Grouping Pursuit in regression

Xiaotong Shen

School of Statistics
University of Minnesota

Email: xshen@stat.umn.edu

Joint with Hsin-Cheng Huang (Sinica, Taiwan)

Workshop in honor of John Hartigan: Innovation and Inventiveness in Statistics Workshop, May 15-17, Yale

Introduction

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

$$Y_i \equiv \mu(\boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{1}$$

  where $\mu(\boldsymbol{x}_i) \equiv \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\varepsilon_i \sim N(0, \sigma^2)$.

- Goal: Identify all potential groupings for optimal predication of Y, especially when $p >> n$.
- *Grouping pursuit* amounts to estimating grouping $\mathcal{G}^0 = (\mathcal{G}_1^0, \ldots, \mathcal{G}_K^0)^T$ as well as $\boldsymbol{\alpha}^0 = (\alpha_1^0, \ldots, \alpha_K^0)^T$ given $\mathcal{G}^0$ when true $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_p^0)^T =$ or $\approx (\alpha_1^0 1_{|\mathcal{G}_1^0|}, \ldots, \alpha_K^0 1_{|\mathcal{G}_K^0|})^T$ with $1_{|\mathcal{G}_1^0|}$ denoting a vector of 1's with length $|\mathcal{G}_1^0|$.

Grouping pursuit

- *Essential* to high-dimensional analysis is seeking a certain low-dimensional structure.
  - Homogenous subgroups. Variable selection seeks only two homogenous groups: zero-coefficient group vs non-zero-coefficient group.
  - Projection pursuit, $\cdots$.
- *Main idea*: Group coefficients of roughly the same value or size.
- *Benefits*: Variance reduction, which goes beyond variable selection. Simpler model with higher predictive power. Can be thought of as one kind of supervised clustering.
- Challenges: Complexity for identifying the best grouping is the pth order *Bell number*: $B_p = \frac{1}{e}\sum_{k=0}^{\infty}\frac{k^p}{k!}$–order $e^{e^{p^a}}$ for some $0 < a < 1$.

Relevant literature and motivation

- Literature:
  - Grouping in series order (F-Lasso, TSRZK, 05): $\sum_{j=1}^{p} |\beta_j - \beta_{j+1}|$.
  - Grouping in size (Bondell & Reich, 08): $\sum_{i<j} \max(|\beta_i|, |\beta_j|)$.
  - Grouping pursuit is one kind of supervised clustering,.....
- Motivating example:

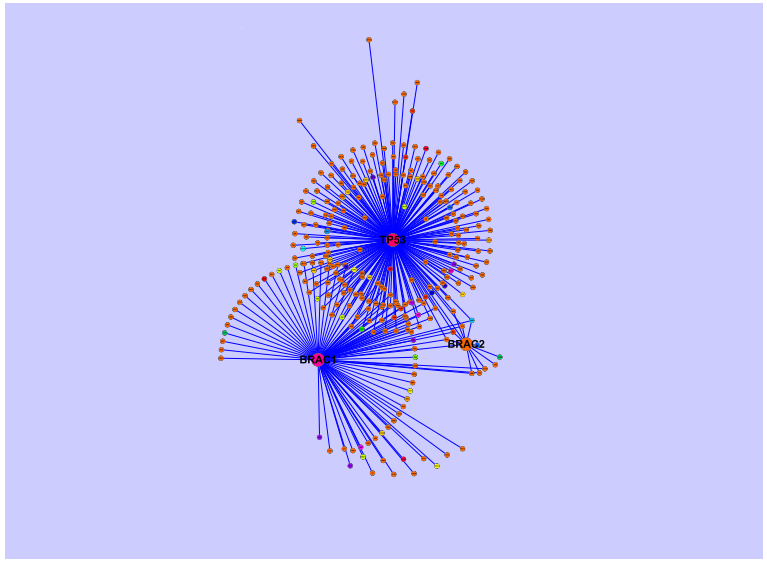


Figure 1: Plot of the PPI gene subnetwork for breast cancer data

Grouping

- Enumeration
  - Partition $\{1, \cdots, p\}$ into $\mathcal{G} = (G_1, \cdots, G_k)$. Given $\mathcal{G}$, compute OLS through regression of $\boldsymbol{Y}$ on grouped $Z_{G_1} \equiv \boldsymbol{X}_{\mathcal{G}_1} 1, \cdots, Z_{G_k} \equiv \boldsymbol{X}_{\mathcal{G}_k} 1$.
  - Choose the best grouping from all possible groupings.
  - Computation is infeasible, i.e., $p = 10$ requires *115975* enumerations (Bell number)—much worst than that in variable selection.
- Our objectives
  - Accurate grouping.
  - Computational efficiency.
  - Reconstruction of *true grouping* & *unbiased OLS* based on it simultaneously.

Grouping pursuit–our approach

- Regularization through designed nonconvex penalty

$$S(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda_1 J(\boldsymbol{\beta}); J(\boldsymbol{\beta}) = \sum_{j<j'} G(\beta_j - \beta_{j'}), \tag{2}$$

  where $\lambda_1 > 0$ is a regularization parameter, $G(z) = \lambda_2$ if $|z| > \lambda_2$ and $G(z) = |z|$ otherwise, and $\lambda_2 > 0$ is a thresholding parameter.

- Role of $G(z)$
  - Piecewise linear for computational advantage through *grouped subdifferentials* and *difference convex (DC) programming*.
  - Three non-differentiable points: (a) $z = 0$ for grouping pursuit; (b) $z = \pm\lambda_2$ for computation and for theoretical advantages.

Grouped subdifferentials

- Subdifferential of convex $S(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$ is the set of all subgradients at $\boldsymbol{\beta}$.
- Subgradient of $|\beta_j - \beta_{j'}|$ wrt β_j at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is $b_{jj'}(\boldsymbol{\lambda})$.

$$= \begin{cases} \text{Sign}(\hat{\beta}_j(\boldsymbol{\lambda}) - \hat{\beta}_{j'}(\boldsymbol{\lambda})) & \text{if } 0 < |\hat{\beta}_j(\boldsymbol{\lambda}) - \hat{\beta}_{j'}(\boldsymbol{\lambda})| \\ |b_{jj'}(\boldsymbol{\lambda})| \leq 1 & \text{if } \hat{\beta}_j(\boldsymbol{\lambda}) - \hat{\beta}_{j'}(\boldsymbol{\lambda}) = 0. \end{cases}$$

- Due to overcompleteness of the penalty, $b_{jj'}(\boldsymbol{\lambda})$ can not be estimated.
- *Subgradient of j wrt group $\mathcal{G}_k(\boldsymbol{\lambda})$*: $B_j(\boldsymbol{\lambda}) \equiv \sum_{j' \in \mathcal{G}_k(\boldsymbol{\lambda}) \setminus \{j\}} b_{jj'}(\boldsymbol{\lambda})$, with $\sum_{j \in \mathcal{G}_k(\boldsymbol{\lambda})} B_j(\boldsymbol{\lambda}) = 0$, because $b_{jj'} = -b_{j'j}$ for $j \neq j'$.
- *Subgradient of subset A wrt group $\mathcal{G}_k(\boldsymbol{\lambda})$*:
  $B_A(\boldsymbol{\lambda}) \equiv \sum_{j \in A} B_j(\boldsymbol{\lambda}) = \sum_{(j,j') \in A \times (\mathcal{G}_k(\boldsymbol{\lambda}) \setminus A)} b_{jj'}(\boldsymbol{\lambda})$, with
  $|B_A(\boldsymbol{\lambda})| \leq |A| \big(|\mathcal{G}_k(\boldsymbol{\lambda})| - |A| \big)$.

Solution surface via DC programming

- *Decompose* $S(\boldsymbol{\beta})$ in (2) into a difference of two convex functions $S_1(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \lambda_1 \sum_{j<j'} |\beta_j - \beta_{j'}|$ and $S_2(\boldsymbol{\beta}) = \lambda_1 \sum_{j<j'} G_2(\beta_j - \beta_{j'})$, through a DC decomposition of $G(\cdot) = G_1(\cdot) - G_2(\cdot)$ with $G_1(z) = |z|$ & $G_2(z) = (|z| - \lambda_2)_+$.

