Graphlet Screening (GS)

Jiashun Jin

Carnegie Mellon University

April 11, 2014

同 ト く ヨ ト く ヨ ト

Alphabetically:

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

æ

白 ト イヨト イヨト

Rare/Weak signals in LSI

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



/ 1988 Analose Statement Association and the Analose Technika for Factor Factoria TECHNOMETRICS, FEBRUARY 1986, VOL. 28, NO. 1

Editor's Actor: This stricts was presented at the Technomenics Season of the 28th Annuel Fall Technical Conference of the American Society for Duality Central (Cheminal and Process Understein Schedung and Season Society) and the American Statistican Association (Eastern on Physical and Expiration) and the American Statistican Association (Eastern on Physical and Expiration) Economy Jan Tech. On Joury 18:455–4555.

An Analysis for Unreplicated Fractional Factorials

George E. P. Bax Center for Quality and Productivity Improvement University of Wassensin Madison, WI 53208

ality and Lubricel Cosposition sprovement Wickliffs, DH 44092 Wisconin # 82208

Loss of markets to support has reasoning several advances to yours to the concentre probability and other manufacturing presess, and become for the probability of the probability and other manufacturing presess, and become for the present of network probability of the the concentre areas of conference presents on the theory for the factor of the properties of the presess withhele 10 band downstreams of "hanne present" emergence presents of the presess withhele 10 band downstreams of "hanne present" emergence presents of the presess withhele 10 band downstreams of "hanne present" emergence presents of the presess withhele 10 band downstreams of "hanne present" emergence presents of the presess withhele 10 band downstreams of the theory of the presents of the presess of the present to the theory and present and the the theory of the present of the present of the theory of the present of the theory of the presents of the present of the theory of the present of the theory of the present of the present of the theory of the present of the theory of the theory of the present of the present of the theory of the theory of the theory of the theory of the present of the theory of the present of the theory of the theory of the theory of the theory of the present of the theory of theory of the theory of the th

1. INTRODUCTION

Altered by foreign competition, management at last seems within to be ded how who have long aftercoald matrixed design as a key to improvement of produces and presents. The probabilit importance of functional factorial designs in industrial applications seems to have been first recognisit sense 50 parts ago (Tepper) 1932; also see Fisher 1966; p. 80, Thpper) severability originated a 125 function of a 32 functional as a scenario design to discover the neuro of difficulties to endow speciality matchine. A guarant explained by a small proportion of the process variables. This sparsity hypothesis has implications for both design and analysis. Concerning the design aspect, consider, for exam-

Concerning the design index (consider, for their picture of the design index (considered on the design of the constant sign fractions are not been design of the sign of the constant of the design of the constant on the design of the desi

イロト イヨト イヨト イヨト

Box and Meyer (1986)

2

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

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

個 と く ヨ と く ヨ と …

Linkage Disequilibrium (LD)



Jiashun Jin Grap

Idea

æ

・ロン ・回 と ・ ヨン ・ モン

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

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

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

$$egin{aligned} &(Y,x_j) = \sum_{\ell=1}^p (x_\ell,x_j)eta_\ell + (z,x_j) \ &= eta_j + \sum_{\ell\in \mathcal{S}(eta)ackslash \{j\}} (x_j,x_\ell)eta_\ell + \mathcal{N}(0,1), \end{aligned}$$

'signal cancellation' may happen as

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

Exhaustive Multivariate Screening (EMS)

Fix $m_0 \ge 1$ and a small $\pi_0 > 0$ • For $m = 1, 2, ..., m_0$ and any subset $\mathcal{J} = \{j_1, j_2, ..., j_m\}, \qquad j_1 < j_2 < ... < j_m$ Project Y onto $\{x_{j_1}, x_{j_2}, ..., x_{j_m}\}$: $Y \mapsto P^{\mathcal{J}}Y$

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

$$P(\chi_m^2(0) \ge \|P^{\mathcal{J}}Y\|^2) \ge \pi_0$$

Fit the model with all retained nodes

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

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

- 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 ∈ J} together

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 $\mathcal{G} = (V, E)$:

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

$$|(x_i, x_j)| \ge \delta, \qquad (\delta = \frac{1}{\log(p)}, \text{ say})$$

•
$$G = X'X$$
 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

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

Mutual orthogonality

G: there is an edge between i and $j \iff |(x_i, x_j)| \ge 1/\log(p)$ $S = S(\beta) = \{1 \le i \le p : \beta_i \ne 0\}$

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

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

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

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

- ◆ □ ▶ ◆ 三 ▶ ◆ □ ● ● ○ ○ ○ ○

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

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

Lemma. If $I\{\beta_j \neq 0\} \stackrel{iid}{\sim} \text{Bernoulli}(\epsilon)$, then $P(m_0^*(\beta, G, \delta) > m) \leq p(ed\epsilon)^{m+1}, \quad \forall m > 1,$ 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

 $\begin{aligned} \mathcal{A}(m_0) &= \{ \text{All connected subgraphs of } \mathcal{G}; \text{ size } \leq m_0 \}, \\ \mathcal{G}_S &= \mathcal{G}_{S,1} \cup \mathcal{G}_{S,2} \ldots \cup \mathcal{G}_{S,M}; \text{ maximum size } m_0^*(\beta, G, \delta) \end{aligned}$

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

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

Approximately, no 'signal cancellation' for

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

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



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

・ロ・・ (日・・ (日・・ (日・)

æ

Methods

æ

・ロン ・回 と ・ ヨン ・ ヨン

Application I: rank features

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

Example.

