NETWORK-BASED PENALIZED REGRESSION

ALEXEJ GOSSMANN

TULANE UNIVERSITY

DEPT. OF GLOBAL BIOSTATISTICS AND DATA SCIENCE

2016/11/9

BACKGROUND

- Linear model $\mathbf{y} = X\mathbf{b} + \mathbf{z}$, where $\mathbf{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\mathbf{b} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{z}) = 0$.
- Possibly $n < p$.
- *Estimation/Prediction*: Find best predictions for $\mathbf{y}$.
- *Feature selection*: Find which b_i are non-zero.

MOTIVATION

- Typically penalized regression approaches **ignore** any relationships among the features $\mathbf{x}_i$.
- In biomedical applications features are related.
- A network of relationships between the features $\mathbf{x}_i$
  - can be constructed from the data (e.g., graphical model),
  - may be given as biological prior knowledge (e.g., genetic pathways from KEGG, etc.)

⇝ Utilize the network in the regression model!

CURRENTLY AVAILABLE METHODS

1. Bondell and Reich (2008) OSCAR
2. Yang et. al. (2013) GOSCAR
3. Li and Li (2008, 2010) Grace, aGrace
4. Pan et. al. (2010) Incorporating Predictor Network in Penalized Regression with Application to Microarray Data
5. Kim et. al. (2013) Network-based penalized regression with application to genomic data
6. Kim and Xing (2009) GFlasso
7. Zhu et. al. (2013) Simultaneous grouping pursuit and feature selection over an undirected graph
8. Yu and Liu (Oct. 2016) SRIG

OSCAR

H. Bondell and B. Reich (2008) "Simulataneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR"

OSCAR

Octagonal shrinkage and clustering algorithm:

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2$$

subject to

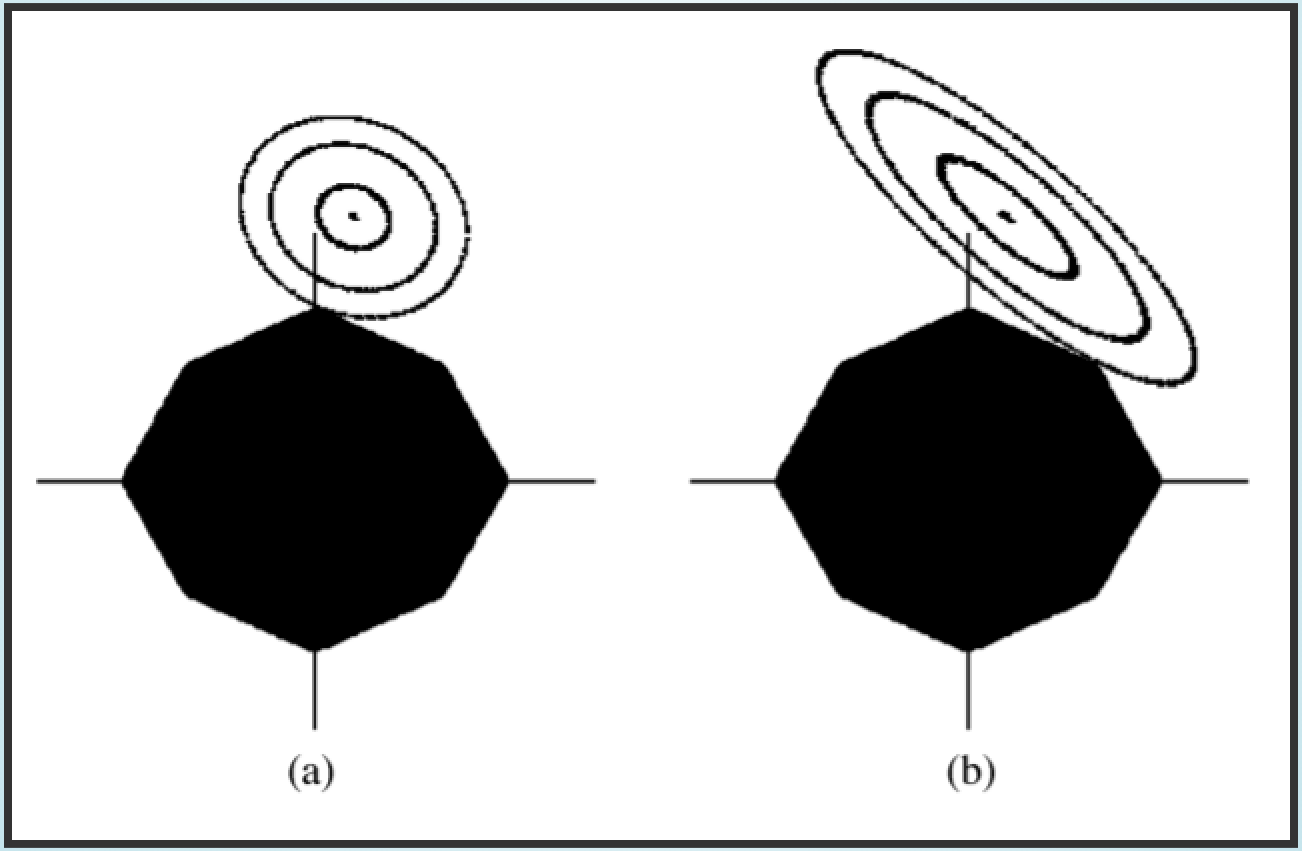
$$\sum_{j=1}^{p} |b_j| + c \sum_{j<k} \max\{|b_j|, |b_k|\} \leq t.$$