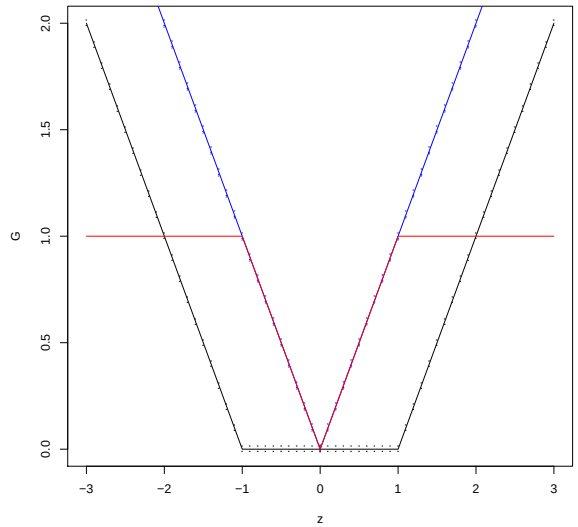


Figure 2: DC decomposition of $G(z)$.

Solution surface via DCP, continued

- Linearize $S_2(\boldsymbol{\beta})$ at iteration m by its affine minorization from iteration $m-1$, leading to an upper convex approximating function at iteration m:

$$S^{(m)}(\boldsymbol{\beta}) = S_1(\boldsymbol{\beta}) - S_2(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda})) - (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda}))^T \nabla S_2(\hat{\boldsymbol{\beta}}^{(m-1)}(\boldsymbol{\lambda})), \quad (3)$$

  - ∇ : the subgradient operator; $\hat{\beta}_k^{(m-1)}(\boldsymbol{\lambda})$: minimizer of (3) at iteration $m-1$.

- Solve (3) iteratively until it converges.

- No need to seek global solution—DC solution has desired optimality of a global solution in grouping (Theorem), and can be computed much efficiently (Theorem).

Homotopy method+DCP

- *Key*: homotopy via subdifferentials and DCP for solution $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ of (3).
  - *Optimality* through subdifferentials: $\nabla S^{(m)}(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})} = 0$.
- *Major challenges*: (1) (Discontinuity) $\hat{\boldsymbol{\beta}}^{(m)}(\boldsymbol{\lambda})$ may contain jumps in $(\boldsymbol{Y}, \lambda_2)$; (2) (Overcompleteness) computing $B_j^{(m)}(\boldsymbol{\lambda})$ via enumerations over $\{b_{jj'}\}$ is infeasible, (Bell number).
- Homotopy (1) piecewise linear and continuous in λ_1 given $(\boldsymbol{Y}, \lambda_2)$ with piecewise linear penalty and designed support point. Transition conditions: (1) Combining groups $\mathcal{G}_k(\boldsymbol{\lambda})$ with $\mathcal{G}_k(\boldsymbol{\lambda})$: $\alpha_k(\boldsymbol{\lambda}) = \alpha_l(\boldsymbol{\lambda})$; (2) Splitting group $\mathcal{G}_k(\boldsymbol{\lambda})$: $|B_A(\boldsymbol{\lambda})| \leq |A|\big(|\mathcal{G}_k(\boldsymbol{\lambda})| - |A|\big)$.
- Overcompleteness: Use piecewise linear property of $B_j^{(m)}(\boldsymbol{\lambda})$ for searching (order: $O(p^2 \log p)$).
- Homotopy Algorithm for computing $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ as a function of $\boldsymbol{\lambda}$ simultaneously.

Homotopy method+DCP, continued

Property: terminate finitely and converge rapidly. Control at one $\boldsymbol{\lambda}_0$ implies the entire surface.

