The Graph-guided Group Lasso

Zi Wang

Imperial College London, United Kingdom

8th July 2013

Outline

1 Introduction
- Bioinformatics motivation
- Penalized regression
- Incorporating prior knowledge

Outline

1 Introduction
- Bioinformatics motivation
- Penalized regression
- Incorporating prior knowledge

2 The Graph-guided Group Lasso
- GGGL-1
- GGGL-2
- Estimation algorithms

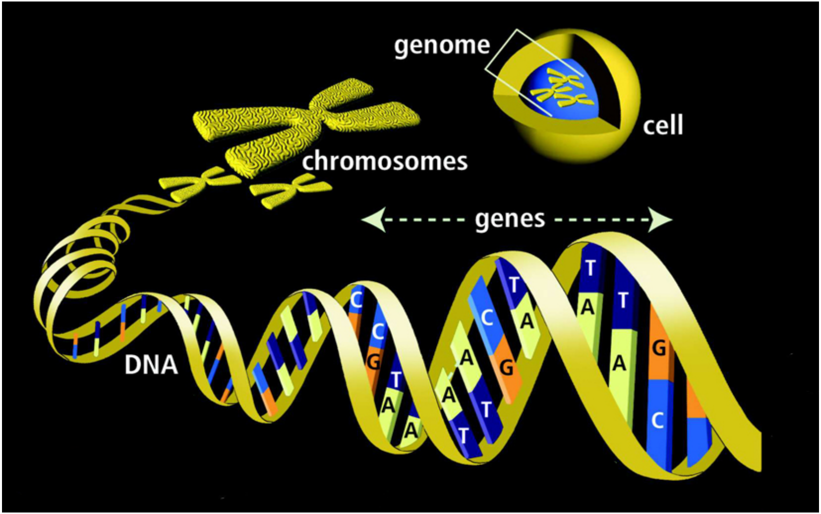
Outline

1. Introduction
   - Bioinformatics motivation
   - Penalized regression
   - Incorporating prior knowledge
2. The Graph-guided Group Lasso
   - GGGL-1
   - GGGL-2
   - Estimation algorithms
3. Preliminary results

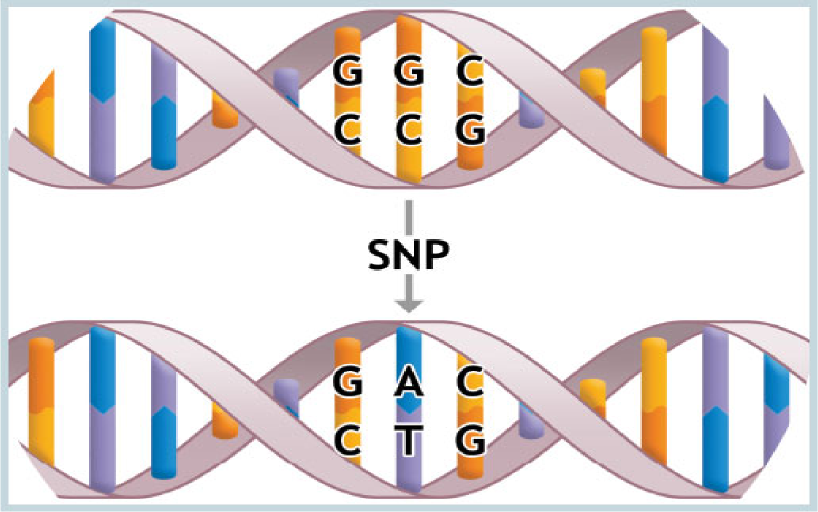
Outline

1. Introduction
   - Bioinformatics motivation
   - Penalized regression
   - Incorporating prior knowledge
2. The Graph-guided Group Lasso
   - GGGL-1
   - GGGL-2
   - Estimation algorithms
3. Preliminary results
4. Future works

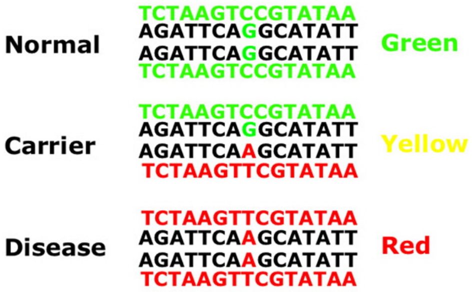
A 30s introduction to the biology



Single-nucleotide polymorphisms (SNPs)



Genome-wide association study (GWAs)



Objective: To identify important predictors (e.g. SNPs), that account for the variability of a quantitative trait.

Notation

- X: $n \times p$ predictor matrix containing n observations on p covariates.
- y: n observations on univariate continuous response.
- β: $p \times 1$ coefficient matrix.
- ϵ: $n \times 1$ matrix. $\mathbb{E}(\epsilon_i) = 0$, $\forall i$.

Use linear regression model:

$$y = X\beta + \epsilon$$

where X and y are columnwise centered, such that the intercept term can be dropped.