OSCAR

- ℓ_1 norm encourages sparsity.
- ℓ_∞ norm encourages equality of coefficients.
- OSCAR encourages grouping of highly correlated variables.
- OSCAR performs sparse regression while simultaneously performing supervised clustering.
- But no networks of graphs envolved!
- Application example: Appalachian Mountains soil data (predicting number of plant species based on soil characteristics).

OSCAR

(a) $\rho = 0.15$ (b) $\rho = 0.85$



OSCAR

AN INTERESTING OBSERVATION

OSCAR is actually a special case of SLOPE:

$$\arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda \sum_{j=1}^{p} (c(p-j)+1)|b|_{(j)},$$

for $|b|_{(1)} \geq \cdots \geq |b|_{(p)}$.

GOSCAR

S. Yang, L. Yuan, Y.-C. Lai, X. Shen, P. Wonka, and J. Ye (2013) "Feature Grouping and Selection Over an Undirected Graph"

GOSCAR

- Let (N, E) be the given undirected graph.
- *Assumption*: If nodes i and j are connected by an edge in E then $|b_i|$ and $|b_j|$ tend to be equal.
- Graph OSCAR:

$$\arg\min_{\mathbf{b}} \frac{1}{2}\|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda_1 \|\mathbf{b}\|_1 + \lambda_2 \sum_{(i,j)\in E} \max\{|b_i|, |b$$

GOSCAR

- When the graph is complete GOSCAR is equivalent to OSCAR.
- GOSCAR is much more challenging to solve than OSCAR.
- GOSCAR encourages equality of absolute values of coefficients for features connected in the graph.
- The ℓ_∞ regularizer can overpenalize large coefficients, resulting in strongly biased estimates.
- Not clear if robust to graph misspecification.
- Application example: Breast cancer data set ($n = 295$ cancer tumors and $p = 566$ genes), where GOSCAR outperforms LASSO, OSCAR and GFlasso.

GRACE AND AGRACE

C. Li and H. Li (2008) "Network-constrained regularization and variable selection for analysis of genomic data"

C. Li and H. Li (2010) "Variable selection and regression analysis for graph-structured covariates with an application to genomics"

GRAPH-CONSTRAINED ESTIMATION (GRACE)

1. Consider a weighted graph $G = (V, E, W)$.
2. Write $u \sim v$ if predictors u and v are linked in the network.
3. $w(u, v)$ denotes the weight of the edge $e = (u \sim v)$.
4. $d_v := \sum_{(u\sim v)} w(u, v)$ denotes the degree of vertex v.
5. The uvth element of the normalized Laplacian matrix L is defined as

$$L(u, v) := \begin{cases} 1 - w(u, v)/d_u, & \text{if } u = v, d_u \neq 0 \\ -w(u, v)/\sqrt{d_u d_v}, & \text{if } u \sim v \\ 0, & \text{otherwise} \end{cases}.$$

6. The smoothness of vector $\mathbf{b}$ with respect to the graph structure can be expressed as

$$\mathbf{b}^T L \mathbf{b} = \sum_{u\sim v} \left(\frac{b_u}{\sqrt{d_u}} - \frac{b_v}{\sqrt{d_v}} \right)^2 w(u, v).$$

