Graphlet Screening (GS)

Jiashun Jin

Carnegie Mellon University

April 11, 2014

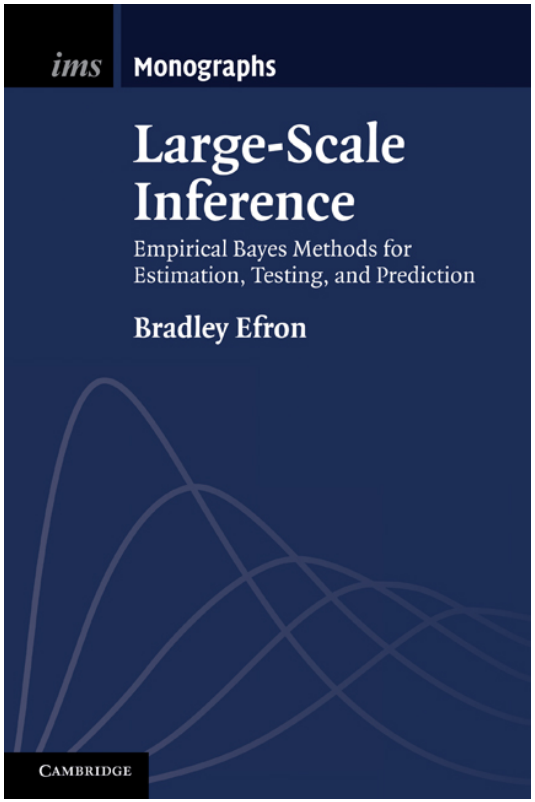
Collaborators

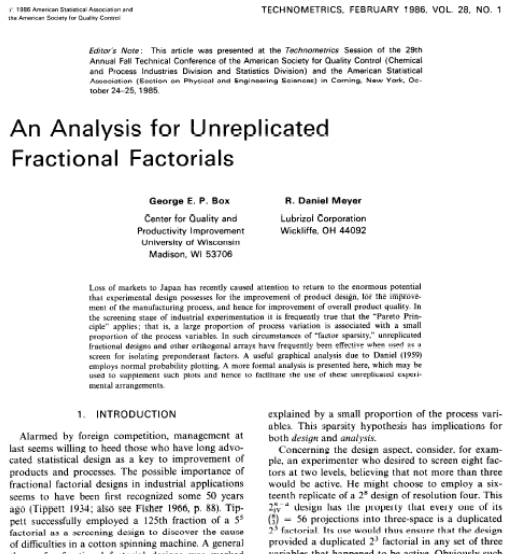
Alphabetically:

Zheng (Tracy) Ke	Princeton
Cun-Hui Zhang	Rutgers
Qi Zhang	University of Wisconsin

Rare/Weak signals in LSI

- **Rare**. Signals sparsely scatter across different observation units; no priori where there are
- **Weak**. Signals are individually weak (new)





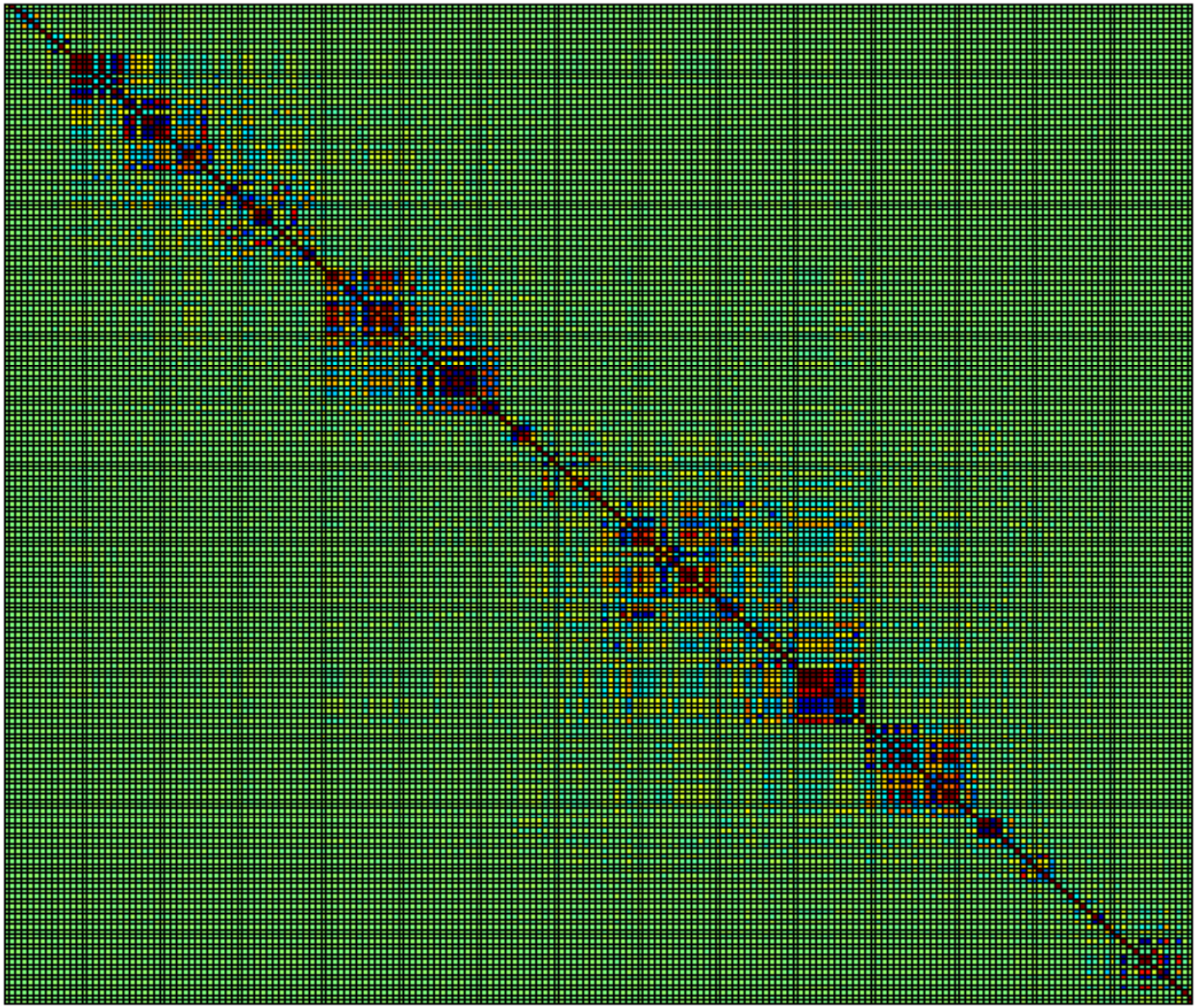
Box and Meyer (1986)

