UPS Delivers Optimal Phase Diagram in Variable Selection

Non-optimal Regions for L^1 and L^0-Penalization Methods

Jiashun Jin (STAT. Carnegie Mellon)

With Pengsheng Ji (Math. Cornell)

June 9, 2011

Variable selection

$$Y = X\beta + z, \qquad X = X_{n,p}, \qquad z \sim N(0, I_n)$$

- $p \gg n \gg 1$
- β is sparse: many coordinates are 0
- Gram matrix (of the column) $X'X$
  - has unit diagonals
  - is sparse (few large coordinates in each row)
- Ex. Compressive Sensing, Genomics ....

Subset selection

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \frac{\lambda^2}{2}\|\beta\|_0$$

- L^0-penalization method
- Variants: Cp, AIC, BIC, RIC
- Computationally challenging

Mallows (1973), Akaike (1974), Schwartz (1978), Foster & George (1994)

The lasso

$$\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

- L^1-penalization method; Basis Pursuit
- Widely used
  - computationally efficient even when p is large
  - if sufficiently sparse and noiseless, equivalent to L^0-penalization

Chen et al. (1998); Tibshirani (1996); Donoho (2006)

Graph of strong correlation

Let $X = [x_1, \ldots, x_p]$, define graph $G = (V, E)$:

- $V = \{1, 2, \ldots, p\}$: each variable is a node
- Nodes i & j are connected if and only if

$$\left|(X'X)(i,j)\right| \geq \frac{1}{\log(p)}, \qquad \text{say}$$

- G is sparse in many applications

Signal sparsity and graph sparsity

- Despite its sparsity, G is usually complicate
- Denote the support of β by

$$S = S(\beta) = \{1 \leq i \leq p,\ \beta_i \neq 0\}$$

  Restricting nodes to S forms a subgraph G_S
- **Key insight**: G_S splits into many **small-size** component that are disconnected to each other

Component: a maximal connected subgraph

For today

Univariate **P**enalization **S**creening (UPS)

U-step: screen by univariate thresholding

P-step: clean by penalized MLE

U-step

$$Y = X\beta + z, \qquad X = X_{n,p}, \qquad z \sim N(0, I_n).$$

- Let $\tilde{Y} = X'Y$. For a threshold $t_p^* > 0$ TBD,

  keep the j-th variable $\iff$ $\tilde{Y}_j > t_p^*$

- Denote the set of survived indices by

$$\mathcal{U}_p(t_p^*) = \{1 \le j \le p,\ \tilde{Y}_j > t_p^*\}.$$

Donoho (2006), Fan and Lv (2008), Genovese et al. (2010)

Use one-sided thresholding for simplicity

Two important properties of U-step

$$\mathcal{U}_p(t_p^*) = \{1 \le j \le p,\ \tilde{Y}_j > t_p^*\}: \qquad \text{survived indices}$$

If both signals and Graph G are sparse:

- **Sure Screening (SS)**: $\mathcal{U}_p(t_p^*)$ retains all *but a small proportion* of signals
- **Separable After Screening (SAS)**: $\mathcal{U}_p(t_p^*)$ splits into many components, each is small in size, and different ones are disconnected

Original problem reduces to **many small-size regression** problems

Reduce to many small-size regression

$$\tilde{Y} = X'Y, \qquad \mathcal{I}_0 \subset \mathcal{U}_p(t_p^*): \quad \text{a component}$$

$(X'X)^{\mathcal{I}_0}$: row restriction; $(X'X)^{\mathcal{I}_0,\mathcal{I}_0}$: row & column restriction

- Restrict regression to $\mathcal{I}_0$

$$Y = X\beta + z \implies \tilde{Y} = X'X\beta + X'z$$
$$\implies \tilde{Y}^{\mathcal{I}_0} = (X'X\beta)^{\mathcal{I}_0} + (X'z)^{\mathcal{I}_0}$$

