2. Iteratively Independent Learning (§8.3)

4.1 Sure Independence Screening

Available in R package: SIS

Independence Screening

Regression: Feature ranking by **marginal corr** $\{|\widehat{\mathrm{corr}}(X_j, Y)|\}$.

★Easily to implement ★Scalable to Big Data

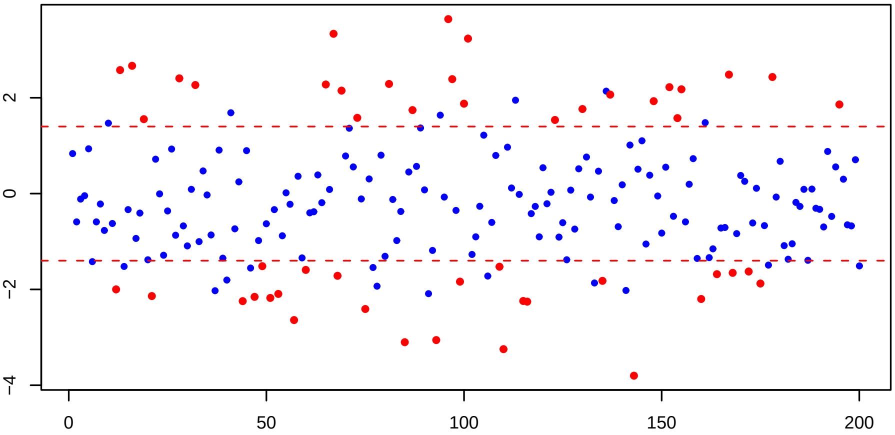
Classification$(Y = \pm 1)$: Feature ranking by two-sample t-tests or other tests.

(Applications: Sentiment analysis — selecting a bag of words related to financial returns *(Ke, Kelly, Xiu, 2019)*)

Sure Screening: Selected $\widehat{\mathcal{S}}$ contains all important variables $\mathcal{S}$.

Sure Independent Screening (SIS): Correlation learning has sure screening property *(Fan and Lv, 2008, JRSS-B)*: $P(\mathcal{S} \subset \widehat{\mathcal{S}}) \to 1$.

An illustration



x-axis: label of variables y-axis: correlation with Y, red: $\mathcal{S}$ blue: $\mathcal{S}^c$

A Framework for Independence Screening

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

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

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

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

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

Dim. reduction: From high to moderate dimensions

200 10000

Square-loss

When $L(Y, \mathbf{X}^T\beta) = (Y - \mathbf{X}^T\beta)^2$, we have *(homework)*

$$\widehat{L}_j = \widehat{r}_j^2 \widehat{L}_0, \qquad \widehat{\beta}_j = \widehat{r}_j \widehat{L}_0^{1/2}$$

Both reduce to the correlation ranking *(Fan and Lv, 2008)*.

Generalized correlation: Use multiple R^2 based on univariate polynomial regression *(Hall and Miller, 09)*.

Nonparametric screening: Use multiple R^2 based on univariate spline regression *(NIS, Fan, Feng, Song, 10)*.

Extensions and Questions

- ★ **Marginal LR** *(Fan, Samworth & Wu, 09)*;
- ★ **MMLE** *(Fan and Song, 10)*; ★**MPLE** *(Zhao & Li, 12)*; ★**D-corr** *(Li, Zhong, Zhu, 12)*; ★**Rank-corr** *(Li, et. al, 12)*;
- ★ **Nonparametric learning** *(Fan, Feng, Song, 10)*

1. Can we have model selection consistency?
2. Can we have sure screening property? In what capacity?
3. How to choose a thresholding parameter?

Choice of thresholding parameter

Threshold parameter: maximum marginal utility under **null model**, estimated by random decoupling, called Principled SIS *(Zhao and Li, 12)*.

- Obtain the decoupled synthetic data $\{(\mathbf{X}_{\pi(i)}, Y_i)\}_{i=1}^n$ —Marginal distributions are untouched;
- Compute $a_n^* = \max_j \widehat{L}_j^*$ based on decoupled data.
  For correlation learning, this becomes $\mathrm{corr}^2(\{(X_{\pi(i)j}, Y_i)\})$.
- Choose the top α-quantile of a_n^* as ν_n.