A linear model

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

- $p \gg n \gg 1$
- β is sparse: many coordinates are 0
- Gram matrix $G = X'X$
  - has unit diagonals
  - is sparse (few large coordinates in each row)

Linkage Disequilibrium (LD)



Idea

Univariate Screening

$$Y = X\beta + z, \qquad X = X_{n,p} = [x_1, x_2, \ldots, x_p]$$

For a threshold $t > 0$, apply Hard-Thresholding:

$$\hat{\beta}_j^{HT} = \begin{cases} (Y, x_j), & |(Y, x_j)| \geq t, \\ 0, & \text{otherwise} \end{cases}$$

Signal cancellation

Even if β_j is large, $E[(x_j, Y)]$ could be small

Denote the support of β by

$$S = S(\beta) = \{1 \le j \le p : \beta_j \neq 0\}$$

$$(Y, x_j) = \sum_{\ell=1}^{p} (x_\ell, x_j)\beta_\ell + (z, x_j)$$

$$= \beta_j + \sum_{\ell \in S(\beta)\setminus\{j\}} (x_j, x_\ell)\beta_\ell + N(0,1),$$

'signal cancellation' may happen as

$$x_j \quad \not\perp \quad \{x_\ell : \ell \in S(\beta) \setminus \{j\}\}$$

Exhaustive Multivariate Screening (EMS)

Fix $m_0 \geq 1$ and a small $\pi_0 > 0$

- For $m = 1, 2, \ldots, m_0$ and any subset

$$\mathcal{J} = \{j_1, j_2, \ldots, j_m\}, \qquad j_1 < j_2 < \ldots < j_m$$

Project Y onto $\{x_{j_1}, x_{j_2}, \ldots, x_{j_m}\}$:

$$Y \mapsto P^{\mathcal{J}} Y$$

- Retain the nodes in $\mathcal{J}$ if and only if

$$P\big(\chi^2_m(0) \geq \|P^{\mathcal{J}} Y\|^2\big) \quad \geq \quad \pi_0$$

- Fit the model with all retained nodes

Rationale

Hope: maybe for some small-size set $\mathcal{J}$,

$$\{x_j : j \in \mathcal{J}\} \quad \perp \quad \{x_\ell : \ell \in S(\beta) \setminus \mathcal{J}\};$$

$$Y = \sum_{j \in \mathcal{J}} \beta_j x_j + \sum_{j \in S(\beta) \setminus \mathcal{J}} \beta_j x_j + z$$

- Possible 'signal cancellation' if we look at any single projected coefficients (Y, x_j)
- No 'signal cancelation' if look at the projected coefficients $\{(Y, x_j) : j \in \mathcal{J}\}$ together

Problems

- Computationally infeasible:

$$\sum_{m=1}^{m_0} \binom{p}{m}$$

- Inefficiency: include too many candidates for screening; signals need to be stronger than necessary to survive the screening

Our proposal: Graphlet Screening (GS)

Graph of Strong Dependence (GOSD)

Define GOSD as the graph $\mathcal{G} = (V, E)$:

- $V = \{1, 2, \ldots, p\}$: each variable is a node
- Nodes i and j have an edge iff

$$\big|(x_i, x_j)\big| \geq \delta, \qquad (\delta = \tfrac{1}{\log(p)}, \text{ say})$$

- $G = X'X$ sparse $\Longrightarrow \mathcal{G}$ sparse

Graphlet Screening (GS)

$\mathcal{A}(m_0) = \{\text{All connected subgraphs of } \mathcal{G}\text{; size} \leq m_0\}$

GS: same as EMS, except for

- EMS exhaustively screens all

$$\mathcal{J} = \{j_1, j_2, \ldots, j_m\}, \qquad j_1 < j_2 < \ldots < j_m$$

- GS screens $\mathcal{J}$ if and only if $\mathcal{J} \in \mathcal{A}(m_0)$

Lemma. If d be the maximum degree of $\mathcal{G}$, then

$$|\mathcal{A}(m_0)| \leq Cp(ed)^{m_0}; \qquad e = 2.718$$

Mutual orthogonality

$\mathcal{G}$: *there is an edge between* i *and* $j \iff |(x_i, x_j)| \geq 1/\log(p)$

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

Restricting nodes to S forms a subgraph $\mathcal{G}_S$, and

$$\mathcal{G}_S = \mathcal{G}_{S,1} \cup \ldots \cup \mathcal{G}_{S,M}: \qquad \mathcal{G}_{S,\ell}\text{: components}$$

By how $\mathcal{G}$ is defined, approximately,

$$\{x_j : j \in \mathcal{G}_{S,1}\} \perp \{x_j : j \in \mathcal{G}_{S,2}\} \perp \ldots \perp \{x_j : j \in \mathcal{G}_{S,M}\}$$