- $(X'z)^{\mathcal{I}_0} \sim N(0, (X'X)^{\mathcal{I}_0,\mathcal{I}_0})$ since $z \sim N(0, I_n)$
- **Key**: $(X'X\beta)^{\mathcal{I}_0} \approx (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}$
- **Result**: many small-size regression:

$$\tilde{Y}^{\mathcal{I}_0} \approx N\big((X'X)^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}, (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\big)$$

Reduce to small-size regression, II

Why $(X'X\beta)^{\mathcal{I}_0} = (X'X)^{\mathcal{I}_0}\beta \approx (X'X)^{\mathcal{I}_0,\mathcal{I}_0}\beta^{\mathcal{I}_0}$?

$$\left[\ (X'X)^{\mathcal{I}_0,\mathcal{I}_0} \quad \blacksquare \quad (X'X)^{\mathcal{I}_0,\mathcal{J}_0} \quad \blacksquare \quad \dots\ \right] \begin{bmatrix} \beta^{\mathcal{I}_0} \\ 0 \\ \beta^{\mathcal{J}_0} \\ 0 \\ \dots \end{bmatrix}$$

- $\mathcal{I}_0, \mathcal{J}_0 \subset \mathcal{U}_p(t_p^*)$: components
- By SS property, $\beta^{\blacksquare} = 0$
- By SAS property, $(X'X)^{\mathcal{I}_0,\mathcal{J}_0} \approx 0$

P-step

$$Y = X\beta + z, \qquad z \sim N(0, I_n)$$

- $\tilde{Y} = X'Y$; $\mathcal{I}_0 \subset \mathcal{U}_p(t_p^*)$: a component
- $\beta^{\mathcal{I}_0}$: restricting rows of β to $\mathcal{I}_0$
- M_0: restricting rows/columns of $X'X$ to $\mathcal{I}_0$

Fixing (λ^{ups}, u^{ups}),

- estimate $\beta^{\mathcal{I}_0}$ via minimizing

$$[\tilde{Y}^{\mathcal{I}_0} - M_0\beta^{\mathcal{I}_0}]' M_0^{-1} [\tilde{Y}^{\mathcal{I}_0} - M_0\beta^{\mathcal{I}_0}] + (\lambda^{ups})^2 \|\beta^{\mathcal{I}_0}\|_0,$$

  where $\beta^{\mathcal{I}_0}$ takes values from $\{0, u^{ups}\}$
- $j \notin \mathcal{U}_p(t_p^*)$: set $\hat{\beta}_j = 0$

Sparse model

$$Y = X\beta + z, \qquad z \sim N(0, I_n)$$

- Sparse signals:

$$\beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \pi_p$$

where ν_0 is point mass at 0 and

$$\epsilon_p = p^{-\vartheta}, \qquad 0 < \vartheta < 1$$

Measuring errors with Hamming distance:

$$\mathrm{Hamm}_p(\hat{\beta}; \epsilon_p, \pi_p | X) = E\left[\sum_{j=1}^{p} 1\{\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)\}\right]$$

Universal lower bound

$$\beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \pi_p, \qquad \epsilon_p = p^{-\vartheta}$$

Theorem 1. Fix $r > 0$ and let $\tau_p = \sqrt{2r \log p}$. If the support of π_p is contained in $[-\tau_p, 0) \cup (0, \tau_p]$, then as $p \to \infty$,

$$\frac{\mathrm{Hamm}_p(\hat{\beta}; \epsilon_p, \pi_p | X)}{p\epsilon_p} \gtrsim \begin{cases} L_p \cdot p^{-(r-\vartheta)^2/(4r)}, & r > \vartheta, \\ 1, & r < \vartheta, \end{cases}$$

where L_p is a generic multi-log(p) term.

