Graphlet Screening (GS)
Achieves Optimal Rate in Variable Selection

Jiashun Jin
Carnegie Mellon University

Collaborated with Cun-Hui Zhang (Rutgers)
Qi Zhang (Univ. of Pittsburgh)
Variable selection

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

- \( p \gg n \gg 1 \)
- signals are \textbf{rare and weak}
- let \( G = X'X \) be the Gram matrix
  - diagonals of \( G \) are normalized to 1
  - \( G \) is \textbf{sparse} (few large entries each row)
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

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
  - \textbf{in the noiseless case}, if signals sufficiently sparse, equivalent to \( L^0 \)-penalization

\textit{Chen et al. (1998); Tibshirani (1996); Donoho (2006)}
Ex. $Y = X\beta + z$, $z \sim N(0, I_n)$, $\beta_j$ take values from $\{0, \tau\}$ and

$$G = X'X = \begin{pmatrix} D & 0 & \ldots & 0 \\ 0 & D & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & D \end{pmatrix}, \quad D = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix}$$

$\{1, 2, \ldots p\}$ partitions into 3 types of $2 \times 2$ blocks:

- I. No signal
- II. One signal
- III. Two signals
Limitation of $L^0$ penalization, II

- one-stage method
- one tuning parameter
- does not exploit ‘local’ graphical structure

Therefore, many penalization methods (e.g. lasso, SCAD, MC+, Dantzig selector) are non-optimal, as $L^0$-penalization is the ‘idol’ these methods mimic

‘local’: neighboring nodes in geodesic distance of a graph (TBD)
Where are the signals?


John Wilder Tukey (1915-2000)
GOSD is the graph $\mathcal{G} = (V, E)$:
- $V = \{1, 2, \ldots, p\}$: each variable is a node
- An edge between nodes $i$ and $j$ iff
  \[ |G(i, j)| \geq \frac{1}{\log(p)}, \text{ say} \]
- $G = X'X$ sparse $\implies \mathcal{G}$ sparse
Despite its sparsity, $\mathcal{G}$ is usually complicated.

Denote the support of $\beta$ by $S = S(\beta) = \{1 \leq i \leq p, \beta_i \neq 0\}$

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

**Key insight:** $\mathcal{G}_S$ decomposes into many small-size components that are disconnected to each other.

*Component*: a maximal connected subgraph
Graphlet Screening (GS):

- **gs**-step: graphlet screening by sequential $\chi^2$-tests
- **gc**-step: graphlet cleaning by Penalized MLE
- **Focus**: rare and weak signals
Graphlet screening (gs-step), Initial stage

\[ Y = X\beta + z, \quad X = X_{n,p}, \quad z \sim N(0, I_n); \quad G : GOSD \]

- Fix \( m \geq 1 \) (small)
- Let \( \{G_t : 1 \leq t \leq T\} \) be all connected subgraphs of \( G \) with size \( \leq m \)
- arranged by size, ties breaking lexicographically:

\[ p = 10, \quad m = 3, \quad T = 30; \]

\[ \{G_t, 1 \leq t \leq T\}: \]

\[ \{1\}, \{2\}, \ldots, \{10\} \]
\[ \{1, 2\}, \{1, 7\}, \ldots, \{9, 10\} \]
\[ \{1, 2, 4\}, \{1, 2, 7\}, \ldots, \{8, 9, 10\} \]
gs-step, II. Updating stage

\[ X = [x_1, x_2, \ldots, x_p], \quad \{G_t\}_{t=1}^T: \text{ all connected subgraphs with size } \leq m \]

For \( t = 1, 2, \ldots, T \)

- \( S_{t-1} \): set of retained indices in last stage
- Define \( T(Y; D, F) = \| P^{G_t} Y \|^2 - \| P^F Y \|^2 \)
  - \( F = G_t \cap S_{t-1} \): nodes accepted previously
  - \( D = G_t \setminus F \): nodes currently under investigation
  - \( P^F \): projection from \( R^n \) to subspace \( \{x_j : j \in F\} \)