Maximum component size $m_0^*(\beta, G, \delta)$

Surprise. $m_0^*(\beta, G, \delta)$ is small in many cases, where

$$m_0^*(\beta, G, \delta) = \max_{1 \le \ell \le M} |\mathcal{G}_{S,\ell}|$$

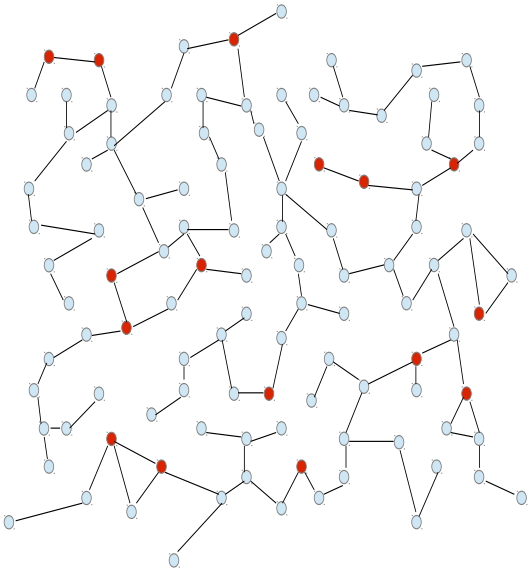
Lemma. If $I\{\beta_j \neq 0\} \overset{iid}{\sim} \text{Bernoulli}(\epsilon)$, then

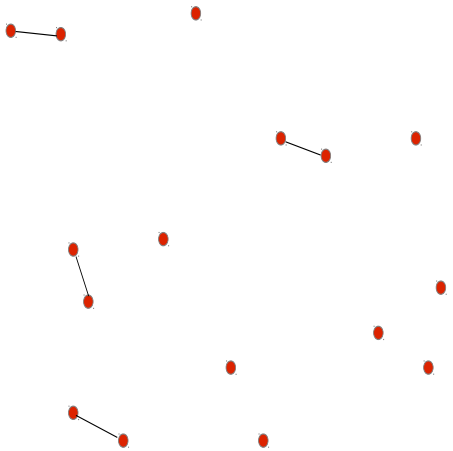
$$P\big(m_0^*(\beta, G, \delta) > m\big) \le p(ed\epsilon)^{m+1}, \qquad \forall\ m > 1,$$

where $d = d(\mathcal{G})$ is maximum degree of $\mathcal{G}$

Signal archipelago

- Signals split into many small-size disconnected islands
- Columns indexed by different islands: mutual orthogonal
- No 'signal cancellation' when we screen each island individually





Why GS works

$\mathcal{A}(m_0) = \{\text{All connected subgraphs of } \mathcal{G}\text{; size} \le m_0\}$,

$\mathcal{G}_S = \mathcal{G}_{S,1} \cup \mathcal{G}_{S,2} \ldots \cup \mathcal{G}_{S,M}$; maximum size $m_0^*(\beta, G, \delta)$

- GS screens a set $\mathcal{J} \Longleftrightarrow \mathcal{J} \in \mathcal{A}(m_0)$
- If $m_0 \ge m_0^*(\beta, G, \delta)$

$$\mathcal{G}_{S,\ell} \in \mathcal{A}(m_0): \qquad \text{on our screen list!}$$

- Approximately, no 'signal cancellation' for

$$\{x_j : \mathcal{G}_{S,\ell}\} \qquad \perp \qquad \{x_j : \mathcal{G}_S \setminus \mathcal{G}_{S,\ell}\}, \qquad \text{approx.}$$

$$Y = \sum_{j \in \mathcal{G}_{S,\ell}} \beta_j x_j + \sum_{j \in \mathcal{G}_S \setminus \mathcal{G}_{S,\ell}} \beta_j x_j + z$$

Summary

Method	US	EMS	GS
Computation	p	$\binom{p}{m_0}$	$Cp(ed)^{m_0}$
Efficiency	yes	no	yes
Robust to Signal Cancellation	no	yes	yes

Methods

Application I: rank features

For feature j, find all $\mathcal{J} \ni j$, calculate P-values; use $\pi_j =$ minimum of all such P-values for ranking

Example.

- Rows of $X = X_{n,p}$ are iid $N(0, \Omega)$
- Ω block-wise diagonal, with building blocks

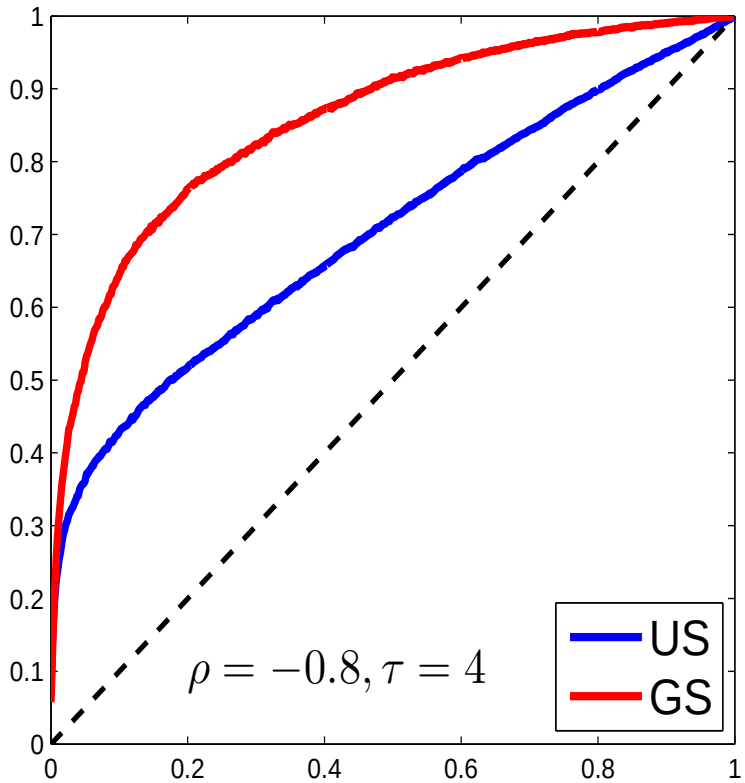
$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