Random design model

$$Y = X\beta + z, \qquad X = \begin{pmatrix} X_1' \\ \dots \\ X_n' \end{pmatrix}, \qquad X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega)$$

- Ω: unknown correlation matrix
- $n = n_p = p^\theta$ and $(1 - \vartheta) < \theta < 1$, so that

$$p\epsilon_p \ll n_p \ll p$$

- Ex: Compressive Sensing, Computer Security

Dinur and Nissim (2004), Nowak et al. (2007)

Additional Assumptions

$$\beta_j \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p \pi_p, \quad \epsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r\log p}$$

$$\eta = \min\{\frac{2\vartheta}{r},\ 1-\frac{\vartheta}{r},\ \sqrt{2-2\omega_0}-1+\frac{\vartheta}{r}\}$$

- Support of $\pi_p \subset [\tau_p, (1+\eta)\tau_p], \quad r > \vartheta \qquad (*)$
- $\sum_{j=1}^{p} |\Omega(i,j)|^{\gamma} \le C,\ \gamma \in (0,1)$
- $\max\{\|U(\Omega)\|_1, \|U(\Omega)\|_\infty\} \le \omega_0 < 1/2 \qquad (*)$
- $\Omega(i,j) \ge 0$ or $r/\vartheta \le 3+2\sqrt{2} \qquad (*)$