- Adding nodes in \( D \) to \( S_{t-1} \) iff
  \[ T(Y; D, F) > t(D, F), \quad t(D, F): \text{ threshold TBD} \]

Once accepted, a node is kept until the end of gs-step
Comparison with marginal regression (computational complexity)

- Marginal screening
  - ineffective (neglects ‘local’ graphical structure)
  - ‘brute-forth’ $m$-variate screening is computationally challenging: $O(p^m)$

- $gs$-step
  - only screens connected subgraphs of $G$
  - if maximum degree of $G \leq K$, then there are $\leq C(eK)^m p$ such subgraphs

Two important properties of gs-step

\[ S^* \equiv S_T: \text{ set of survived nodes in the end of gs-step} \]

If both signals and Graph \( \mathcal{G} \) are sparse:

- **Sure Screening (SS):** \( S^* \) retains all *but a small proportion* of signals
- **Separable After Screening (SAS):** \( S^* \) decomposes into many small-size components
Reduce to many small-size regression, I

\[ G = X'X, \quad \mathcal{I}_0 \subset S^* : \text{a component} \]

\[ G_{\mathcal{I}_0} : \text{row restriction;} \quad G_{\mathcal{I}_0, \mathcal{I}_0} : \text{row \& column restriction} \]

- **Restrict regression to** \( \mathcal{I}_0 \)

\[ Y = X\beta + z \implies X'Y = X'X\beta + X'z \]
\[ \implies (X'Y)_{\mathcal{I}_0} = (G\beta)_{\mathcal{I}_0} + (X'z)_{\mathcal{I}_0} \]

- \((X'z)_{\mathcal{I}_0} \sim N(0, G_{\mathcal{I}_0, \mathcal{I}_0})\) since \( z \sim N(0, I_n) \)

- **Key:** \((G\beta)_{\mathcal{I}_0} \approx G_{\mathcal{I}_0, \mathcal{I}_0} \beta_{\mathcal{I}_0} \)

- **Result:** many small-size regression:

\[ (X'Y)_{\mathcal{I}_0} \approx N(G_{\mathcal{I}_0, \mathcal{I}_0} \beta_{\mathcal{I}_0}, G_{\mathcal{I}_0, \mathcal{I}_0}) \]
Reduce to small-size regression, II

Why \((G\beta)^{I_0} \equiv G^{I_0} \beta \approx G^{I_0,J_0} \beta^{I_0}\)?

\[
G^{I_0}\beta = \begin{bmatrix}
G^{I_0,I_0} & \mathbf{0} & \mathbf{0} & \cdots
\end{bmatrix}
\begin{bmatrix}
\beta^{I_0} \\
0 \\
\beta^{J_0} \\
\cdots
\end{bmatrix}
\]

- \(I_0, J_0 \subseteq S^*\): components
- By SS property, \(\beta^{\mathbf{0}} = 0\)
- By SAS property, \(G^{I_0,J_0} \approx 0\)
Graphlet cleaning (gc-step)

\[ Y = X\beta + z, \quad z \sim N(0, I_n) \]

- \( \mathcal{I}_0 \): a component of \( S^* \); \( S^* \): set of all survived nodes
- \( \beta^{\mathcal{I}_0} \): restricting rows of \( \beta \) to \( \mathcal{I}_0 \)
- \( X^{*,\mathcal{I}_0} \): restricting columns of \( X \) to \( \mathcal{I}_0 \)