7. A Gaussian Markov random field prior can be assumed for $\mathbf{b}$:

$$f(\mathbf{b}) \propto \exp\left(-\frac{1}{2\sigma^2} \mathbf{b}^T L \mathbf{b} \right).$$

GRAPH-CONSTRAINED ESTIMATION (GRACE)

- Grace:

$$\arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda_1 \|\mathbf{b}\|_1 + \lambda_2 \sum_{u\sim v} \Bigg(\frac{b_u}{\sqrt{d_u}} - \frac{b_v}{\sqrt{d_v}}$$

- Similar to the fused lasso (Tibshirani et. al. 2005), but utiliz network structure and ℓ_2 norm on the differences.
- The last term penalizes the vector $\mathbf{b}$, if it differs too much predictors that are linked in the graph.
- What if $u \sim v$, but b_u and b_v have different signs? (e.g., on neighboring genes is upregulated while the other is down

ADAPTIVE GRACE

aGrace

$$\arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda_1 \|\mathbf{b}\|_1$$

$$+ \lambda_2 \sum_{u\sim v} \left(\frac{\operatorname{sgn}(\tilde{b}_u) b_u}{\sqrt{d_u}} - \frac{\operatorname{sgn}(\tilde{b}_v) b_v}{\sqrt{d_v}} \right)$$

where $\tilde{\mathbf{b}}$ is an initial estimate obtained from LS, Ridge or Enet (2-step procedure).

GRACE APPLICATION EXAMPLE

- Glioblastoma microarray gene-expression data.
- $n = 50$ patients in the training data, $n = 61$ patients in the test data.
- $p = 1533$ genes, organized in a network of 33 KEGG pathways.
- Logarithm of time to death used as the response variable.
- Grace outperforms LASSO and Enet.

AGRACE APPLICATION EXAMPLE

- Analysis of gene expression data measured in human brains of individuals of different ages.
- Logarithm of age of $n = 30$ individuals as the response variable.
- Expression levels of $p = 1305$ genes as the predictors.
- KEGG network with 5288 edges.
- aGrace outperforms LASSO, Enet, and Grace.

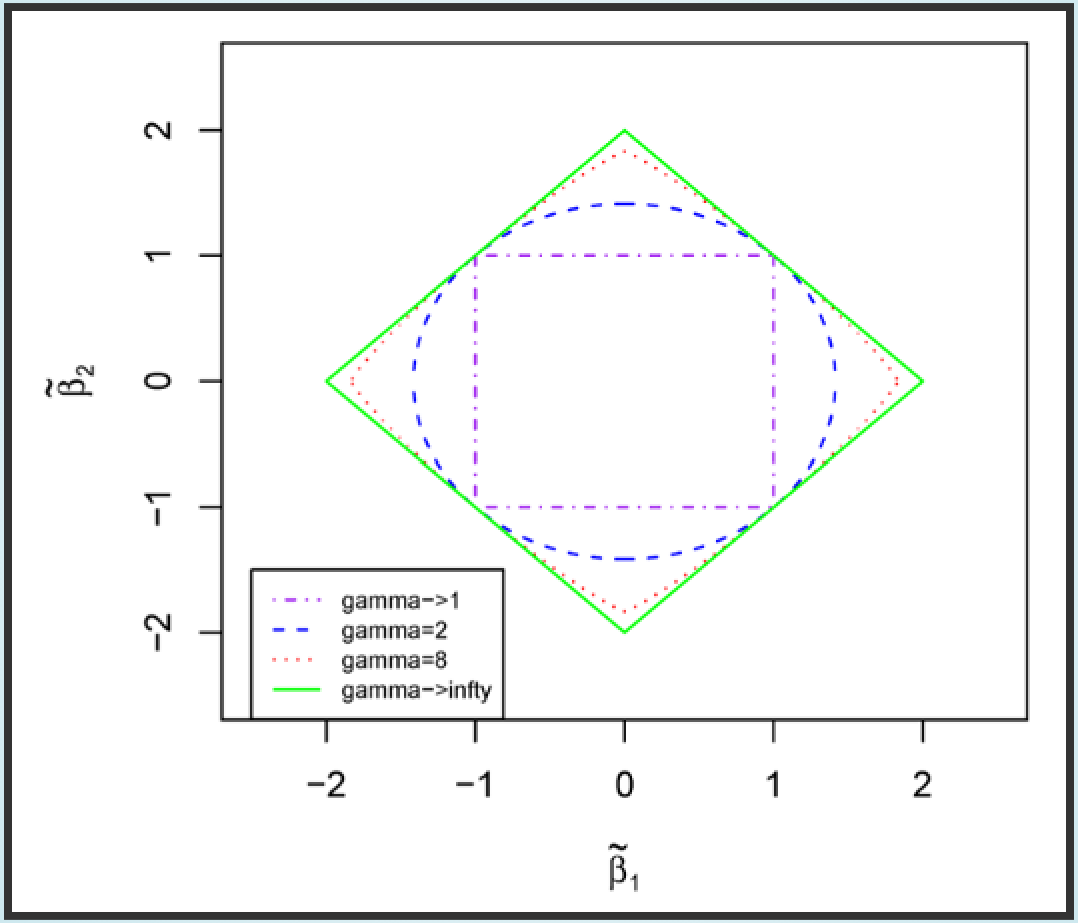
W. Pan, B. Xie, X. Shen (2010) "Incorporating Predictor Network in Penalized Regression with Application to Microarray Data"