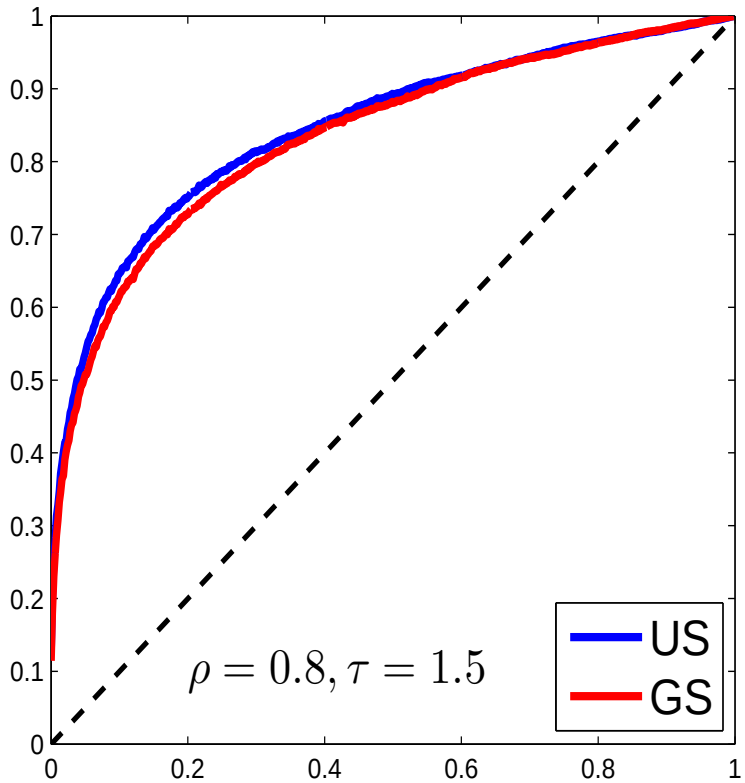
- β: partitioned to blocks of 2,

$$(\beta_{2k-1}, \beta_{2k}) = \begin{cases} (0,0), & \text{prob. } (1-\epsilon), \\ (0,\tau), & \text{prob. } \epsilon/4, \\ (\tau,0), & \text{prob. } \epsilon/4, \\ (\tau,\tau), & \text{prob. } \epsilon/2 \end{cases}$$

ROC of US vs ROC of GS ($m_0 = 2$)

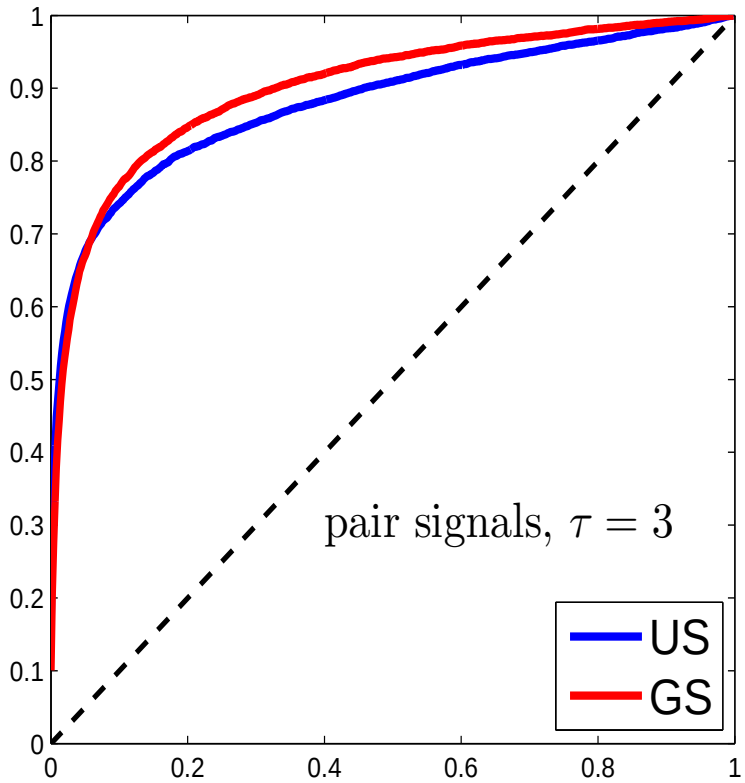
$p = 1000,\ n = 500,\ \epsilon = 0.05$

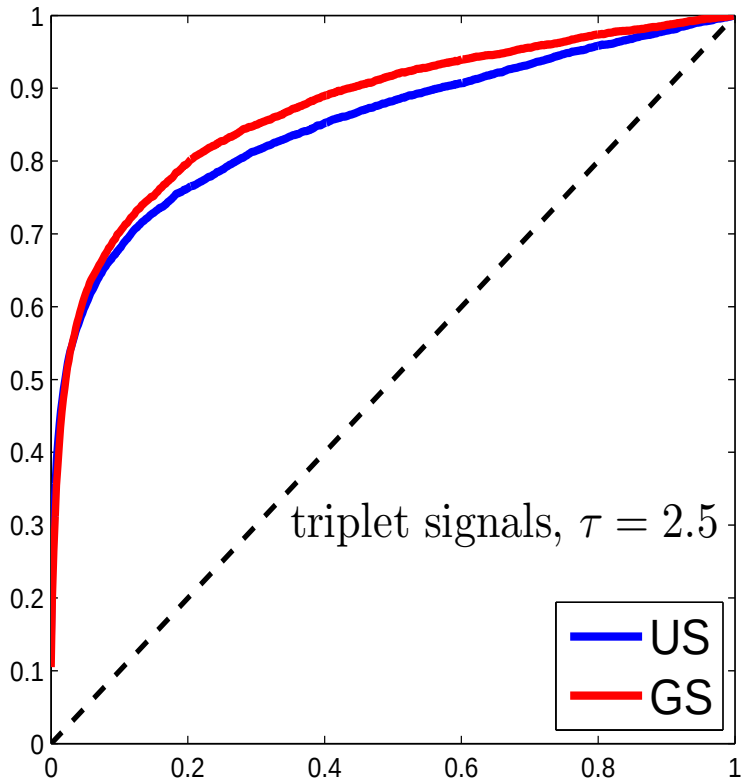




SNP data (chromosome 21)