Fixing \( (u^{gs}, v^{gs}) \),

- \( j \notin S^* \): set \( \hat{\beta}_j = 0 \)
- \( j \in S^* \): estimate \( \beta^{\mathcal{I}_0} \) via minimizing

\[
\| P^{\mathcal{I}_0}(Y - X^{*,\mathcal{I}_0}\theta) \|^2 + (u^{gs})^2 \| \theta \|_0,
\]

where an entry of \( \theta \) is 0 or \( \geq v^{gs} \) in magnitude
Random design model

\[ Y = X\beta + z, \quad X = \begin{pmatrix} X_1' \\ \vdots \\ X_n' \end{pmatrix}, \quad X_i \sim_{iid} N(0, \frac{1}{n}\Omega) \]

- \(\Omega\): unknown correlation matrix
- Ex: Compressive Sensing, Computer Security

Dinur and Nissim (2004), Nowak et al. (2007)
Rare and Weak signal model

\[ Y = X\beta + z, \quad z \sim N(0, I_n) \]

\[ \beta = b \odot \mu, \quad b_i \sim \text{Bernoulli}(\epsilon), \quad \mu \in \Theta^*_p(\tau, a) \]

- \( b \odot \mu \in \mathbb{R}^p: (b \odot \mu)_j = b_j \mu_j \)
- \( \Theta^*_p(\tau, a) = \{ \mu \in \mathbb{R}^p : \tau \leq |\mu_j| \leq a\tau \}, \ a > 1 \)
- Two key parameters:
  - \( \epsilon \): sparsity
  - \( \tau \): (minimum) signal strength
Asymptotic framework

Use $p$ as driving asymptotic parameter, and tie $(\epsilon, \tau, n)$ to $p$ by fixed parameters

- **Signal rarity:**
  \[ \epsilon = \epsilon_p = p^{-\vartheta}, \quad 0 < \vartheta < 1 \]

- **Signal weakness:**
  \[ \tau = \tau_p = \sqrt{2r \log(p)}, \quad r > 0 \]

- **Sample size:**
  \[ n = p^\theta, \quad (1 - \vartheta) < \theta < 1, \]

so that $p\epsilon_p \ll n_p \ll p$
Limitation of ‘Oracle Property’

Oracle property or probability of exact support recovery is a widely used criterion for assessing optimality in variable selection.

However, when signals are rare and weak, it is usually impossible to have exact recovery.
Measuring errors with Hamming distance:

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

Minimax Hamming distance:

\[ \text{Hamm}_p^*(\vartheta, \theta, r, a, \Omega) = \inf_{\hat{\beta}} \sup_{\mu \in \Theta_p^*(\tau_p, a)} H_p(\hat{\beta}, \epsilon_p, \mu; \Omega) \]
Exponent $\rho_j^* = \rho_j^*(\vartheta, r, \Omega)$

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

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

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 $(\theta, a)$ (mild regularity cond.)
- computable; has explicit form for some $\Omega$
Define sets of least favorable configuration at site $j$

$$(S_{0j}^*, S_{1j}^*) = \arg\max\{(S_0, S_1): j \in S_0 \cup S_1\} \left\{ \rho(S_0, S_1; \vartheta, r, a, \Omega) \right\}$$

**Definition.** GOLF is the graph $\mathcal{G}^\Diamond = (V, E)$ where $V = \{1, 2, \ldots, p\}$ and there is an edge between $i$ and $j$ if and only if $(S_{0j}^* \cup S_{1j}^*) \cap (S_{0k}^* \cup S_{1k}^*) \neq \emptyset$
Lower bound

\[ \beta = b \circ \mu, \quad b_j \overset{iid}{\sim} \text{Bernoulli}(\epsilon_p), \quad \mu \in \Theta^*_p(\tau_p, a) \]

\[ \epsilon_p = p^{-\vartheta}, \quad \tau_p = \sqrt{2r \log(p)} \]

**Theorem 1.** Let \( d(G^\odot) \) be the maximum degree of GOLF. As \( p \to \infty \),

\[ \text{Hamm}^*_p(\vartheta, \theta, r, a, \Omega) \geq \frac{L_p \sum_{j=1}^{p} p^{-\rho_j^*}}{d_p(G^\odot)} \]

where \( L_p \) is a generic multi-log\((p)\) term.
Main result: GS is asymptotic minimax