- L_γ penalized regression (class of penalties):

$$\arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda 2^{\frac{1}{\gamma'}} \sum_{i \sim j} \left(\frac{|b_i|^\gamma}{w_i} + \frac{|b_j|^\gamma}{w_j} \right)^{\frac{1}{\gamma}},$$

- $1/\gamma' + 1/\gamma = 1$ and $\gamma > 1$.
- Each term is a weighted group penalty ⇒ connected features are likely to have similar effects.
- w_i determines what to smooth:
  - $w_i = d_i$ encourages $|b_i| \approx |b_j|$ if $i \sim j$.
  - $w_i = d_i^{(\gamma+1)/2}$ encourages $\frac{|b_i|}{\sqrt{d_i}} \approx \frac{|b_j|}{\sqrt{d_j}}$ if $i \sim j$.
- Larger γ ⇒ more smoothing.

L_γ-norm ball:



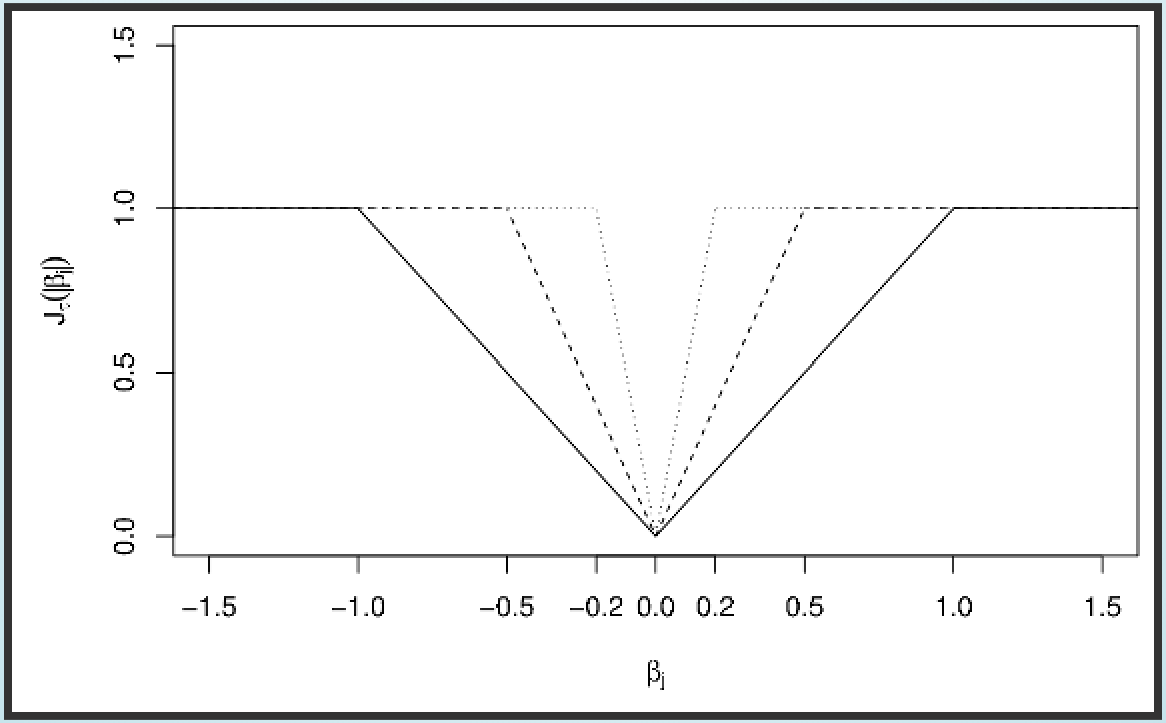
Larger $\gamma \Rightarrow$ more smoothing.