(*): relaxable; Jin and Zhang (2011), in progress

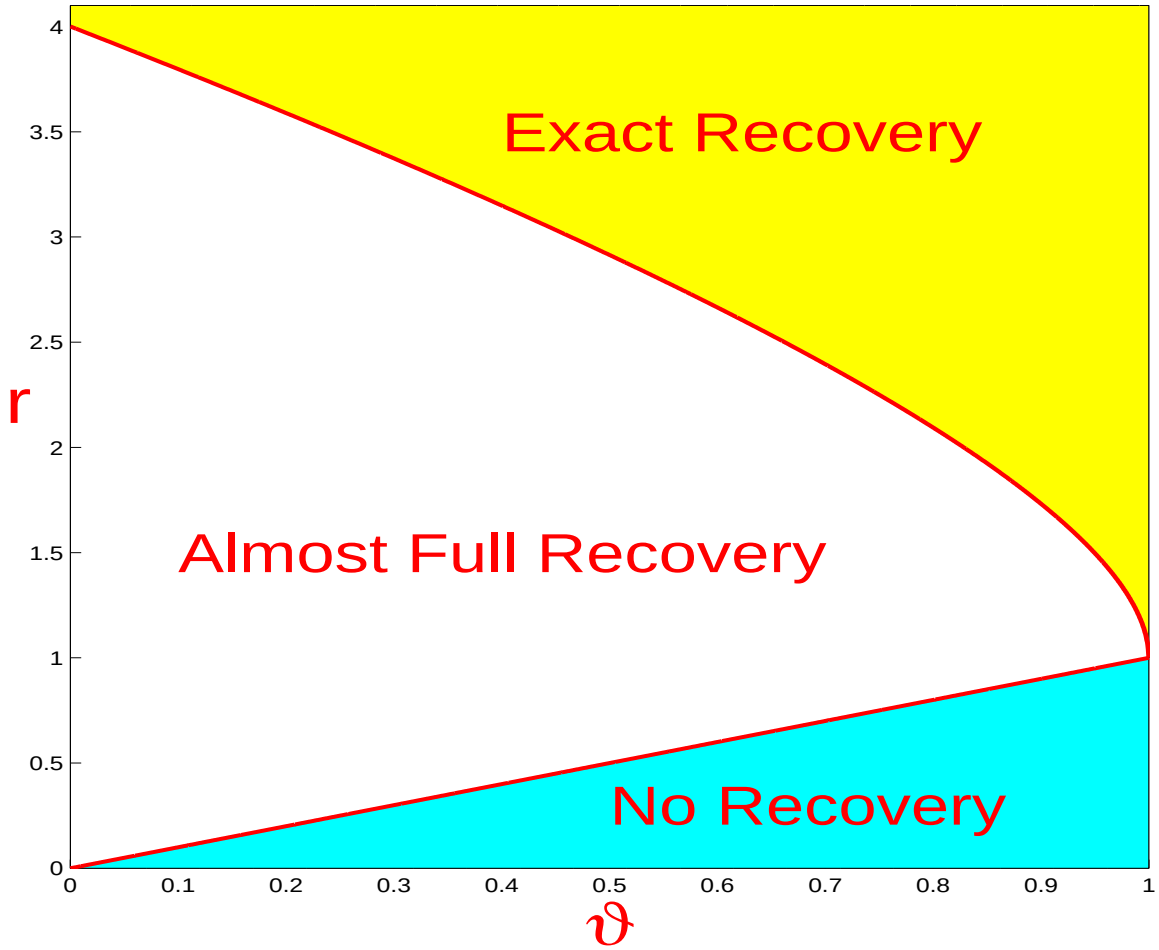
Main Results

Theorem 2. Set $t_p^* = \sqrt{2q \log p}$, $q \in (0, \frac{(\vartheta+r)^2}{4r}]$, $\lambda^{ups} = \sqrt{2\vartheta \log p}$, and $u^{ups} = \tau_p$. As $p \to \infty$,

- Both SS and SAS property hold
- UPS achieves optimal rate of convergence

$$\frac{E_\Omega[\mathrm{Hamm}_p(\hat{\beta}^{ups}; \epsilon_p, \pi_p | X)]}{p\epsilon_p} \leq L_p p^{-\frac{(r-\vartheta)^2}{4r}}.$$

UPS delivers optimal phase diagram



$\epsilon_p = p^{-\vartheta}$; $\tau_p = \sqrt{2r \log p}$; line: $r = \vartheta$; curve: $r = (1 + \sqrt{1 - \vartheta})^2$

Non-optimality of lasso & subset selection

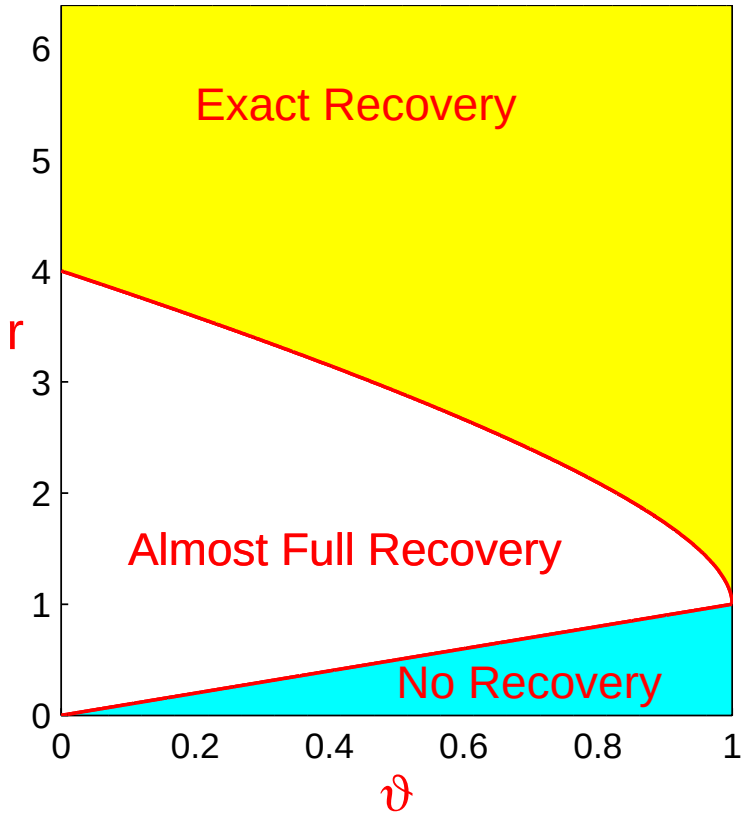
$$\epsilon_p = p^{-\vartheta}, \qquad \tau_p = \sqrt{2r\log p}, \qquad X_i \overset{iid}{\sim} N(0, \frac{1}{n}\Omega)$$