GS: $m_0 = 2$; $p = 3937$, $n = 16179$. *Left:* $(\beta_{2k-1}, \beta_{2k}) = (\tau, \tau)$, *prob.* 0.01. *Right:* $(\beta_{3k-2}, \beta_{3k-1}, \beta_{3k}) = (\tau, \tau, \tau)$, *prob.* 0.01.

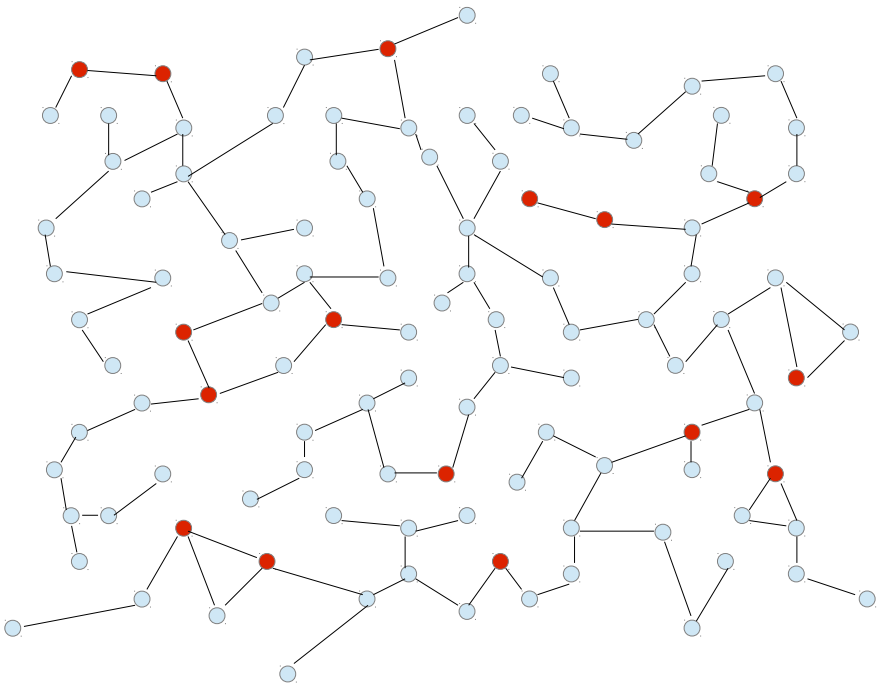


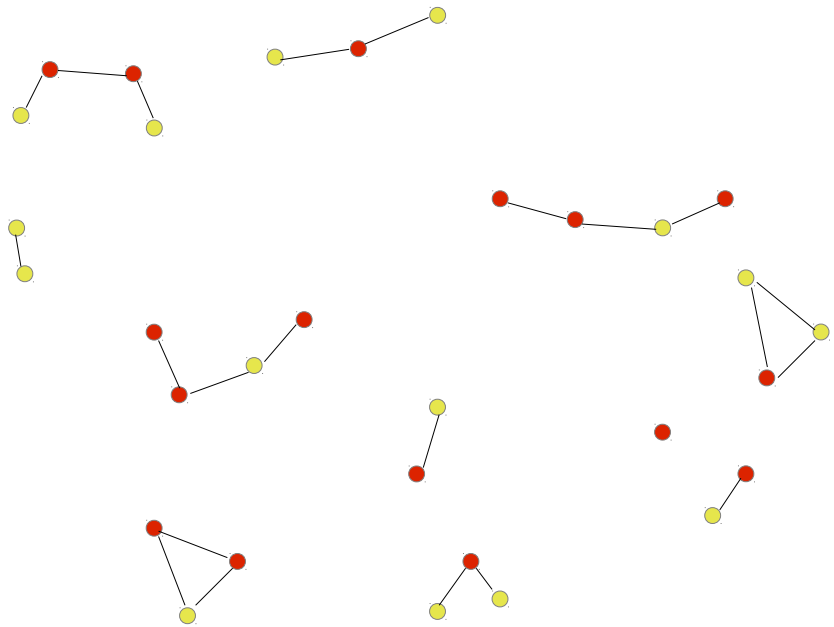


Application II. Variable Selection

A two-stage screen and clean procedure:

- Apply GS, with small modifications
- Clean





Screen step

- $\mathcal{A}(m_0) = \{\text{All connected subgraphs of } \mathcal{G} \text{ with size} \le m_0\}$
- Arrange by size, ties breaking lexicographically:

$$\mathcal{A}(m_0) = \{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_T\}$$

- Initializing with $\mathcal{S}_0 = \emptyset$
- For $t = 1, 2, \ldots, T$, letting $\mathcal{S}_{t-1}$ be the set of retained indices in stage $t-1$, update $\mathcal{S}_{t-1}$ by

$$\mathcal{S}_t = \begin{cases} \mathcal{S}_{t-1} \cup \mathcal{J}_t, & \text{if } \|P^{\mathcal{J}_t} Y\|^2 - \|P^{\mathcal{J}_t \cap \mathcal{S}_{t-1}} Y\|^2 \ge 2q \log(p), \\ \mathcal{S}_{t-1}, & \text{otherwise} \end{cases}$$

