Simultaneous Supervised Clustering and Feature Selection over a Graph

Xiaotong Shen

University of Minnesota

International Workshop on Perspectives on High-Dimensional Data Analysis, June 9-11, 2011

Joint with Hsin-Cheng Huang, Academia Sinica, and Wei Pan, U of Minnesota.

Outline

Introduction

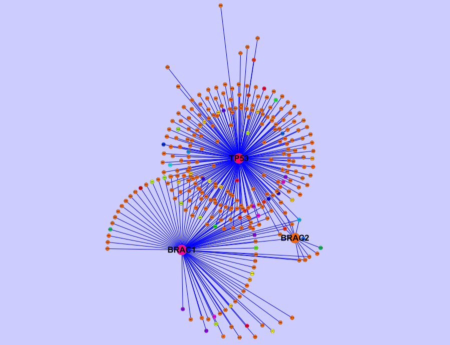
- Response: $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$
- Predictors: p-dimensional $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$.
- Regression model

$$Y_i = \boldsymbol{x}_i^T \beta + \varepsilon_i; \quad \varepsilon_i \sim N(0, \sigma^2); i = 1, \ldots, n, \tag{1}$$

  where each predictor corresponds to a node in a given undirected graph, and an edge of the graph indicates possible clustering between two predictors.
- Identifying two kinds low-dimensional structures simultaneously:
  - Supervised clustering: Estimate homogenous and collapsing clusters of predictors.
  - Feature selection: Estimate nonzero coefficients of predictors.

Motivating example

- Identifying subnetworks relevant to breast cancer survival.
- Y = log survival time, X: Clinical variables as well as gene expression profiles.
- Gene network: protein-protein interaction network (Chang, et al., 07), and describes dependency structure of genes.



Simultaneous supervised clustering & feature selection

- Homogeneity: Partition $\{1, \ldots, p\}$ into clusters:

$$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \approx (0\mathbf{1}_{|\mathcal{G}_0|}, \alpha_1 \mathbf{1}_{|\mathcal{G}_1|}, \ldots, \alpha_K \mathbf{1}_{|\mathcal{G}_K|})'.$$

- Goal: Over the graph, estimate true $\mathcal{G}^0 = (\mathcal{G}^0_0, \mathcal{G}^0_1, \ldots, \mathcal{G}^0_{K_0})'$ & $\boldsymbol{\alpha}^0 = (0, \alpha^0_1, \ldots, \alpha^0_{K_0})'$.
- Benefits
  - Structure: Explore sparseness & clustering by leveraging dependency structures given by a graph.
  - Estimation: Higher accuracy is due to variance reduction.
  - Selection: Overcome feature selection instability by grouping and collapsing highly positively correlated predictors, & remove redundant clusters by feature selection. Higher accuracy for both.
  - Interpretability: Simpler model with higher predictive power.

Objectives and Challenges

- Objectives
  - Reconstructing biased OLS based on $\mathcal{G}^0$.
  - Accurate identification of clusters & parameter estimation/prediction.
  - Developing an efficient computational algorithm for large problems.
- Challenges
  - More difficult than the problem of feature selection alone & supervised clustering alone.
  - Complexity for enumeration over a complete graph is the Bell number: $B(p) = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^p}{k!} \approx O\left(e^{e^{p^a}}\right)$.

Literature

- Graph:
  - Fused Lasso(TSRZK, 05): $\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j - \beta_{j+1}|$. (QP).
  - LL(08): $\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j\sim j'}^{p} (\frac{\beta_j}{w_j} - \frac{\beta_{j+1}}{w_{j+1}})^2$. (QP).
  - TT(11): Glasso: $\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j\sim j'}^{p} |\beta_j - \beta_{j'}|$. (Homotopy).
- Non-graph:
  - SH (10) $\lambda \sum_{j,j'=1}^{p} J(|\beta_j - \beta_{j'}|)$. (Homotopy)
  - JKLDY (11) $\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j,j'=1}^{p} |\beta_j - \beta_{j'}|$. (QP).
- Other types: Clustering in sizes: BR (08), XPS(09). Not a Glasso problem.
- Huge literature for feature selection and encouraging grouping in feature selection.......

Challenges

- Computation: For large problems, QP is infeasible, and a homotopy method may be inefficient. Coordinate decent method breaks down even for F-lasso (special algorithm). Need efficient methods for large p over an arbitrary undirected graph
- Theory: F-lasso: Rinaldo (2009); clustering: SH (10). Lack of theory to guide practice.