Sparse solution

Note:

$$\hat{\beta}_i = 0 \Leftrightarrow X_i \text{ is excluded from the model}$$

Thus, if there are only a handful of i such that: $\hat{\beta}_i \neq 0$, then the set:

$$\{X_i : \hat{\beta}_i \neq 0\}$$

corresponds to the set of "important" predictors (causal SNPs).

Penalized linear regression

An ordinary least square estimate minimizes:

$$\|y - X\beta\|_2^2$$

Penalized linear regression

An ordinary least square estimate minimizes:

$$\|y - X\beta\|_2^2$$

A penalized linear regression estimate minimizes:

$$\|y - X\beta\|_2^2 + P(\beta)$$

where $P(\beta)$ is called “the penalty term”.

Some notable penalties that impose sparsity

Lasso:

$$P(\beta) = \lambda \cdot \|\beta\|_1$$

Elastic-net:

$$P(\beta) = \lambda_1 \cdot \|\beta\|_2 + \lambda_2 \cdot \|\beta\|_1$$

Incorporating prior biological knowledge - Variable grouping

- Multiple SNPs from one gene often jointly carry out genetic functionalities.

[1] Association screening of common and rare genetic variants by penalized regression. (*Bioinformatics 26(19): 2375-2382. 2010.*)

[2] Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. (*Bioinformatics 28(2): 229-237. 2012.*)

Incorporating prior biological knowledge - Variable grouping

- Multiple SNPs from one gene often jointly carry out genetic functionalities.
  $\Rightarrow$ SNPs grouped into genes

[1] Association screening of common and rare genetic variants by penalized regression. (*Bioinformatics 26(19): 2375-2382. 2010.*)

[2] Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. (*Bioinformatics 28(2): 229-237. 2012.*)

Incorporating prior biological knowledge - Variable grouping

- Multiple SNPs from one gene often jointly carry out genetic functionalities.
  $\Rightarrow$ SNPs grouped into genes
- Prior information: Partition of predictors into groups.

[1] Association screening of common and rare genetic variants by penalized regression. (*Bioinformatics 26(19): 2375-2382. 2010.*)

[2] Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. (*Bioinformatics 28(2): 229-237. 2012.*)

Incorporating prior biological knowledge - Variable grouping

- Multiple SNPs from one gene often jointly carry out genetic functionalities.
  $\Rightarrow$ SNPs grouped into genes
- Prior information: Partition of predictors into groups.
- Desired sparsity pattern:

$$\hat{\beta} = (\underbrace{[0.2, 0, 0]}_{\text{group 1}}, \underbrace{[0, 0, 0, ..., 0]}_{\text{group 2}}, \underbrace{[0, 0.5, 0, 0, 0, 0.1]}_{\text{group 3}}, ...)$$

[1] Association screening of common and rare genetic variants by penalized regression. (*Bioinformatics 26(19): 2375-2382. 2010.*)

[2] Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. (*Bioinformatics 28(2): 229-237. 2012.*)

Incorporating prior biological knowledge - Variable grouping

- Multiple SNPs from one gene often jointly carry out genetic functionalities.
  $\Rightarrow$ SNPs grouped into genes
- Prior information: Partition of predictors into groups.
- Desired sparsity pattern:

$$\hat{\beta} = (\underbrace{[0.2, 0, 0]}_{\text{group 1}}, \underbrace{[0, 0, 0, ..., 0]}_{\text{group 2}}, \underbrace{[0, 0.5, 0, 0, 0, 0.1]}_{\text{group 3}}, ...)$$

- e.g. Zhou *et al.* [1], H. Wang *et al.* [2]

[1] Association screening of common and rare genetic variants by penalized regression. (*Bioinformatics 26(19): 2375-2382. 2010.*)

[2] Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. (*Bioinformatics 28(2): 229-237. 2012.*)

Incorporating prior biological knowledge - Network

- Genes belonging to the same pathway are often expressed similarly in response.

[3] Network-constrained regularization and variable selection for analysis of genomic data. (*Bioinformatics. Vol. 24 no. 9, pages 1175-1182 2008.*)

Incorporating prior biological knowledge - Network

- Genes belonging to the same pathway are often expressed similarly in response.
  $\Rightarrow$ Gene regulatory network

[3] Network-constrained regularization and variable selection for analysis of genomic data. (*Bioinformatics. Vol. 24 no. 9, pages 1175-1182 2008.*)

Incorporating prior biological knowledge - Network

- Genes belonging to the same pathway are often expressed similarly in response.
  $\Rightarrow$ Gene regulatory network
- Prior information: Pairwise relations on predictors encoded in a network.

[3] Network-constrained regularization and variable selection for analysis of genomic data. (*Bioinformatics. Vol. 24 no. 9, pages 1175-1182 2008.*)

Incorporating prior biological knowledge - Network

- Genes belonging to the same pathway are often expressed similarly in response.
  $\Rightarrow$ Gene regulatory network
- Prior information: Pairwise relations on predictors encoded in a network.
- Desired sparsity pattern: connected variables are encouraged to be selected together.