- $\hat{S} = \mathcal{S}_T$: all retained nodes

Clean step

Fixing tuning parameters (u^{gs}, v^{gs}),

- $j \notin \hat{S}$: set $\hat{\beta}_j^{gs} = 0$
- $j \in \hat{S}$: decompose

$$\mathcal{G}_{\hat{S}} = \mathcal{G}_{\hat{S},1} \cup \mathcal{G}_{\hat{S},2} \cup \ldots \cup \mathcal{G}_{\hat{S},\hat{M}},$$

and estimate $\{\beta_j : j \in \mathcal{G}_{\hat{S},\ell}\}$ by minimizing

$$\|P^{\mathcal{G}_{\hat{S},\ell}}(Y - \sum_{j \in \mathcal{G}_{\hat{S},\ell}} \beta_j X_j)\|^2 + (u^{gs})^2 \|\beta\|_0,$$

subject to either $\beta_j = 0$ or $|\beta_j| \geq v^{gs}$

Tuning parameters of GS

- Feature ranking: need (δ, m_0); choices of which can be guided by G and computation capacity
- Variable selection: also need (q, u^{gs}, v^{gs})
  - q: flexible and insensitive
  - u^{gs} is relatively easy to estimate
  - v^{gs} is relatively hard to estimate

Minimax Theory

Sparse model

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

$$\beta = b \circ \mu, \qquad b_i \overset{iid}{\sim} \text{Bernoulli}(\epsilon), \qquad \mu \in \mathcal{M}_p^*(\tau, a)$$

- $b \circ \mu \in \mathbb{R}^p$: $(b \circ \mu)_j = b_j \mu_j$
- $\mathcal{M}_p^*(\tau, a) = \{\mu \in \mathbb{R}^p : \tau \le |\mu_j| \le a \cdot \tau\}$
- As $p \to \infty$, link (ϵ, τ) to p by fixed parameters:

$$\epsilon = \epsilon_p = p^{-\vartheta}, \qquad \tau = \tau_p = \sqrt{2r \log(p)}$$

Sparse model, II. Random design

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

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

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

Minimax Hamming distance

Measuring errors with Hamming distance:

$$H_p(\hat{\beta}, \epsilon_p, \mu; \Omega) = E\left[\sum_{j=1}^{p} 1\{\mathrm{sgn}(\hat{\beta}_j) \neq \mathrm{sgn}(\beta_j)\}\right]$$

Minimax Hamming distance:

$$\mathrm{Hamm}_p^*(\vartheta, \theta, r, a, \Omega) = \inf_{\hat{\beta}} \sup_{\mu \in \mathcal{M}_p^*(\tau_p, a)} H_p(\hat{\beta}, \epsilon_p, \mu; \Omega)$$

Exponent $\rho_j^* = \rho_j^*(\vartheta, r, \Omega)$

Define $\omega = \omega(S_0, S_1; \Omega) = \inf_\delta\{\delta'\Omega\delta\}$ where

$$\delta \equiv u^{(0)} - u^{(1)}: \quad \begin{cases} u_i^{(k)} = 0, & i \notin S_k \\ 1 \le |u_i^{(k)}| \le a, & i \in S_k \end{cases}, \quad k = 0, 1$$

Define

$$\rho(S_0, S_1; \vartheta, r, a, \Omega) = \frac{|S_0| + |S_1|}{2}\vartheta + \frac{\omega r}{4} + \frac{(|S_1| - |S_0|)^2\vartheta^2}{4\omega r}$$

Minimax rate critically depends on the exponents:

$$\rho_j^* = \rho_j^*(\vartheta, r; \Omega) = \min_{(S_0, S_1): j \in S_0 \cup S_1} \rho(S_0, S_1, \vartheta, r, a, \Omega)$$

- not dependent on (θ, a) (mild regularity cond.)
- computable; has explicit form for some Ω

Asymptotic minimaxity of GS

- Assume $\sum_{j=1}^{p} |\Omega(i,j)|^{\gamma} \leq C, \quad \gamma \in (0,1), \quad 1 \leq i \leq p$
- Screen-step: q is a properly small number
- Clean-step: set $u^{gs} = \sqrt{2\vartheta \log p}$, and $v^{gs} = \tau_p$

Theorem GS achieves optimal rate of convergence:

$$\sup_{\mu \in \mathcal{M}_p^*(\tau_p, a)} H_p(\hat{\beta}^{gs}, \epsilon_p, \mu, \Omega) \leq L_p \left[\Big(\sum_{j=1}^{p} p^{-\rho_j^*} \Big) + p^{1-(m_0+1)\vartheta} \right]$$

$$\leq L_p \left[\mathrm{Hamm}_p^*(\vartheta, \theta, r, a, \Omega) + p^{1-(m_0+1)\vartheta} \right]$$

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

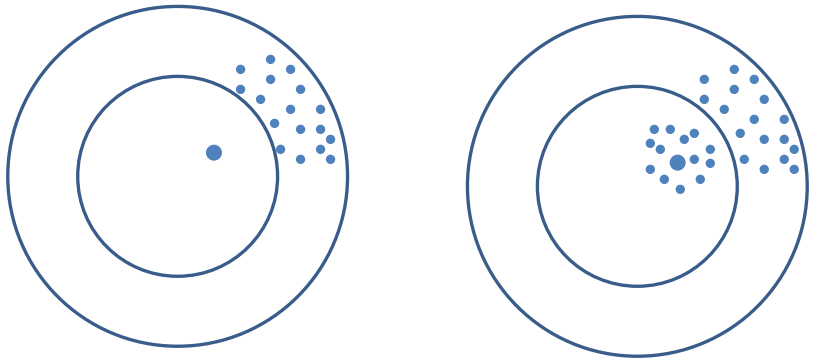
L^0-penalization method

Donoho and Starck (1989)

$$\|Y - X\beta\|^2 + \lambda\|\beta\|_0 : \qquad \lambda > 0\text{: tuning parameter}$$