- S. Kim, W. Pan, and X. Shen (2013) "Network-based penalized regression with application to genomic data"
- Y. Zhu, X. Shen, and W. Pan (2013) "Simultaneous grouping and feature selection over an undirected graph"

- All previously shown methods assume that $\frac{|b_i|}{w_i} \approx \frac{|b_j|}{w_j}$ if $i \sim j$.
- Too strong an assumption?
- Relaxation: b_i and b_j are likely to be zero or non-zero at the same time (if $i \sim j$).

- Prior assumption: $I(b_i \neq 0) = I(b_j \neq 0)$ if $i \sim j$.
- Truncated Lasso Penalty (Shen et. al. 2012):

$$J_\tau(|z|) = \min\left(\frac{|z|}{\tau}, 1\right) \to I(z \neq 0), \quad \text{as } \tau \to 0^+.$$



- $TTLP_I$

$$\arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda_1 \sum_{i=1}^{p} J_\tau(|b_i|) + \lambda_2 \sum_{i\sim j} \left| J_\tau\left(\frac{|b_i|}{w_i}\right) - J_\tau\left(\frac{|b_j|}{w_j}\right)\right|,$$

- $LTLP_I$

$$\arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda_1 \sum_{i=1}^{p} |b_i| + \lambda_2 \sum_{i\sim j} \left| J_\tau\left(\frac{|b_i|}{w_i}\right) - J_\tau\left(\frac{|b_j|}{w_j}\right)\right|,$$

- Non-convex (use difference convex programming).

- Simulation studies:
  - $TTLP_I$ and $LTLP_I$ produce less biased estimates than Grace, aGrace, and ℓ_∞ based methods.
  - $TTLP_I$ and $LTLP_I$ are robust to misspecified weights and misspecified network, when compared to Grace, aGrace, and ℓ_∞ based methods.
- Breast cancer gene expression data:
  - $n = 286 + 295$ patients from two studies.
  - Binary outcome variable (metastasis).
  - Prior gene network of $p = 294$ genes and 326 edges.
- eQTL data

GFLASSO

S. Kim and E. P. Xing (2009) "Statistical Estimation of Correlated Genome Associations to a Quantitative Train Network"

GRAPH-GUIDED FUSED LASSO FOR MULTIPLE CORRELATED TRAITS (GFLASSO)

- Captures a network among multiple response variables:

$$\arg\min_{\mathbf{b}} \sum_k \|\mathbf{y}_k - X\mathbf{b}_k\|_2^2 + \lambda_1 \sum_k \sum_{i=1}^{p} |b_{ki}| + \lambda_2 \sum_{m\sim l} \sum_{i=1}^{p} |b_{mi} - \operatorname{sgn}(\rho_{ml}) b_{li}|.$$

- Utilizes a quantitative trait network in a multivariate regression model, in order to identify pleiotropic genes/SNPs.

GFLASSO

- When there is only one response variable, GFlasso can be used to capture a network between features:

$$\arg\min_{\mathbf{b}} \|\mathbf{y} - X\mathbf{b}\|_2^2 + \lambda_1 \|\mathbf{b}\|_1 + \lambda_2 \sum_{i\sim j} |b_i - \operatorname{sgn}(\rho_{ij}) b_j|.$$

SRIG

G. Yu and Y. Liu (2016) "Sparse Regression Incorporating Graphical Structure Among Predictors"

SRIG

- $\mathbf{y} = X\mathbf{b} + \mathbf{z}$, where $\mathbf{y} \in \mathbb{R}^n$, $X = [X_1, X_2, \ldots, X_p] \in \mathbb{R}^{n\times p}$, $\mathbf{b} \in \mathbb{R}^p$, $\mathbb{E}(\mathbf{z}) = 0$.
- Predictor graph G.
- *Random design setting*: For each *row* $\mathbf{x}_i$ of X assume that $\mathbb{E}(\mathbf{x}_i) = 0$ and $\mathbb{Var}(\mathbf{x}_i) = \Sigma$.
- Denote $\Omega = (\omega_{ij}) := \Sigma^{-1}$.
- Denote $\Sigma_{xy} = (c_1, c_2, \ldots, c_p)^T := \mathbb{Cov}(X_k, y_k)$.