[3] Network-constrained regularization and variable selection for analysis of genomic data. (*Bioinformatics. Vol. 24 no. 9, pages 1175-1182 2008.*)

Incorporating prior biological knowledge - Network

- Genes belonging to the same pathway are often expressed similarly in response.
  $\Rightarrow$ Gene regulatory network
- Prior information: Pairwise relations on predictors encoded in a network.
- Desired sparsity pattern: connected variables are encouraged to be selected together.
- e.g. Li and Li [3]

[3] Network-constrained regularization and variable selection for analysis of genomic data. (*Bioinformatics. Vol. 24 no. 9, pages 1175-1182 2008.*)

Incorporating prior knowledge at multiple levels

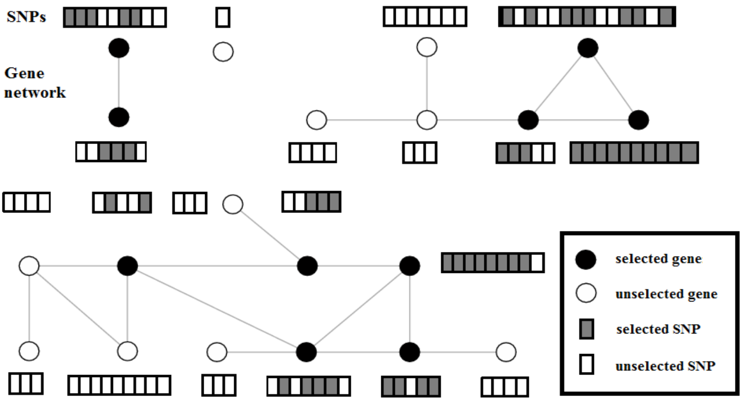


Figure : Sparsity pattern of the proposed "Graph-guided Group Lasso" (GGGL)

The between-group relations

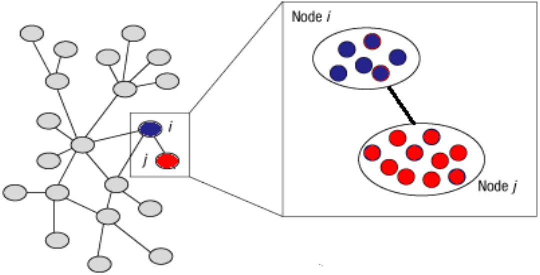


Figure : The key part of GGGL: How to incorporate information at heterogeneous levels

Notation

- X, y, β as defined before. Further require the columns of X to have Euclidean norm 1.
- Let $\mathcal{R} = \{R_1, R_2, ...\}$ be a partition of the predictors. Denote the size of R_I by $|R_I|$, the the $n \times |R_I|$ sub-matrix of X by X_I, and the i^{th} column of X by X_i
- Let $\mathcal{G} = \mathcal{G}(V, E)$ be the given network whose vertex set V corresponds to the groups in $\mathcal{R}$. The weight of the edge $K - L$ is denoted by w_{KL} ($w.l.o.g.$ $w_{KL} \geq 0$), which can be either binary or continuous.

GGGL-1: Illustration

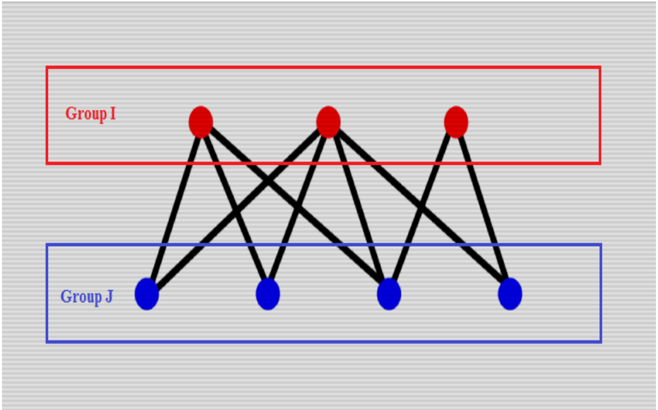


Figure : GGGL-1: If $R_I \sim R_J$, then reformulate a complete bipartite graph with vertex sets R_I and R_J. Edge weights $w_{ij} = W_{IJ} \; \forall i \in R_I, \forall j \in R_J$.

GGGL-1: The model

GGGL-1 minimizes the following objective function on β:

$$\frac{1}{2}\|y - X\beta\|_2^2 + P_1(\beta) + P_2(\beta) + P_3(\beta)$$

where:

$$P_1(\beta) = \lambda_1 \sum_{I:R_I \in \mathcal{R}} \sqrt{|R_I|} \cdot \|\beta_I\|_2, \qquad P_2(\beta) = \lambda_2 \cdot \|\beta\|_1$$

$$P_3(\beta) = \frac{1}{2}\,\mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}\,(\beta_i - \beta_j)^2$$

GGGL-1: Smoothing effect

Proposition (1)

For fixed μ, let $\hat{\beta}$ be the vector that minimizes:

$$\|y - X\beta\|_2^2 + \mu \sum_{k,l:X_k \in R_K, X_l \in R_L} w_{KL} \, (\beta_k - \beta_l)^2$$

Define the following:

$$\rho_{ij} = X_i' X_j, \quad C_I = \sum_{K \sim I} w_{IK} |R_K|, \quad \Gamma_I = \frac{\sum_{k \in R_K, K \sim I} w_{IK} \hat{\beta}_k}{C_I}$$