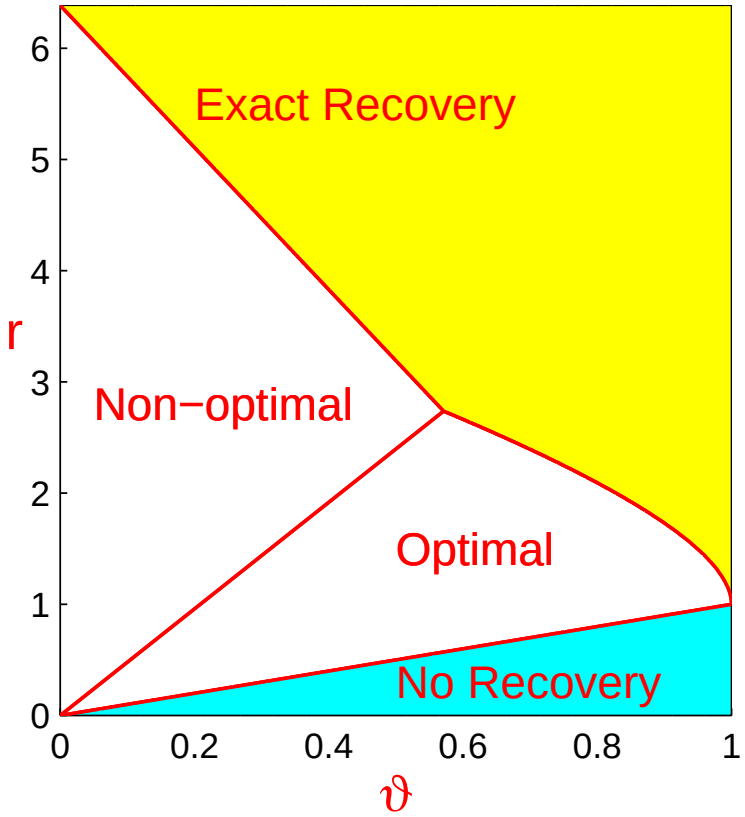
$$\beta_j = \begin{cases} \tau_p, & \text{prob. } \epsilon_p, \\ 0, & \text{prob. } 1-\epsilon_p \end{cases}$$

$$\Omega(i,j) = 1\{i=j\} + a\cdot 1\{|i-j|=1\}, \quad a \in (0,1/2)$$

- UPS achieves the optimal phase diagram
- The lasso and the subset selection do not, even with tuning parameters set ideally

Non-optimal region (lasso)