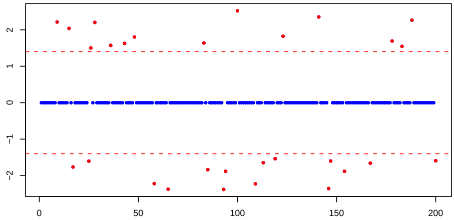
Remark: We can take a_n^* based on one permutation.

Theoretical Basis

Marginal utility: $L_j^\star = EL(Y, \beta_0^M) - \min EL(Y, \beta_0 + \beta_j X_j)$. **Likelihood-ratio** (*Fan and Song, 10)* **True model**: $\mathcal{M}_\star = \{j : \beta_j^\star \neq 0\}$.

Theorem 4.1: 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** $\mathbf{X}_{\mathcal{M}_\star}$ indep of **inactive** $\mathbf{X}_{\mathcal{M}_\star^c}$, then $L_j^\star = 0, j \notin \mathcal{M}_\star$

$\Longrightarrow$ model sel consistency, if gap is wide enough.

Sure independence screening

Thm 4.2: 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!

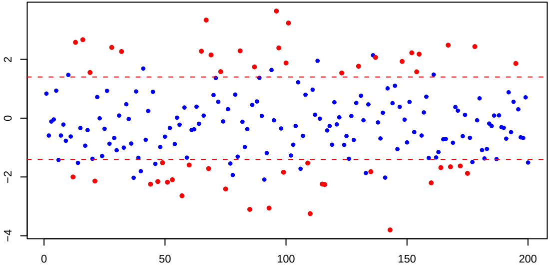
- Screening using **Wald stat** $\widehat{\beta}_j^M$ has also SS property.

Controlling number of features

Theorem 4.3: If $\log p_n = o(n^{1-2\kappa})$
,

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

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



Performance of Independence Screening

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

■Consistent cond for Lasso is stringent: $\|(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_{2,j}\|_1 < 1$.

Design 1: $\{X_j = \frac{\varepsilon_j + a_j\varepsilon}{\sqrt{1+a_j^2}}\}_{j=1}^q$, w/ correlation $\frac{a_j^2}{1+a_j^2}$, rest indep.
a_j generated from $N(a,1)$ with $\rho = \frac{a^2}{1+a^2}$

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)

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

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

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

Regression Coefs: $\beta^\star = (1,-1,1,-1,\cdots)^T$

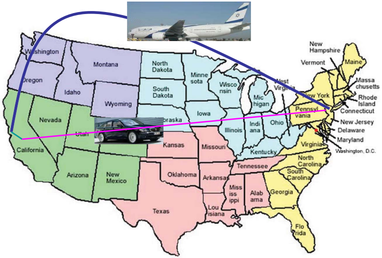
(RSD = $\frac{IQR}{1.35}$)

s	M-$\lambda_{\max}$(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)*

4.2 Iteratively SIS Method

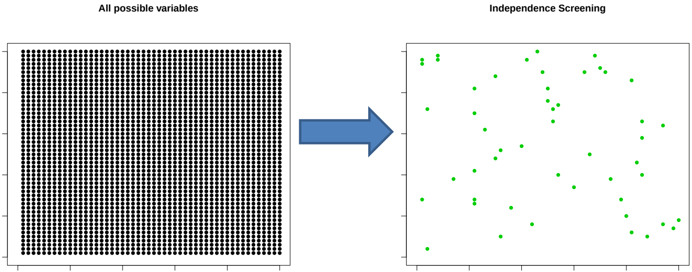
a two-scale framework

(Fan, Samworth, Wu, 2009, JMLR)



Large scale-screening

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



Potential Drawbacks

- **False Negative**: What if X_1 marginally uncorrelated with Y, but jointly correlated with Y?

$$Y = \beta_1 \mathbf{X_1} + X_2 + X_3 + X_4 + X_5 + \varepsilon \quad \text{s.t.} \quad \text{cov}(Y, X_1) = 0.$$

★e.g. $\text{corr}(X_i, X_j) = 0.8$ for all $i, j < p$. $\text{cov}(Y, X_1) = \beta_1 + 4 * .8$. With $\beta = -3.2$, X_1 can not survive screening.