Idea: Where there is no noise

- Infinite solutions to $Y = X\beta$
- But only one is very sparse
- Hope: the sparsest solution is the truth (Occam's Razor)



L^0/L^1-penalization methods

L^1-solution $\approx$ L^0-solution $\approx$ truth

Method	L^0/L^1-penalization	CASE
Regime	Rare/Strong	Rare/Weak
Loss	$I\{\mathrm{sgn}(\hat{\beta}) \neq \mathrm{sgn}(\beta)\}$	$\mathrm{Hamm}(\mathrm{sgn}(\hat{\beta}), \mathrm{sgn}(\beta))$ †
Optimality	Not	Yes
Motivation	Imaging/Engineering	Genetics/Genomics
Design	Controllable/Nice	Uncontrollable/Bad
Key idea	One-stage global method	Multi-stage local method

† *Hamm: Hamming distance*

Donoho and Stark (1989), Tibshiraini (1996), and many others

Phase Diagram

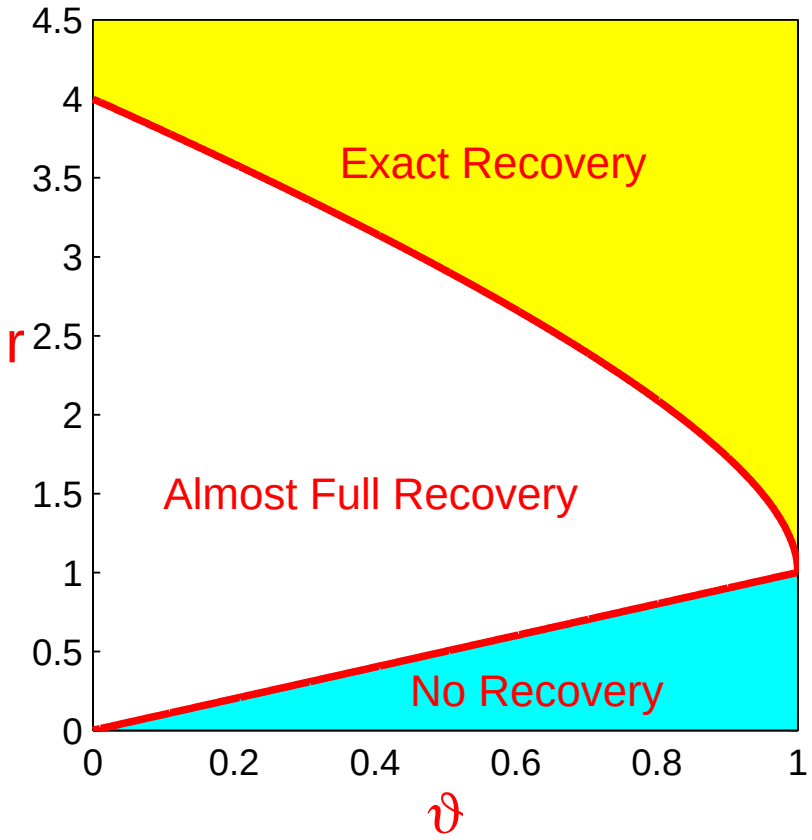
A Corollary

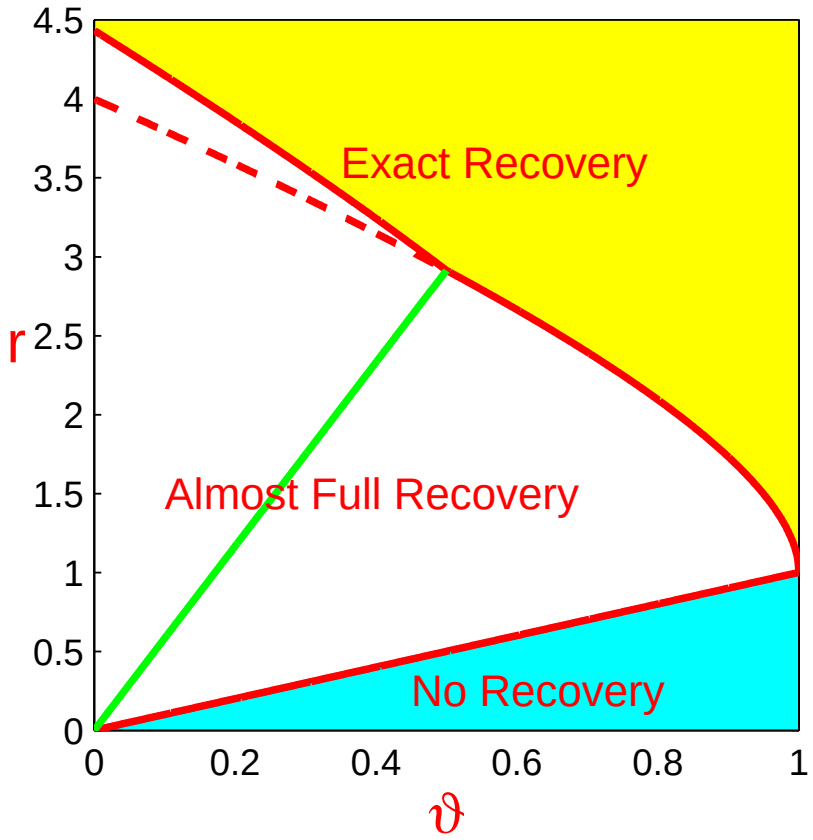
Suppose the conditions of the theorem hold. If additionally, $|\Omega(i,j)| \leq 4\sqrt{2}-5$, $\forall\, i \neq j$, then

$$\frac{\mathrm{Hamm}_p^*(\vartheta,\theta,r,a,\Omega)}{p\epsilon_p} = \begin{cases} 1+o(1), & r<\vartheta, \\ L_p p^{-\frac{(\vartheta-r)^2}{4r}}, & 1<\frac{r}{\vartheta}<3+2\sqrt{2} \end{cases}$$