Then:

$$|(\hat{\beta}_i - \hat{\beta}_j) - (\Gamma_I - \Gamma_J)| \leq \frac{\|y\|_2}{\mu} \left(\frac{\sqrt{2(1-\rho_{ij})}}{C_I} + \left| \frac{1}{C_I} - \frac{1}{C_J} \right| \right)$$

GGGL-1: A potential side effect

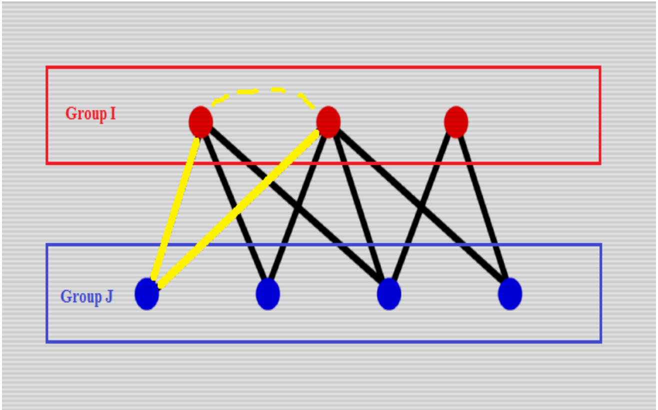


Figure : GGGL-1: Smoothing the coefficients of variables belonging to the same group may be undesirable.

GGGL-2: Another interpretation

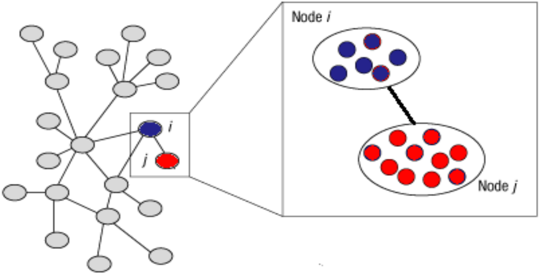


Figure : GGGL-2: encourage connected groups to be selected together $\neq$ every pair of variables should be encouraged to be selected together

GGGL-2: The model

In the objective function of GGGL-1, $P_3(\beta)$ is taken as:

$$P_3(\beta) = \frac{1}{2}\,\mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}\,(\beta_i - \beta_j)^2$$

For GGGL-2, replace it by:

$$P_3(\beta) = \frac{1}{2}\,\mu \cdot \sum_{I \sim J} w_{IJ}(\bar{\beta}_I - \bar{\beta}_J)^2$$

where $\bar{\beta}_I = \frac{1}{|R_I|} \sum_{i:\ i \in R_I} \beta_i$

GGGL-2: The model

In the objective function of GGGL-1, $P_3(\beta)$ is taken as:

$$P_3(\beta) = \frac{1}{2}\,\mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}\,(\beta_i - \beta_j)^2$$

For GGGL-2, replace it by:

$$P_3(\beta) = \frac{1}{2}\,\mu \cdot \sum_{I \sim J} w_{IJ}(\bar{\beta}_I - \bar{\beta}_J)^2$$

where $\bar{\beta}_I = \frac{1}{|R_I|} \sum_{i:\ i \in R_I} \beta_i$
With constraint: $\beta_i \geq 0,\ \forall i$.

GGGL-2: Smoothing effect

Proposition (2)

For fixed μ, let $\hat{\beta}$ be the vector that minimises:

$$\|y - X\beta\|_2^2 + \mu \sum_{K\sim L} w_{KL}(\bar{\beta}_K - \bar{\beta}_L)^2$$

Let d_I be the vertex degree of group R_I in $\mathcal{G}$ and define:

$$\Theta_I = \sum_{K\sim I} \frac{w_{IK}}{d_I}\bar{\hat{\beta}}_K, \quad D_\mu(I,J) = |(\bar{\hat{\beta}}_I - \bar{\hat{\beta}}_J) - (\Theta_I - \Theta_J)|$$

Then:

$$D_\mu(I,J) \leq \frac{\|y\|_2}{\mu}\left(\frac{2|R_I|}{d_I} + \left|\frac{|R_I|}{d_I} - \frac{|R_J|}{d_J}\right|\right)$$

GGGL-2: Within-group effect

Corollary (3)

Assuming X_i and X_j belong to the same group and defining the partial residual $\hat{r}_{ij} = y - \sum_{k \neq i,\, j} X_k \hat{\beta}_k$, the estimated coefficients $\hat{\beta}$ satisfy:

$$|\hat{\beta}_i - \hat{\beta}_j| = \frac{|(X_i' - X_j')\hat{r}_{ij}|}{1 - \rho_{ij}}$$

Comparison: GGGL-1 and GGGL-2 smoothing effect

GGGL-1 penalty:

$$P(\beta) = \lambda_1 \sum_{I:R_I \in \mathcal{R}} \sqrt{|R_I|} \cdot \|\beta_I\|_2 + \frac{1}{2}\,\mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}\,(\beta_i - \beta_j)^2$$