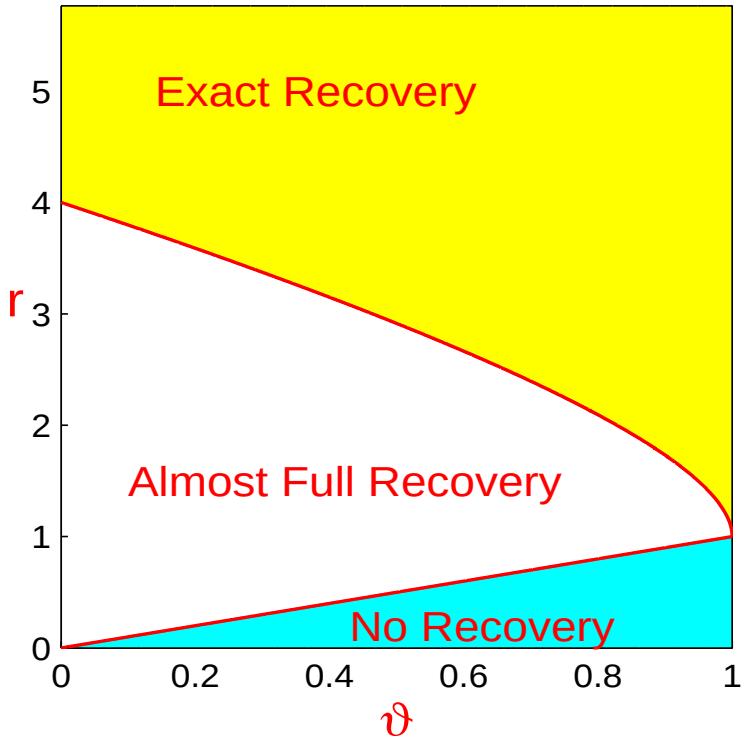


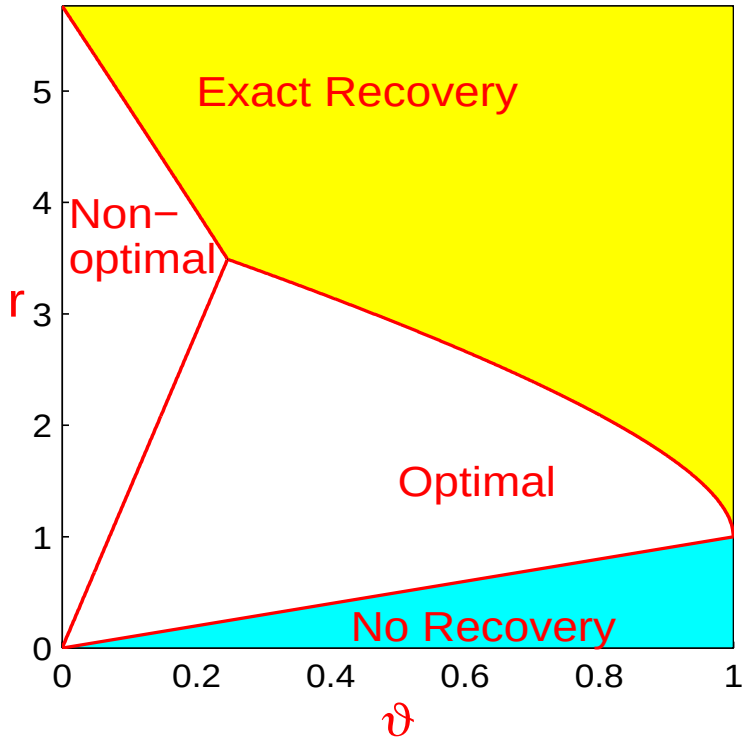


$\epsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r\log p}, \quad a = 0.4; \quad \lambda_p^{lasso} = \max\{\frac{r+\vartheta}{2r}, \ \frac{1}{1+\sqrt{\frac{1-a}{1+a}}}\}\tau_p;$

Two lines: $r = \left(\frac{1+\sqrt{1-a^2}}{a}\right)\vartheta$; $r = (1+\sqrt{\frac{1+a}{1-a}})^2(1-\vartheta)$

Non-optimal region (subset selection)





$\epsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r\log p}, \quad a = 0.4; \quad \lambda_p^{SS} = \min\{\frac{r+\vartheta}{2r}, \frac{r(1-a^2)+2\vartheta}{2r\sqrt{1-a^2}}\}\tau_p;$

Line: $r = \big((2-\sqrt{1-a^2})/[\sqrt{1-a^2}-(1-a^2)]\big)\vartheta;$

Curve: $\sqrt{r} = \frac{1}{v}\big(\sqrt{1-2\vartheta}+\sqrt{1-2\vartheta+\vartheta v}\big), \quad v \equiv 2\sqrt{1-a^2}-1$

Understanding lasso & subset selection

$$Y = X\beta + z \qquad \Longrightarrow \qquad \tilde{Y} \approx N(\Omega\beta, \Omega)$$

Major components:

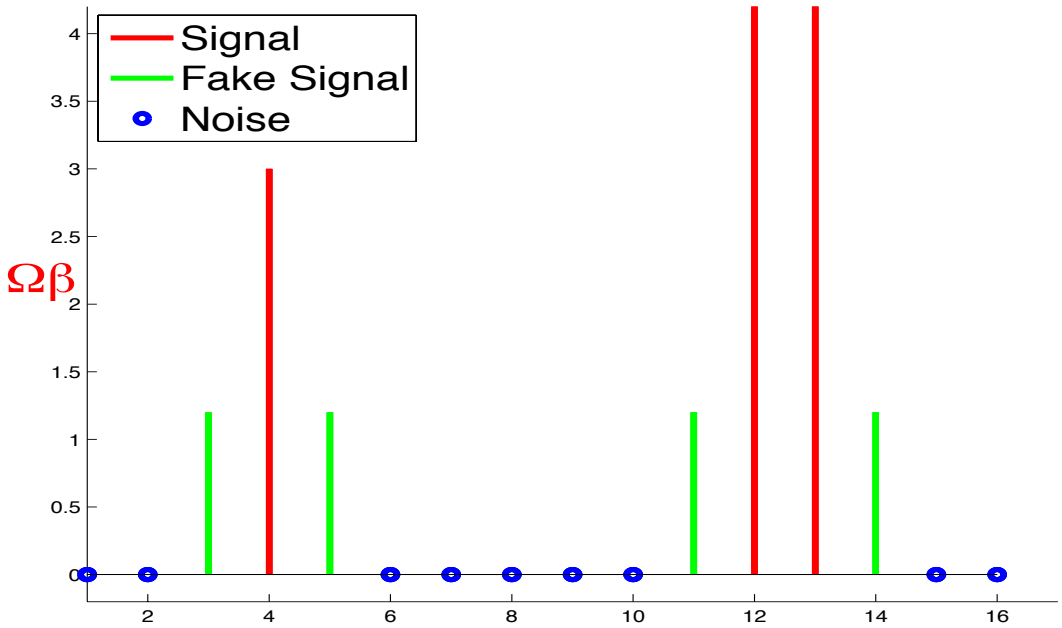
- Signal singleton: $\beta_j = \tau_p$
- Signal pairs: $\beta_j = \beta_{j+1} = \tau_p$
- Pure noise: $\beta_j = (\Omega\beta)_j = 0$
- Fake signals: $\beta_j = 0$, $(\Omega\beta)_j \neq 0$

Key Insight:

- The lasso is too **loose** on fake signals
- Subset selection is too **harsh** on signal pairs