Right hand side: rate when $\Omega = I_p$; $4\sqrt{2}-5 \approx 0.66$

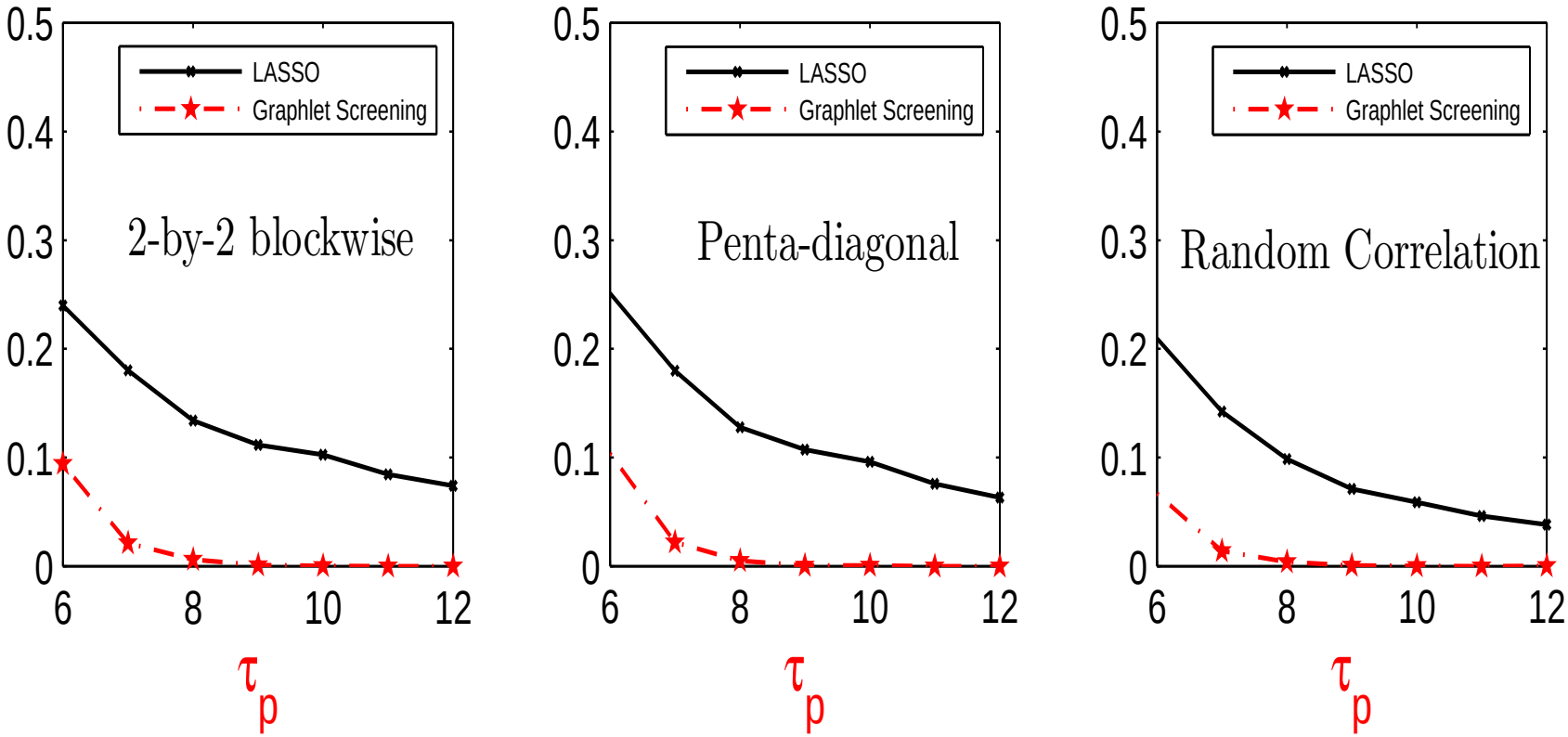
Phase Diagram





Left: $\Omega = I_p$; line $r = \vartheta$; curve $r = (1 + \sqrt{1 - \vartheta})^2$. Right: Ω as in the corollary; green line: $r/\vartheta = 3 + 2\sqrt{2}$

Simulation comparison



$p = 5000$, $n = 4000$, $p\epsilon_p = 250$; $\tau_p = 6, 7, \ldots, 12$. Left to right: G is block-wise, penta-diagonal, randomly generated ('sprandsym' in matlab).

Take-home messages

- GS is a computationally feasible approach that overcomes the challenge of 'signal cancellation'
- Key insight:
  - $\mathcal{G}_S$ splits into different small-size signal islands $\mathcal{G}_{S,\ell}$, and $\{x_j : j \in \mathcal{G}_{S,\ell}\}$ are mutual orthogonal (approx).
  - minimax rate depends on X 'locally' so we have to act 'locally'
- Optimal in variable selection, while penalization methods are not
- A flexible idea and that is useful in many different situations

Main References

Genovese C, Jin J, Wasserman L, and Yao Z (2012) A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.*, **13** 2107-2143.

Ji P and Jin J (2012) UPS delivers optimal phase diagram in high-dimensional variable selection *Ann. Statist*. **40**(1), 73-103.

Jin J, Zhang C-H and Zhang Q (2012) Optimality of Graphlet Screening in High Dimensional Variable Selection. *arXiv:1204.6452*

Ke T, Jin J and Fan J (2012) Covariance assisted screening and estimation. *arXiv:1205.4645*.