$$\Rightarrow \Sigma_{xy} = \mathbb{E}(X^T Y/n) = \mathbb{E}(X^T X\mathbf{b}/n) + \mathbf{E}(X^T \mathbf{z}/n) = \Sigma\mathbf{b}$$

$$\Rightarrow \mathbf{b} = \Sigma^{-1}\Sigma_{xy} = \Omega\Sigma_{xy}$$

SRIG

- From $\mathbf{b} = \Omega\Sigma_{xy}$ we have that

$$\begin{aligned} b_1 &= c_1\omega_{11} + c_2\omega_{12} + \cdots + c_p\omega_{1p} \\ b_2 &= c_1\omega_{21} + c_2\omega_{22} + \cdots + c_p\omega_{2p} \\ &\vdots \\ b_p &= c_1\omega_{p1} + c_2\omega_{p2} + \cdots + c_p\omega_{pp} \end{aligned}$$

- Notice that $\mathbf{b}$ consists of p additive parts.
- If X_i is uncorrelated with $\mathbf{y}$, then $c_i = 0$ and $(c_1\omega_1, c_2\omega_2, \ldots, c_p\omega_p) = 0$.
- If $c_i \neq 0$, then the support of $(c_i\omega_{1i}, c_i\omega_{2i}, \ldots, c_i\omega_{pi})$ is determined by the neighborhood of node i.

SRIG

- Adjacency matrix E (0-1-valued, convention: $E_{i,i} = 1$).
- $\mathcal{N}_i := \{j : E_{ij} \neq 0\}$ = "ith node and its neighbors".
- Change of variables:

$$\begin{aligned} b_1 &= V_1^{(1)} E_{11} + V_1^{(2)} E_{12} + \cdots + V_1^{(p)} E_{1p} \\ b_2 &= V_2^{(1)} E_{21} + V_2^{(2)} E_{22} + \cdots + V_2^{(p)} E_{2p} \\ &\vdots \\ b_p &= V_p^{(1)} E_{p1} + V_p^{(2)} E_{p2} + \cdots + V_p^{(p)} E_{pp} \end{aligned}$$

⇝ Decomposition $\mathbf{b} = V^{(1)} + V^{(2)} + \cdots + V^{(p)}$, where $V^{(i)} = 0$ for some i, and $\operatorname{supp}(V^{(i)}) \subset \mathcal{N}_i$ for all i.

SPARSE REGRESSION INCORPORATING GRAPHICAL STRUCTURE AMONG PREDICTORS (SRIG)

$$\min_{\mathbf{b},V^{(1)},\ldots,V^{(p)}} \frac{1}{2n}\|\mathbf{y}-X\mathbf{b}\|_2^2+\lambda\sum_{i=1}^{p}\tau_i\left\|V^{(i)}\right\|_2,$$

subject to

$$\sum_{i=1}^{p}V^{(i)}=\mathbf{b},$$

$$\mathrm{supp}(V^{(i)})\subseteq\mathcal{N}_i,\quad \forall i=1,\ldots,p.$$

SRIG

- Node-by-node rather than edge-by-edge.
- Adaptive LASSO (no edges in G), group LASSO (G consists of disconnected complete subgraphs), and Ridge Regression (G is a complete graph) are special cases.
- Theoretical finite sample bounds for prediction and estimation.
- Model selection consistency.

SRIG

- Simulation results
  - SRIG performs well for estimation, prediction, and feature selection.
  - SRIG generally outperforms LS, LASSO, aLASSO, Ridge, Enet, PCR, SPLS, GOSCAR, GRACE, under the assumption that connected variables in the graph act together.
  - If the intersection between the neighborhoods of relevant and irrelevant predictors is big, then LASSO outperforms SRIG.

SRIG

ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE (ADNI)

- Mini Mental State Examination (MMSE) score (0-30 points) is predicted from structural MRI.
- 51 AD patients, 52 controls; total: $n = 103$.
- 93 regions of interest (ROI); for each ROI, volume of GM tissue used as a feature; total $p = 93$.
- G estimated by the graphical Lasso (Friedman et. al. 2008), has 419 edges.
- SRIG outperforms LASSO, Ridge, aLasso, Enet, GOSCAR, GRACE, PCR, SPLS in terms of MSE on test data.