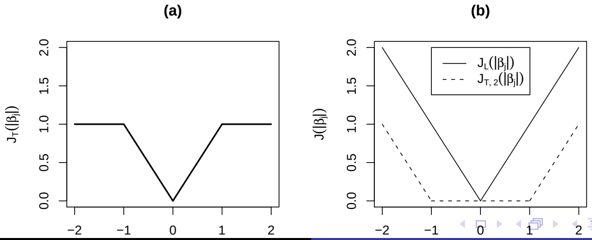
Constrained Least Squares

- Constrained least squares criterion over a graph:

$$\frac{1}{2n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\beta)^2, \quad \text{subj to} \sum_{j=1}^{p} J(|\beta_j|) \le s_1,$$

$$\sum_{j<j':(j,j')\in\mathcal{E}} J(|\beta_j - \beta_{j'}|) \le s_2,$$

- Surrogate of the L_0: $J(z) = \min\left(\frac{|z|}{\lambda_3}, 1\right)$.
- Tuning: (s_1, s_2, λ_3). Clustering: (s_1, s_2); Threshold: $\lambda_3 > 0$.



Nonconvex minimization

- Theorem: A global minimizer of constrained LS is a local minimizer of $S(\boldsymbol{\beta}) = (2n)^{-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^{p} J(|\beta_j|)\lambda_2 \sum_{j<j':(j,j')\in\mathcal{E}} J(|\beta_j - \beta_{j'}|)$, where $\lambda_j \to s_j$; $j = 1, 2$.
- Local optimality: $j = 1, \cdots, p$,

$$-(\boldsymbol{x}^{(j)})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_1}{\lambda_3} b_j + \frac{\lambda_2}{\lambda_3} \sum_{j':(j,j')\in\mathcal{E}} b_{jj'} = 0, \qquad (2)$$

  where $b_j = sign(\beta_j)I(|\beta_j| < \lambda_3)$ if $\beta_j \neq 0$, $b_{jj'} = sign(\beta_j - \beta_{j'})I(|\beta_j - \beta_{j'}| < \lambda_3)$ if $\beta_j - \beta_{j'} \neq 0$, & $b_j = \emptyset$ if $|\beta_j| = \lambda_3$, & $b_{jj'} = \emptyset$ if $|\beta_j - \beta_{j'}| = \lambda_2$, are the regular subdifferentials of $\min(|\beta_j|, \lambda_3)$ & $\min(|\beta_j - \beta_{j'}|, \lambda_3)$ at β_j.
- Strategy: Obtaining a local minimizer of (2) through DC programming, the augmented Lagrangian and coordinate decedent methods.

Difference Convex Programming

- Decomp $S = S_1 - S_2$ into a diff of two convex functions.
- Construct a sequence of upper convex approximations iteratively by replacing S_2 at iteration m, by its affine minorization at iteration $m-1$:

$$S^{(m)}(\beta) = (2n)^{-1}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i^T\beta)^2 + \lambda_1 \sum_{j:j\in\mathcal{F}^{(m-1)}} |\beta_j| + \lambda_2 \sum_{j<j':(j,j')\in\mathcal{E}^{(m-1)}} |\beta_j - \beta_{j'}|, \quad (3)$$

- $\mathcal{E}^{(m-1)} = \{(j,j') \in \mathcal{E}, |\hat{\beta}_j^{(m-1)} - \hat{\beta}_{j'}^{(m-1)}| < \lambda_3\}$ & $\mathcal{F}^{(m-1)} = \{j \in \mathcal{F}, |\hat{\beta}_j^{(m-1)}| < \lambda_3\}$.
- $\hat{\beta}^{(m-1)}$ is the minimizer of $S^{(m-1)}(\beta)$.
- $\hat{\beta}$: DC estimate $\hat{\beta}^{(m)}$ after convergence.

Solution of (3)

- Major challenges:
  - (Stationary points) Coordinate decent method fails for (3).
  - (Graph structure+overcomplete) A graph can be arbitrary.
- Efficient method: Introduce slack variables $\beta_{jj'} = \beta_j - \beta_{j'}$ for $j \neq j'$ for an equivalent augmented problem of (3) in $\zeta = (\beta_1, \cdots, \beta_p, \beta_{12}, \ldots, \beta_{1p}, \ldots, \beta_{(p-1)p})^T$:

$$\tilde{S}^{(m)}(\zeta) = (2n)^{-1} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda_2 \sum_{j<j':(j,j') \in \mathcal{E}^{(m-1)}} |\beta_{jj'}| + \lambda_1 \sum_{j:j \in \mathcal{F}^{(m-1)}} |\beta_j|, \quad (4)$$

subj to linear constraints $\boldsymbol{A}\zeta = \boldsymbol{0}$.