GGGL-2 penalty:

$$P(\beta) = \lambda_1 \sum_{I:R_I \in \mathcal{R}} \sqrt{|R_I|} \cdot \|\beta_I\|_2 + \frac{1}{2}\,\mu \cdot \sum_{I \sim J} w_{IJ}(\bar{\beta}_I - \bar{\beta}_J)^2$$

Comparison: GGGL-1 and GGGL-2 smoothing effect

GGGL-1 penalty:

$$P(\beta) = \lambda_1 \sum_{I:R_I \in \mathcal{R}} \sqrt{|R_I|} \cdot \|\beta_I\|_2 + \frac{1}{2}\,\mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}\,(\beta_i - \beta_j)^2$$

GGGL-2 penalty:

$$P(\beta) = \lambda_1 \sum_{I:R_I \in \mathcal{R}} \sqrt{|R_I|} \cdot \|\beta_I\|_2 + \frac{1}{2}\,\mu \cdot \sum_{I \sim J} w_{IJ}(\bar{\beta}_I - \bar{\beta}_J)^2$$

Tune λ_1 so that both models select the same number of groups. Tune μ such that $\sum_{I \sim J} w_{IJ}(\bar{\beta}_I - \bar{\beta}_J)^2$ are about equal for both models.

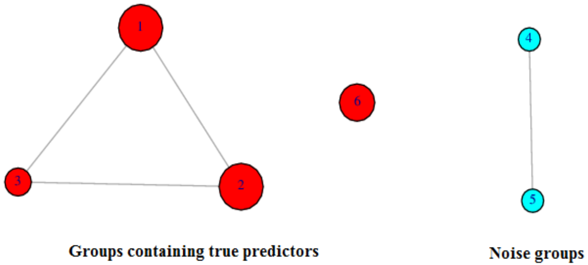
Data generation: key settings

$n = 200, \quad p = 60, \quad$ partitioned into 6 equal groups

Data generation: key settings

$n = 200, \quad p = 60,$ partitioned into 6 equal groups

Specified network:



Comparison: small μ for GGGL-1

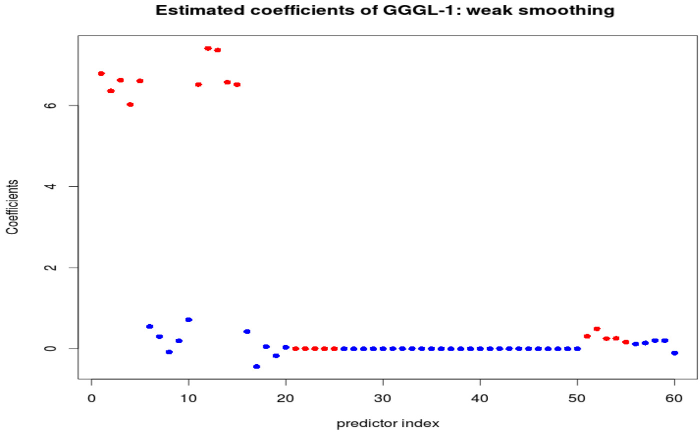


Figure : Red dots represent true variables, blue dots represent noise variables.

Comparison: large μ for GGGL-1

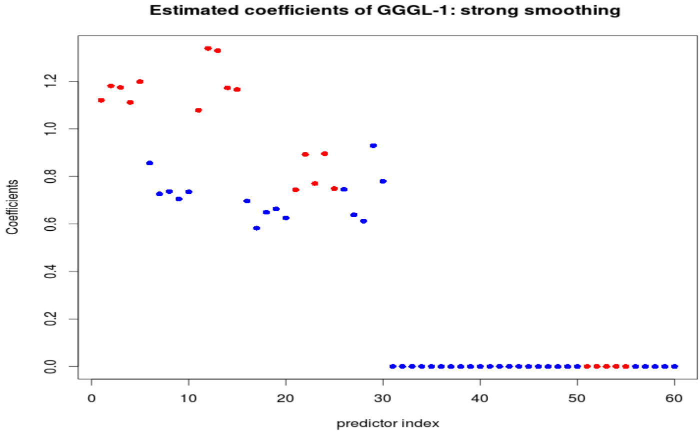


Figure : Red dots represent true variables, blue dots represent noise variables.

Comparison: small μ for GGGL-2

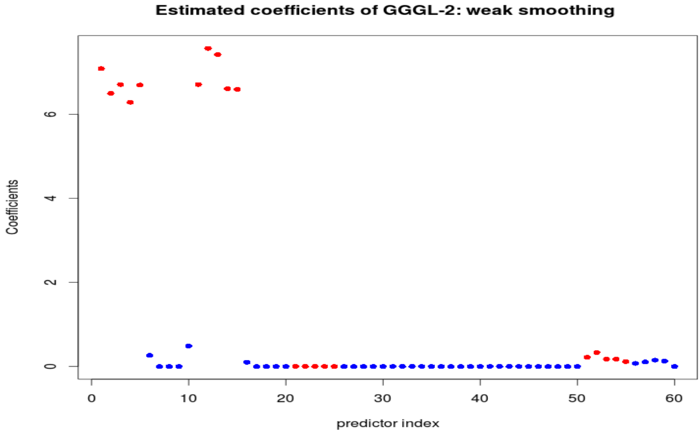


Figure : Red dots represent true variables, blue dots represent noise variables.