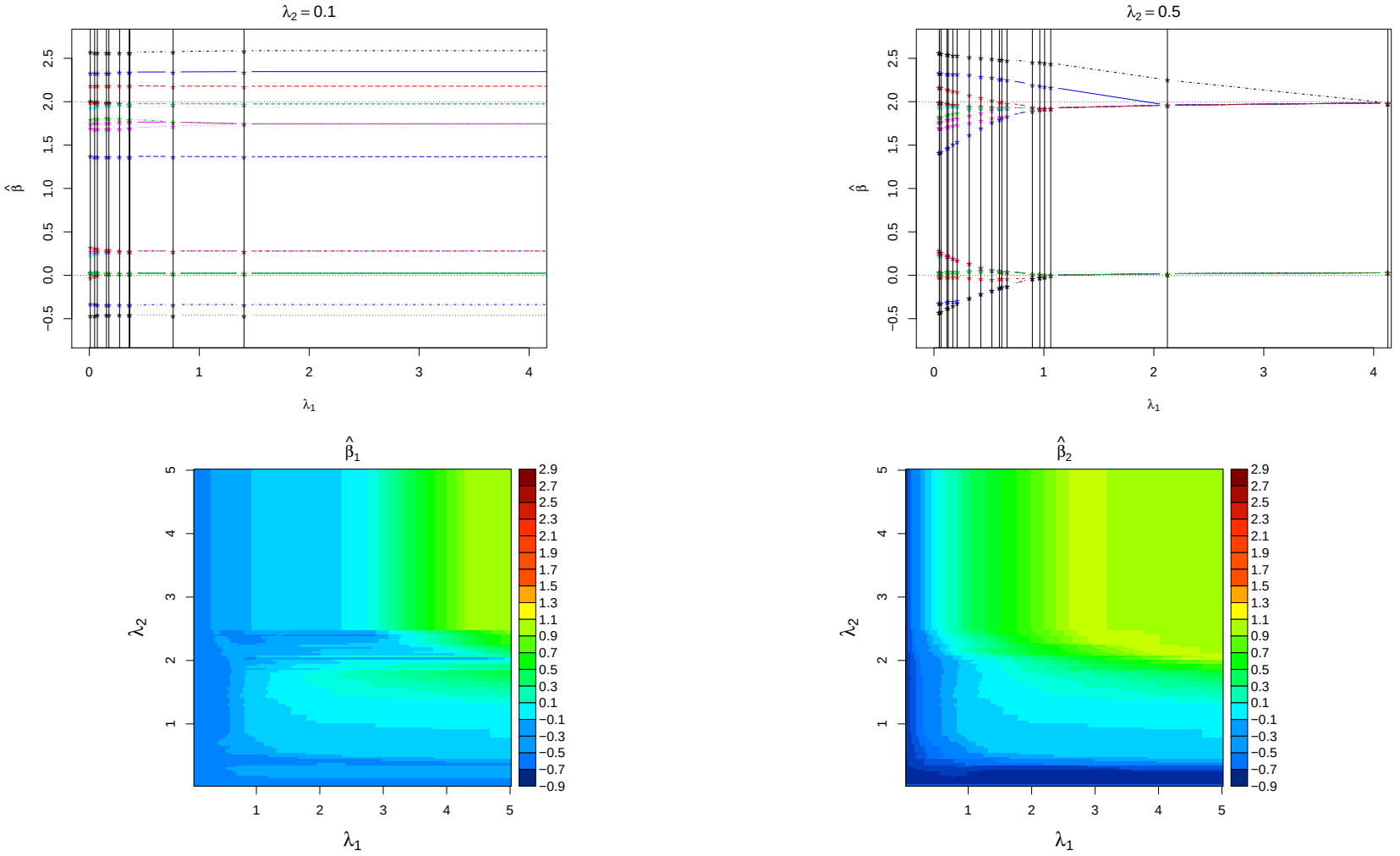


Figure 3: Regularization solution path/surface.