Augmented Lagrangian + coordinate decent for (4)

- Augmented Lagrangian for (4) by solving its unconstrained version iteratively: At iteration t, minimize

$$\bar{S}^{(m)}(\zeta) = \tilde{S}^{(m)}(\boldsymbol{\beta}) + \sum_{j<j':(j,j')\in\mathcal{E}} \tau_{jj'}^{(t)}(\beta_j - \beta_{j'} - \beta_{jj'}) + \frac{1}{2}\sum_{j<j':(j,j')\in\mathcal{E}} \nu_{jj'}^{(t)}(\beta_j - \beta_{j'} - \beta_{jj'})^2, \quad (5)$$

- $(\tau_{jj'}^{(t)}, \nu_{jj'}^{(t)})$ are Lagrangian multipliers for $\boldsymbol{A}\zeta = \mathbf{0}$ and for expediting convergence.
- Solve (5) through analytic updating formula and coordinate decent method.

Path-following Algorithm

For given λ_3,

Step 1 (Initialization) Specify evaluation points for tuning parameters. Supply a good initial estimate $\hat{\beta}^{(0)}$. Set tolerance error for convergence,& $\tau_{jj'}^{(0)} = 1$ & $\nu_{jj'}^{(0)} = 1$.

Step 2 (Iteration) Iteration begins with $m = 1$. At iteration m, compute $\hat{\beta}^{(m)}$ by solving (5) through coordinate descent over active sets.

Step 3 (Stopping) Terminate when $S(\hat{\beta}^{(m-1)}) - S(\hat{\beta}^{(m)}) \leq 0$. The estimate $\hat{\beta} = \hat{\beta}^{(m_0)}$, where m_0 is the termination index.

Computational properties

- DC programming converges fast and finitely. This is due to the three non-differentiable points of $J(\cdot)$.
- The augmented Lagrangian method converges super-linearly.
- The coordinate descent method is efficient when integrated with the augmented Lagrangian method.
- Can handle a problem of size $p = 3000 - 4000$ easily—complexity for constraint terms is order of p^2.

Notation

- $C_{\min}$: $\inf_{\mathcal{G}\neq\mathcal{G}^0} \frac{1}{n}\|(I - \boldsymbol{P}_{\mathcal{G}\setminus\mathcal{G}_0^0})\boldsymbol{X}_{\mathcal{G}\setminus\mathcal{G}_0^0}\boldsymbol{\beta}_{\mathcal{G}\setminus\mathcal{G}_0^0}\|^2$ for $\mathcal{G}$ induced by $\mathcal{E}$, $\boldsymbol{P}$ is the projection for collapsed predictors over $\mathcal{G}\setminus\mathcal{G}_0^0$. $C_{\min}$: describes the least favorable situation in the KL-loss.
- $\boldsymbol{X}_{\mathcal{G}\setminus\mathcal{G}_0^0}$ & $\boldsymbol{\beta}_{\mathcal{G}\setminus\mathcal{G}_0^0}$: design matrix of predictors & coefficient vector over $\mathcal{G}\setminus\mathcal{G}_0^0$, $\|\cdot\|$ is the Eucli-norm in $\mathcal{R}^n$.
- Oracle estimator $\hat{\beta}^{ols}$: $\left(\hat{\beta}_1^{ols},\ldots,\hat{\beta}_p^{ols}\right)^T = \left(0_{|\mathcal{G}_0^0|},\hat{\alpha}_1^{ols}1_{|\mathcal{G}_1^0|},\ldots,\hat{\alpha}_{K^0}^{ols}1_{|\mathcal{G}_{K^0}^0|}\right)^T$ given $\mathcal{G}^0$; $\hat{\boldsymbol{\alpha}}^{ols}\equiv\left(\hat{\alpha}_1^{ols},\ldots,\hat{\alpha}_{K^0}^{ols}\right)^T = \left(\boldsymbol{X}^T_{\mathcal{G}^0\setminus\mathcal{G}_0^0}\boldsymbol{X}_{\mathcal{G}^0\setminus\mathcal{G}_0^0}\right)^{-1}\boldsymbol{X}^T_{\mathcal{G}^0\setminus\mathcal{G}_0^0}\boldsymbol{Y}$.
- Graph parameters: $\bar{K}$–max # clusters allowed; S^*–number of possible distinct clusters, for the given graph.