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

$$\left(\begin{array}{cc} 1 & \rho \\ \rho & 1 \end{array}\right)$$

• 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 ($m_0 = 2$)

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



Jiashun Jin Graphlet Screening (GS)

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.



Jiashun Jin Graphlet Screening (GS)

Application II. Variable Selection

A two-stage screen and clean procedure:

- Apply GS, with small modifications
- Clean



Screen step

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

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

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

$$\mathcal{S}_t = \left\{ egin{array}{ll} \mathcal{S}_{t-1} \cup \mathcal{J}_t, & ext{if } \| \mathcal{P}^{\mathcal{J}_t} Y \|^2 - \| \mathcal{P}^{\mathcal{J}_t \cap \mathcal{S}_{t-1}} Y \|^2 \geq 2q \log(p), \ \mathcal{S}_{t-1}, & ext{otherwise} \end{array}
ight.$$

• $\hat{S} = S_T$: 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

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

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

$$\|\mathcal{P}^{\mathcal{G}_{\mathfrak{S},\ell}}(Y-\sum_{j\in\mathcal{G}_{\mathfrak{S},\ell}}eta_jx_j)\|^2+(u^{gs})^2\|eta\|_0,$$

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

- ► Feature ranking: need (δ, m₀); 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

æ

回 と く ヨ と く ヨ と

Sparse model

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

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

æ

・ロン ・雪 ・ ・ ヨ ・ ・ ヨ ・ ・

Sparse model, II. Random design

$$Y = X\beta + z, \qquad X = \begin{pmatrix} X'_1 \\ \dots \\ X'_n \end{pmatrix}, \qquad X_i \stackrel{iid}{\sim} N(0, \frac{1}{n}\Omega)$$

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

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

æ

▲圖▶ ▲屋▶ ▲屋≯

Minimax Hamming distance

Measuring errors with Hamming distance:

$$H_{p}(\hat{\beta}, \epsilon_{p}, \mu; \Omega) = E\left[\sum_{j=1}^{p} \mathbb{1}\left\{\operatorname{sgn}(\hat{\beta}_{j}) \neq \operatorname{sgn}(\beta_{j})\right\}\right]$$

Minimax Hamming distance:

$$\operatorname{Hamm}_{p}^{*}(\vartheta, \theta, r, a, \Omega) = \inf_{\hat{\beta}} \sup_{\mu \in \mathcal{M}_{p}^{*}(\tau_{p}, a)} H_{p}(\hat{\beta}, \epsilon_{p}, \mu; \Omega)$$

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

Define
$$\omega = \omega(S_0, S_1; \Omega) = \inf_{\delta} \{ \delta' \Omega \delta \}$$
 where
 $\delta \equiv u^{(0)} - u^{(1)} : \begin{cases} u_i^{(k)} = 0, & i \notin S_k \\ 1 \le |u_i^{(k)}| \le a, & i \in S_k \end{cases}, \quad k = 0, 1$

Define

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

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 $\sum_{j=1}^{p} |\Omega(i,j)|^{\gamma} \leq C$, $\gamma \in (0,1)$, $1 \leq i \leq p$
- 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:

$$\sup_{\mu \in \mathcal{M}_{p}^{*}(\tau_{p}, a)} H_{p}(\hat{\beta}^{gs}, \epsilon_{p}, \mu, \Omega) \leq L_{p} \left[\left(\sum_{j=1}^{p} p^{-\rho_{j}^{*}} \right) + p^{1-(m_{0}+1)\vartheta} \right]$$
$$\leq L_{p} \left[\operatorname{Hamm}_{p}^{*}(\vartheta, \theta, r, a, \Omega) + p^{1-(m_{0}+1)\vartheta} \right]$$

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

Donoho and Starck (1989)

 $\|Y - Xeta\|^2 + \lambda \|eta\|_0$: $\lambda > 0$: 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	$I\{\operatorname{sgn}(\hat{eta}) \neq \operatorname{sgn}(eta)\}$	$Hamm(\mathrm{sgn}(\hat{\beta}),\mathrm{sgn}(\beta))$ †
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

▶ < 문 ▶ < 문 ▶</p>

Phase Diagram

æ

(4回) (4回) (4回)

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

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

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: $r/\vartheta = 3 + 2\sqrt{2}$

æ

< ≣⇒

Simulation comparison



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

<= ≣⇒

- GS is a computationally feasible approach that overcomes the challenge of 'signal cancellation'
- Key insight:
 - G_S splits into different small-size signal islands G_{S,ℓ}, and {x_j : j ∈ G_{S,ℓ}} 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

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*.

白 ト イヨト イヨト