- **False Positive**: What if variables highly correlated with important ones, but weakly correlated with Y conditionally? e.g. X_{100} indep of X_i for $i < 99$

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

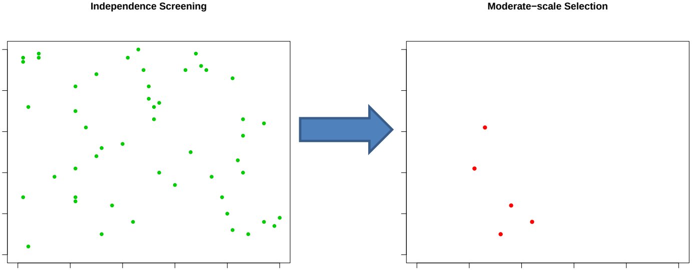
$cov(X_j, Y) = 0.8$, $2 \le j \le 99$ whereas $\text{cov}(X_j, Y) = 0.2$.

Moderate scale selection

Penalized likelihood estimation on survived variables after screening

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

■ Simultaneously estimate coefs and choose variables.



Iterations

Iterative application of

large-scale conditional **screening** and

moderate-scale **selection**.

- Iter-SIS *(Fan & Lv, 08; Fan, Samworth & Wu, 09)*, **available in R**.

Iterative feature selection

1. **(Large-scale screening)**: Apply SIS to pick a set $\mathcal{A}_1$;

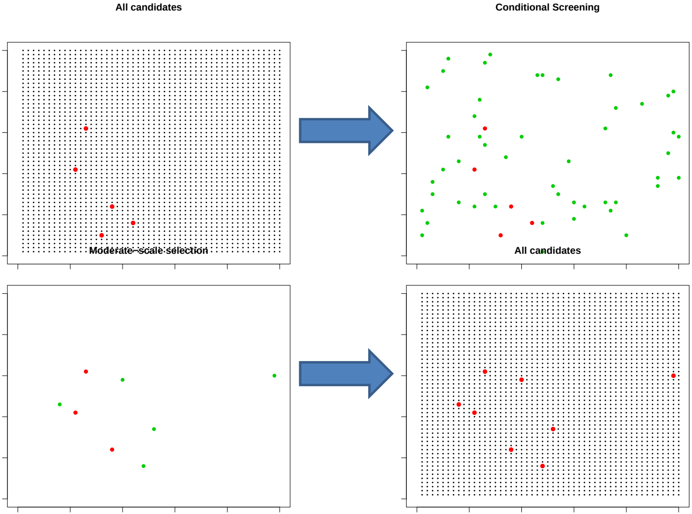
   (Moderate-scale selection): Employ a penalized likelihood to select a subset $\mathcal{M}_1$ of these indices.

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 $\mathcal{A}_2$.

Illustration of Iter-SIS



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**.

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

Applicability of Iter-SIS idea

The idea of Iter-SIS is widely applicable. It can be applied to

- Classification *(Fan, Samworth, & Wu, 09)*.
- Survival analysis *(Fan, Feng, & Wu, 10; Zhao & Li, 12)*.
- Nonparametric learning *(Fan, Feng, & Song, 10)*.
- Robust and quantile regression *(Bradic, Fan, & Wang, 11)*
- Available in R package: **SIS**

Logistic, a difficult case

★$\mathbf{X}_i \sim N(0,1)$, $\mathrm{cov}(X_i, X_j) = 1/\sqrt{2}, i \neq j < p$, indep of X_p.

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

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

	Van-SIS	Iter-SIS	Iter-SIS2	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
$2Q(\widehat{\beta}_0, \widehat{\beta})$(training)	269	188	188	109	N/A
AIC	289	198	199	311	N/A
BIC	337	218	219	714	N/A
$2Q(\widehat{\beta}_0, \widehat{\beta})$ (test)	361	225	226	276	N/A
0-1 test error	**.193**	**.112**	.112	**.146**	**.387**