How lasso/subset selection works

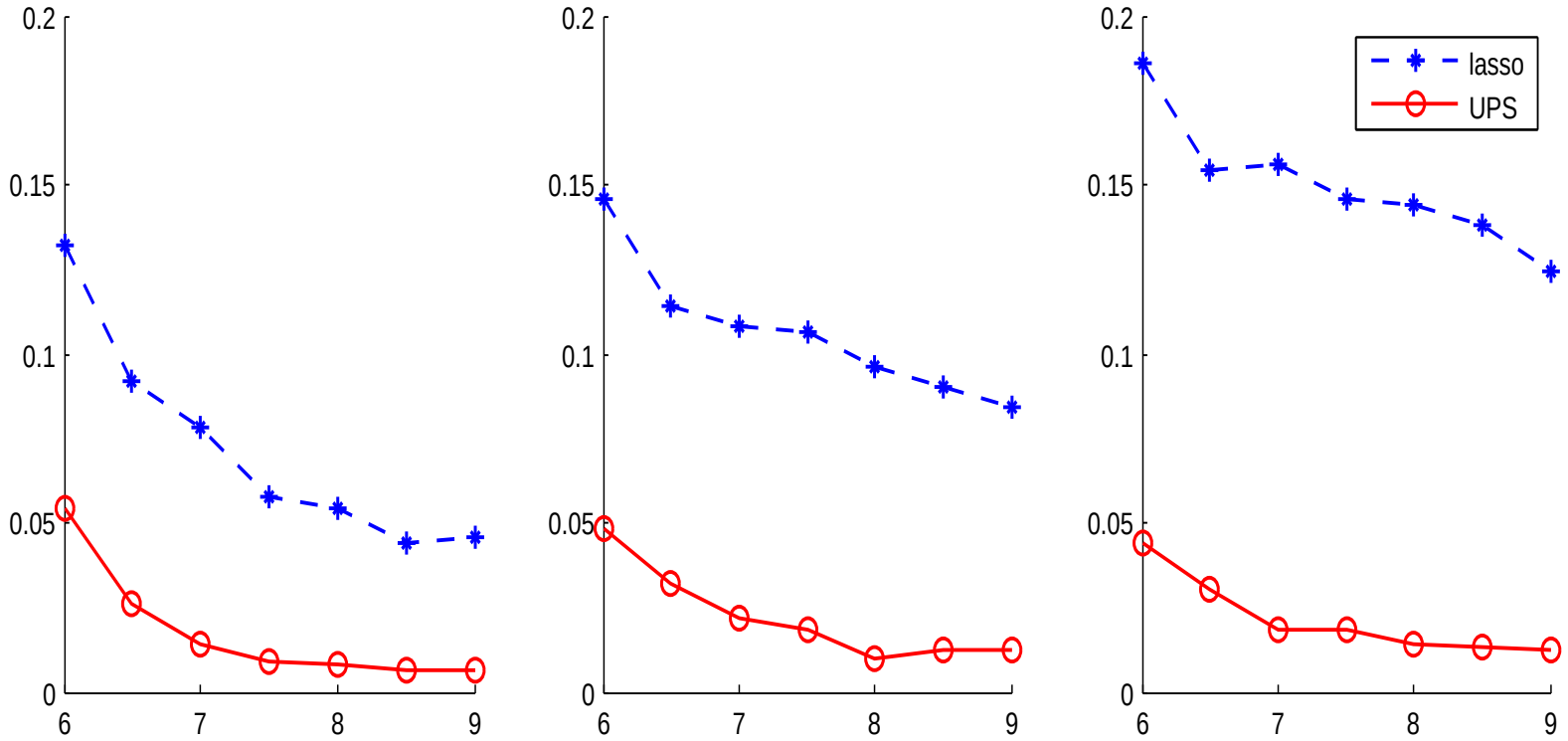
$$Y = X\beta + z \qquad \Longrightarrow \qquad \tilde{Y} \approx N(\Omega\beta, \Omega)$$



For properly set λ_p^{lasso}, many entries of β are estimated 0; remaining entries break into small clusters, where lasso applies independently

Simulation

Hamming error$/(p\epsilon_p)$



Left to right: $\vartheta = 0.2, 0.45, 0.6$. x-axis: τ_p.

$p = 10^4$, $n = 1600$, $\epsilon_p = p^{-\vartheta}$, $\pi_p = \nu_{\tau_p}$, Ω is tridiagonal ($a = 0.45$)

Simulation, II

Hamming errors:

ϑ	τ_p	5	6	7	8	9	10	11
.25	UPS	49	11.1	1.79	**0.26**	0.02	0	0
	lasso	186.7	99.35	58.26	38.53	25.97	18.18	12.94
.50	UPS	10.06	2.11	**0.37**	0.09	0	0	0
	lasso	16.36	5.11	1.47	0.51	0.28	0.33	0.26
.65	UPS	5.49	1.29	**0.33**	0.06	0	0	0
	lasso	7.97	2.43	0.69	0.18	0.07	0.03	.02

$p = 2000$, Ω tridiagonal ($a = 0.45$), $\epsilon_p = p^{-\vartheta}$, $\pi_p = \nu_{\tau_p}$.

Theoretic boundaries for exact recovery: $\tau_p \geq 8.01, 7.32, 7.01$

Take-home messages

- Suggested UPS as a two-stages procedure for variable selection
- Proved that UPS achieves the optimal phase diagram
- Identified non-optimal regions for
  - the lasso
  - the subset selection
- Exposed intuition for the non-optimality of penalization methods

Acknowledgment

Alphabetically:

Peter Bickel, Tony Cai, David Donoho, Jianqing Fan, Stephen Fienberg, Runze Li, Larry Wasserman, Cun-Hui Zhang for inspiration and discussion