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

- **gs-step:** set thresholds at $\sqrt{2q\rho_j^* \log p}$, $0 < q < 1$

- **gc-step:** set $u^{gs} = \sqrt{2\vartheta \log p}$, and $v^{gs} = \tau_p$

**Theorem 2.** As $p \to \infty$,

- Both SS and SAS property hold
- Maximum degree of GOLF $\leq L_p$
- GS achieves optimal rate of convergence:

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

where $L_p$ is a generic multi-log$(p)$ term
GS uses tuning parameters \((\delta, m, u^{gs}, v^{gs})\) and \(Q = \{t(D, F) : D \text{ and } F \text{ as in } gs\text{-step}\}\)

- \((\delta, m)\): flexible (e.g. \(\delta = 1/\log(p), m = 3\))
- \(Q\): only need to be in a certain range

\[
t(D, F) = 2q \log(p), \quad q_0 \leq q \leq q^*(D, F)
\]

- \(u^{gs}\) is relatively easy to estimate
- \(v^{gs}\) is relatively hard to estimate
Example: $\rho_j^*(\vartheta, r, \Omega)$ has simple form

If $\lambda_3^*(\Omega) > 2(5 - 2\sqrt{6})$, $\lambda_4^*(\Omega) > 5 - 2\sqrt{6}$,

$$19 - 8\sqrt{6} < \Omega(i, j) < \frac{\sqrt{1 + \sqrt{6} - \sqrt{2}}}{\sqrt{3/2 + 1}}, \quad \forall i \neq j$$

$$5 - 2\sqrt{6} \approx 0.1, \ 19 - 8\sqrt{6} \approx -0.6, \ \sqrt{1 + \sqrt{6} - \sqrt{2}/(\sqrt{3/2 + 1})} \approx 0.64$$

**Corollary 1.** As $p \to \infty$,

$$\frac{\text{Hamm}_p^*(\vartheta, \theta, r, a, \Omega)}{p \epsilon_p} = \begin{cases} 1 + o(1), & r < \vartheta, \\ L_p p^{-\frac{(\vartheta-r)^2}{4r}}, & 1 < \frac{r}{\vartheta} < 5 + 2\sqrt{6} \end{cases}$$
A three-phase diagram in the phase space

\[ \{(\vartheta, r) : \ 0 < \vartheta < 1, \ r > 0\} \]

to visualize the behavior of a procedure

- I. Region of No Recovery
- II. Region of Almost Full Recovery:
- III. Region of Exact Recovery
Phase diagram of GS (Corollary 1)

Left: $\Omega = I_p$; red curve: $r = (1 + \sqrt{1 - \vartheta})^2$

Right: $\Omega$ as in Corollary 1. Blue line: $\frac{r}{\vartheta} = 5 + 2\sqrt{6}$
Non-optimal regions for $L^0/L^1$ penalization

$G$ is $2 \times 2$ block-wise (diagonal 1, off-diagonal 0.5)

Left: GS. Middle: subset selection. Right: lasso (y-axis is prolonged)

$\epsilon_p = p^{-\vartheta}$, $\tau_p = \sqrt{2r \log p}$, each signal $\geq \tau_p$
Simulation comparison

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

- Main results not tied to Rare Weak model; hold much more broadly
- Extensions to non-random design is mostly straightforward
- Successfully extended to cases where $G$ is non-sparse but **sparsifiable**
  - change-point problem
  - long-memory time series
  - factor model

*Ke, Jin, Fan (2012)*
Take-home messages

- Proposed Graphlet Screening (GS) for variable selection
- Proved optimality of GS
- Key insight:
  - original model is decomposable due to interaction between signal sparsity and graph sparsity
  - minimax rate depends on $X$ ‘locally’ so we have to act ‘locally’
- Exposed intuition for the non-optimality of penalization methods