Global minimizer

Let $\hat{\beta}^{gl}$ be a global minimizer of constrained LS. Let p_0 be $\#$ non-zero predictors.

Theorem

Assume path connectivity for $j, j' \in \mathcal{G}_k^0$ over $\mathcal{E}$. If $(s_1, s_2) = (p - p_0, \sum_{(j,j')\in\mathcal{E}} I(|\beta_j^0 - \beta_{j'}^0| \neq 0)$, $\lambda_3 \leq 2\sigma\sqrt{\frac{\log p}{2np^3\lambda_{\max}\left(\boldsymbol{X}^T\boldsymbol{X}\right)}}$, *then*

$$P\left(\hat{\beta}^{gl} \neq \hat{\beta}^{ols}\right) \leq \exp\left(-\frac{n}{10\sigma^2}\left(C_{\min} - 10\sigma^2\frac{2\log p + \bar{K} + 2\log S^*}{n}\right)\right).$$

Under condition: $C_{min} \geq d_1\sigma^2\frac{2\log p + \bar{K} + 2\log S^*}{n}$ for $d_1 > 10$, there exist tuning parameter values such that oracle properties (A)-(D) hold.

Global minimizer–continued

Estimate: $\hat{\beta} = \hat{\beta}_{\hat{\mathcal{G}}}$; $\hat{\mathcal{G}} = (\hat{\mathcal{G}}_0 = \mathbf{0}, \cdots, \hat{\mathcal{G}}_K)$. Truth: $\beta^0 = \beta^0_{\mathcal{G}^0}$, $\mathcal{G}^0 = (\mathcal{G}^0_0 = \mathbf{0}, \cdots, \mathcal{G}^0_{K^0})$. As $n, p \to \infty$,

(A) (Clustering consistency) $P(\hat{\mathcal{G}} \neq \mathcal{G}^0) \to 0$.

(B) (Parameter estimation) For any β^0,

$$n^{-1} E \|\boldsymbol{X}(\hat{\beta} - \beta^0)\|^2 = \sigma^2 \frac{K_0}{n}.$$

(C) (Normality) With $W_{\hat{A}} = \sigma^2 (\boldsymbol{X}^T_{\hat{\mathcal{G}} \setminus \hat{\mathcal{G}}^0} \boldsymbol{X}_{\hat{\mathcal{G}} \setminus \hat{\mathcal{G}}^0})^{-1}$; $I_{\mathcal{G}_0 \setminus \mathcal{G}^0_0}$ the identity matrix

$$W^{-1/2}_{\hat{\mathcal{G}} \setminus \hat{\mathcal{G}}^0} (\hat{\beta}_{\hat{\mathcal{G}} \setminus \hat{\mathcal{G}}^0} - \beta^0_{\hat{\mathcal{G}} \setminus \hat{\mathcal{G}}^0}) \sim N(0, I_{\mathcal{G}_0 \setminus \mathcal{G}^0}).$$

(D) (Uniformity) (A)-(C) hold uniformly over $B_0(u, l) \equiv \{\beta \in \mathcal{R}^p : K \leq u, C_{\min} \geq l\}$: a L_0-ball of radius $u > 0$ & resolution level $l > 0$. Then $\hat{\beta}$ is asym minimax.

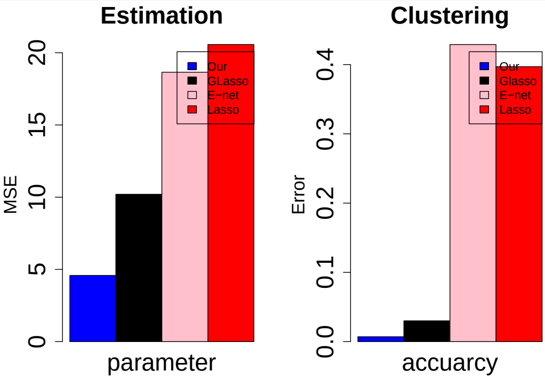
Comments and Local minimizer

- $\bar{K}$ and S^* need to be computed for a given graph. Small graph: Fused, $\bar{K} \leq K_0$; $S^* \leq p_0^{K_0}$, K_0 is the number of true clusters.
- Under smaller conditions, any local minimizer, particularly the one computed from the algorithm, has oracle properties (A)-(D).