Comparison: large μ for GGGL-2

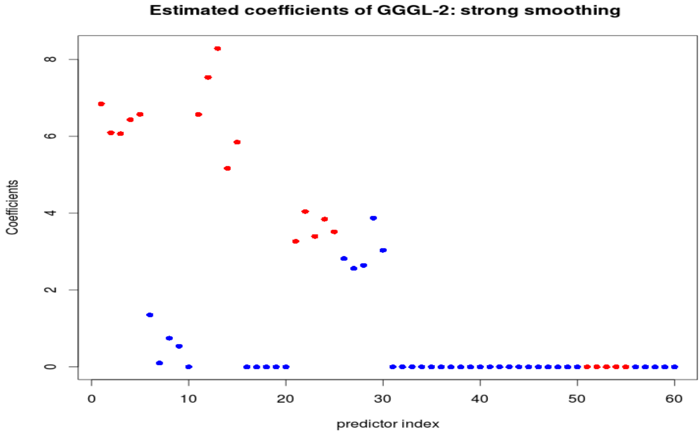


Figure : Red dots represent true variables, blue dots represent noise variables.

Estimation algorithm: GGGL-1

Note:

$$\sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}(\beta_i - \beta_j)^2 = \sum_{i \leq j} w_{ij}(\beta_i - \beta_j)^2$$

where w_{ij} is defined as:

$$w_{ij} = \begin{cases} 0 & \text{if} \quad X_i \text{ and } X_j \text{ belongs to the same group} \\ w_{IJ} & \text{if} \quad X_i \in R_I,\, X_j \in R_J \neq R_I \end{cases}$$

Estimation algorithm: GGGL-1

Note:

$$\sum_{i\in R_I, j\in R_J, I\sim J} w_{IJ}(\beta_i - \beta_j)^2 = \sum_{i\leq j} w_{ij}(\beta_i - \beta_j)^2$$

where w_{ij} is defined as:

$$w_{ij} = \begin{cases} 0 & \text{if} \quad X_i \text{ and } X_j \text{ belongs to the same group} \\ w_{IJ} & \text{if} \quad X_i \in R_I,\, X_j \in R_J \neq R_I \end{cases}$$

Let L be a $p \times p$ matrix whose (i,j)th entry is:

$$(L)_{ij} = \begin{cases} \sum_{j\neq i} w_{ij} & \text{if} \quad i = j \\ -w_{ij} & \text{if} \quad i \neq j \end{cases}$$

Estimation algorithm: GGGL-1

Note:

$$\sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}(\beta_i - \beta_j)^2 = \sum_{i \leq j} w_{ij}(\beta_i - \beta_j)^2$$

where w_{ij} is defined as:

$$w_{ij} = \begin{cases} 0 & \text{if} \quad X_i \text{ and } X_j \text{ belongs to the same group} \\ w_{IJ} & \text{if} \quad X_i \in R_I,\, X_j \in R_J \neq R_I \end{cases}$$

Let L be a $p \times p$ matrix whose (i, j)th entry is:

$$(L)_{ij} = \begin{cases} \sum_{j \neq i} w_{ij} & \text{if} \quad i = j \\ -w_{ij} & \text{if} \quad i \neq j \end{cases}$$

Using L, the right hand side can be re-formulated into:

$$\sum_{i \leq j} w_{ij}(\beta_i - \beta_j)^2 = \beta' L \beta$$

Estimation algorithm: GGGL-1

Up to this point, we have:

$$\|y - X\beta\|_2^2 + \mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}(\beta_i - \beta_j)^2 = \|y - X\beta\|_2^2 + \mu \beta' L \beta$$

Estimation algorithm: GGGL-1

Up to this point, we have:

$$\|y - X\beta\|_2^2 + \mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}(\beta_i - \beta_j)^2 = \|y - X\beta\|_2^2 + \mu\beta' L\beta$$

Note L is positive semi-definite, therefore we can find $p \times p$ matrix U such that: $L = UU'$, using singular value decomposition. We then construct the $(n+p) \times 1$ matrix y^* and the $(n+p) \times p$ matrix $X*$ according to:

$$y* = \begin{pmatrix} y_{n\times 1} \\ 0_{p\times 1} \end{pmatrix}, \qquad X^* = \begin{pmatrix} X \\ \sqrt{\mu}U' \end{pmatrix}$$

Estimation algorithm: GGGL-1

Up to this point, we have:

$$\|y - X\beta\|_2^2 + \mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ}(\beta_i - \beta_j)^2 = \|y - X\beta\|_2^2 + \mu\beta' L\beta$$

Note L is positive semi-definite, therefore we can find $p \times p$ matrix U such that: $L = UU'$, using singular value decomposition. We then construct the $(n+p) \times 1$ matrix y^* and the $(n+p) \times p$ matrix $X*$ according to:

$$y* = \begin{pmatrix} y_{n \times 1} \\ 0_{p \times 1} \end{pmatrix}, \qquad X^* = \begin{pmatrix} X \\ \sqrt{\mu} U' \end{pmatrix}$$