Numerical examples

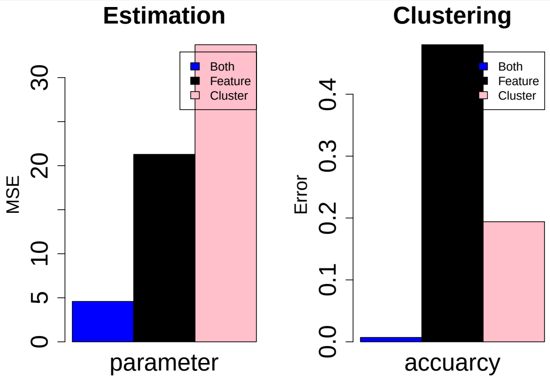
- Ex: (Graph, Li & Li, 08). Consider a network consisting of 200 subnetworks, each with one transcription factor (TF) and its 10 regulatory target genes. In (1), a predictor of each target gene and the TF follows a bivariate normal distribution with correlation $\rho = 0.7$, and target genes are independent $N(0,1)$'s, conditional on the TF; $n = 100$, $p = 2200$, $\sigma^2 = \sum_{j=1}^{p} \beta_j^2/2$.
- $\boldsymbol{\beta} =$
$$(5, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{10}, -5, \underbrace{\frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}}}_{10}, 3, \underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{10}, -3, \underbrace{\frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}}_{10}, \underbrace{0, \ldots, 0}_{p-44})^T.$$
- Mean squares error: averaged over 100 replications.
- Selection error & grouping error:. % committing an error.
- Tuning parameters: are estimated by minimizing MSE over a set of grid points.
- Comparison: GLasso, Elastic Net, & Lasso.

Comparison: Example



- Our method outperforms in both parameter estimation and accuracy of selection. Interestingly, Elastic net performs better than Lasso in parameter estimation but worst than selection.
- The average number of DC iterations is about 4.

Simultaneous supervised clustering & feature selection



- Simultaneous supervised clustering & feature selection performs better than either one alone. Perform relatively better when p gets large.

Network-based eQTL analysis

- Goal: Identify genomic loci (Expression quantitative trait loci , called eQTLs) linked to gene expression traits. Improve power in detecting eQTLs for a group of co-regulated genes.
- Mouse dataset in Lan et al. (06): 60 F2 mice from B6 and BTBR founder strains, where the B6 and BTBR strains are diabetes resistant and non-resistant, respectively. About 45000 gene expression traits are measured with genotypes of 194 markers distributed across the mouse genome with an average marker interval of approximately 10cm.

Network-based eQTL analysis-continued

- Model: $\boldsymbol{Y}_g$: expression of gene g of 60 mice, linking to $\boldsymbol{X}_0$:

$$\boldsymbol{Y}_g = \boldsymbol{X}_0\boldsymbol{\beta}_g + \boldsymbol{\epsilon}_g; \quad \epsilon_g \sim N(0, \sigma^2\boldsymbol{I}); \quad g = 1, \ldots, G, \tag{6}$$

  where X_0 is genotypes of 194 markers across 60 mouse genomes, with each genotype taking $-1, 0, 1$ indicating one of three alleles.
- Estimation of nonzero coefficients $\boldsymbol{\beta}_g$.
- GPCR (G protein-coupled receptor) co-expression subnetwork of 17 positively correlated genes based on Ghazalour et al. (06).

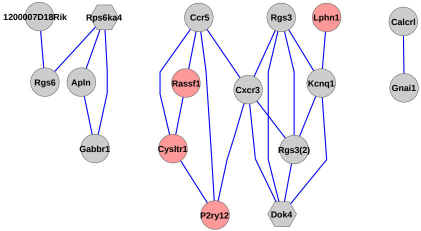
Network-based eQTL analysis-continued

- It is biologically reasonable to assume that genes connected in a co-expression network are likely to share some common eQTLs; i.e, if two genes are connected, their expressions are likely to be associated with the genotypes at some common genomic loci.
- Combined model:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \quad \epsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

- Our method identifies 13 out 17 genes, as opposed to 2 by Lasso. The result is in agreement with that in previous studies (Lan et al., 06).

Network-based eQTL analysis-continued



Take Away Messages

- Supervised clustering and feature selection can reduce estimation variance while retaining the roughly the same amount of bias, leading to better predictive accuracy.
- The method identifies and collapses highly positive correlated predictors in a process of selection.
- Further develop methods for time varying networks.
- Study other types of clustering, e.g., coefficients of similar size not value, which involves the absolute values.