Therefore the optimization problem of GGGL-1 is equivalent to:

$$\|y^* - X^*\beta\|_2^2 + 2\lambda_1 \sum_{I:\ R_I \in \mathcal{R}} \sqrt{|R_I|}\|\beta_I\|_2 + 2\lambda_2\|\beta\|_1$$

Estimation algorithm: GGGL-2

Note:

$$\sum_{I \sim J} w_{IJ}(\bar{\beta}_I - \bar{\beta}_J)^2 = \beta' \mathcal{L} \beta$$

where $\mathcal{L}$ is defined as:

$$(\mathcal{L})_{ij} = \begin{cases} \sum_{\{K:K\sim I\}} \frac{w_{IK}}{|R_I|^2} & \text{if} \quad X_i \in R_I, X_j \in R_I \\ -\frac{w_{IJ}}{|R_I|\cdot|R_J|} & \text{if} \quad X_i \in R_I, X_j \in R_J \end{cases}$$

Estimation algorithm: GGGL-2

Note:

$$\sum_{I \sim J} w_{IJ}(\bar{\beta}_I - \bar{\beta}_J)^2 = \beta' \mathcal{L} \beta$$

where $\mathcal{L}$ is defined as:

$$(\mathcal{L})_{ij} = \begin{cases} \sum_{\{K:K\sim I\}} \frac{w_{IK}}{|R_I|^2} & \text{if} \quad X_i \in R_I, X_j \in R_I \\ -\frac{w_{IJ}}{|R_I|\cdot|R_J|} & \text{if} \quad X_i \in R_I, X_j \in R_J \end{cases}$$

Both optimization problems can be solved using standard block coordinate descent algorithm.

Parallel computation: outline

- For large scale data analysis it is necessary to parallelize.
- In each step, update a subset of the groups in parallel.
- An application of Richtarik and Takac [4]
- Code written in CUDA, to run on graphics processing units (GPUs).
- On a data set where $n = 3000$, $p = 2000$ partitioned into 200 groups, we observed a larger than $10\times$ speed-up compared with the non-parallel algorithm written in C.

[4] Parallel coordinate descent methods for big data optimization. (*arXiv:1212.0873, 2012.*)

Parallel computation: outline

Parallel Coordinate Descent Method

Input: Data, parameters, m groups to update in each step.
Output: column vector $\hat{\beta}$

1. Choose initial estimate $\hat{\beta}^{(0)}$.
2. $k \leftarrow 1$
3. Randomly pick a set of blocks from $\mathcal{R}$: $k_1, k_2, \ldots k_m$.
4. In parallel do: $\hat{\beta}^{(k+1)}_{R_{k_m}} \leftarrow \phi(\hat{\beta}^{(k)}, k_m)$, for $m = 1, 2, \ldots$.
5. Collect estimates from the processors to obtain $\hat{\beta}^{(k+1)}$.
6. Set $k \leftarrow k + 1$ and go back to 3 until convergence.

ϕ is defined so that at each step: $\mathbb{E}[F(\hat{\beta}^{(k+1)})|\hat{\beta}^{(k)}] \leq \mathbb{E}[F(\hat{\beta}^{(k)})]$, where F is the objective function.

Preliminary results

Data generation:

- $n = 200$, $p = 800$, fixed grouping of X's into 80 groups. $X \sim \mathcal{N}(0, \Sigma)$.
- All predictors in R_1, ..., R_{40} are true variables, all predictors in the other groups are noise variables.
- Compute $y = X\beta + \delta \cdot \epsilon$, where β_i's are independently generated from uniform$(0.5, 1)$ distribution for true variables. ϵ_i's are $i.i.d.$ standard normal RVs, δ controls signal-to-noise level to 1.
- X is columnwise normalized and y is centered.

Networks for GGGL

We categorize the networks into 3 types, according to their relevance to the study:

- informative: true variables are connected (not necessarily in one component though) whereas there are very few links between true variables and noise variables.
- uninformative: all pairs of variables are connected with roughly equal probabilities.
- noisy: true variables and noise variables form an almost bipartite graph and the true variables are rarely linked.

Illustration of networks

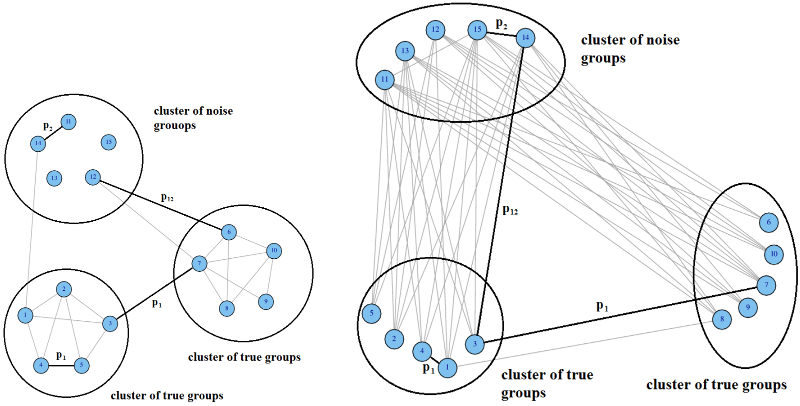


Figure : Left: informative network; Right: noisy network.

Experiment design: GGGL-1 vs Group lasso

Repeat for 200 data sets:

- Generate random network with probabilities of connection: $p_1 = 0.7$ (between true groups), $p_{12} = 0.01$ (between a true group and a noise group), $p_2 = 0.1$ (between noise groups).
- Fix $\mu = 50$ and $\lambda_2 = 0$ in GGGL-1. So the GGGL-1 penalty becomes:

$$P(\beta) = \lambda_1 \sum_{I:R_I \in \mathcal{R}} \sqrt{|R_I|} \cdot \|\beta_I\|_2 + \frac{1}{2} \mu \sum_{i \in R_I, j \in R_J, I \sim J} w_{IJ} \, (\beta_i - \beta_j)^2$$

  with $\mu = 10$, and the group lasso penalty is simply when $\mu = 0$.
- Tune λ_1 so that both models select exactly 40 groups.

Rank the groups according to selection frequencies for each model, and compare using the receiver operating characteristic (ROC) curves.

GGGL-1 vs Group lasso

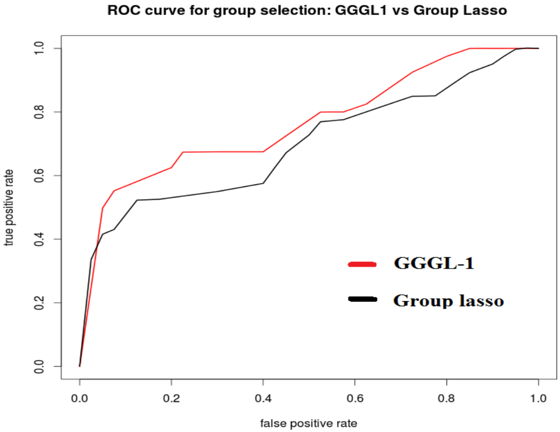


Figure : Comparison of GGGL-1 and Group lasso on group selection using ROC curves, where GGGL-1 shows superior power.

Future works

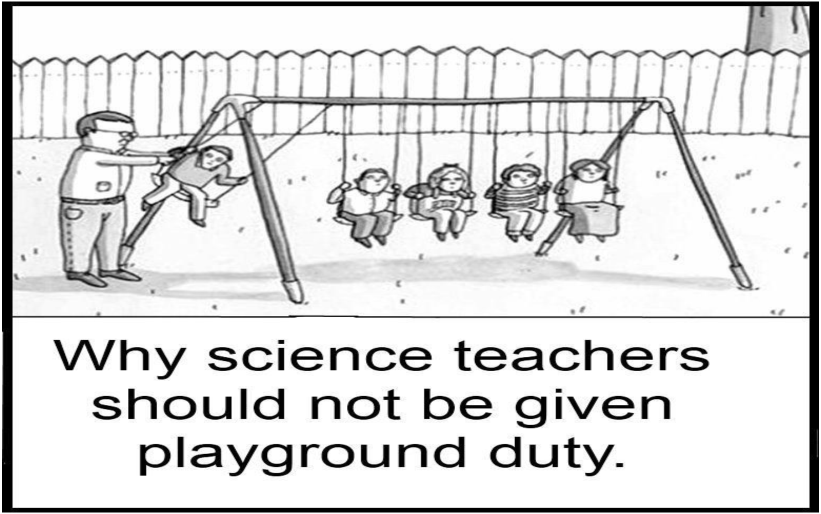
- Complete simulation study on GGGL-2
- Study the performance of GGGL models on the three types of networks.
- Application to tumor data set.

Acknowledgement

- Dr. Giovanni Montana, Imperial College London
- Dr. Ed Curry, Imperial College London
- Peter Nash, Imperial College London

Reference

- Tibshirani. Regression shrinkage and selection via the lasso. *J.R.Statist.* Soc.B, 58:267-288. 1996.
- Zhou *et al.* Association screening of common and rare genetic variants by penalized regression. *Bioinformatics* 26(19): 2375-2382. 2010.
- Wang *et al.* Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the ADNI cohort. *Bioinformatics* 28(2): 229-237. 2012.
- Li and Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics.* Vol. 24 no. 9, pages 1175-1182 2008.
- Yuan and Lin. Model selection and estimation in regression with grouped variables. *J.R.Statist.* Soc.B, 68(1):49-67, 2006.
- Friedman *et al.* A note on the group lasso and a sparse group lasso. *arXiv:*1001.0736. 2010.
- Daye and Jeng. Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics and Data Analysis 53(4), 1284-1298.* 2009.
- Chung. Spectral graph theory. *CBMS regional conference series 92. Amer. Math. Soc., Providence, RI. MR1421568.* 1997.
- Richtárik and Takáč Parallel coordinate descent methods for big data optimization. *arXiv:1212.0873.* 2012


Why science teachers should not be given playground duty.