Model selection for prediction

- Model selection:

$$\widehat{\text{GDF}}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) = \frac{1}{2n}\sum_{i=1}^{n}(Y_i - \hat{\mu}(\boldsymbol{\lambda}, \boldsymbol{x}_i)))^2 + \frac{1}{n}\sigma^2\widehat{df}(\boldsymbol{\lambda}), \qquad (4)$$

  - For smooth $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ ($m = 0$), $\widehat{df}(\boldsymbol{\lambda}) = K(\boldsymbol{\lambda})$ for fast computation, c.f., SURE (Stein, 1981).
  - For piecewise smooth $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ ($m > 0$),

    $\widehat{df}(\boldsymbol{\lambda}) = \frac{\sigma^2}{\tau^2}\sum_{i=1}^{n}$ Cov$^*(Y_i, \hat{\mu}^*(\boldsymbol{\lambda}, \boldsymbol{x}_i))$ and Cov$^*(Y_i, \hat{\mu}^*(\boldsymbol{\lambda}, \boldsymbol{x}_i))$, through *data perturbation* (GSURE, Shen & Ye, 2002).

Theory: Error analysis

- *Performance for grouping pursuit:*
  - Error: (Disagreement) $P(\mathcal{G}(\boldsymbol{\lambda}) \neq \mathcal{G}^0) \leq P(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \neq \hat{\boldsymbol{\beta}}^{(ols)})$.

  $\mathcal{G}(\boldsymbol{\lambda}), \mathcal{G}^0$: estimated and true grouping (uniquely defined). $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \hat{\boldsymbol{\beta}}^{(ols)}$: estimator defined by Algorithm 2 and OLS based on $\mathcal{G}^0$.

  Theorem: $P(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \neq \hat{\boldsymbol{\beta}}^{(ols)})$ is upper bounded by

$$\frac{K(K-1)}{2}\Phi\Big(\frac{-n^{1/2}\big(\gamma_{\min}-\lambda_2\big)}{2\sigma c_{\min}^{-1/2}}\Big) + p\,\Phi\Big(\frac{-n\lambda_1}{\sigma \max_{1\leq j\leq p}\|\boldsymbol{x}_j\|}\Big). \qquad (5)$$

- $\Phi(z)$: CDF of $N(0,1)$.

  $\|\boldsymbol{x}_j\|$: L_2-norm of $\boldsymbol{x}_j$.

  $\gamma_{\min}$: $\min\{|\alpha_k^0 - \alpha_l^0| > 0 : 1 \leq k < l \leq K\}$.

  $c_{\min}$: smallest eigenvalue of $\boldsymbol{Z}_{\mathcal{G}^0}^T \boldsymbol{Z}_{\mathcal{G}^0}/n$.

  K: # of estimated groups, which is no larger than $\min(n,p)$.

Theory: Error analysis, continued

If $\max\left\{\frac{nc_{\min}(\gamma_{\min}-\lambda_2)^2}{8\sigma^2}, \frac{n\lambda_1^2}{2\sigma^2 \max\limits_{1\le j\le p}\|\boldsymbol{x}_j\|^2/n}\right\} - \log p \to \infty,$

- *Grouping Consistency*

$$P(\mathcal{G}(\boldsymbol{\lambda}) \neq \mathcal{G}^0) \le P(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \neq \hat{\boldsymbol{\beta}}^{(ols)}) \to 0, \quad p, n \to +\infty.$$

- *Remarks:*
  - Roughly: $p < \exp(O(n\lambda_1^2))$, $\lambda_1 \to 0$, $n^{1/2}\lambda_1 \to \infty$, $nc_{min}(\gamma_{min}-\lambda_2) \to \infty$. ($\max\limits_{j:1\le j\le p}\|\boldsymbol{x}_j\|^2/n$ bounded, satisfied by standardization).
  - Note that $c_{\min}$ can be independent of (p,n) or $c_{\min} \to 0$ as $p, n \to \infty$, depending on if K increases in (p,n), even though the true model is independent of (p,n).
  - A less sharp bound can be derived under a moment assumption of ε_1.

Theory: Grouping

Let $r_j(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) = \boldsymbol{x}_j^T\left(\boldsymbol{Y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})\right)$, which becomes the sample correlation between $\boldsymbol{x}_j$ and the residual, after standardization of $\{\boldsymbol{x}_j : j = 1, \ldots, p\}$.

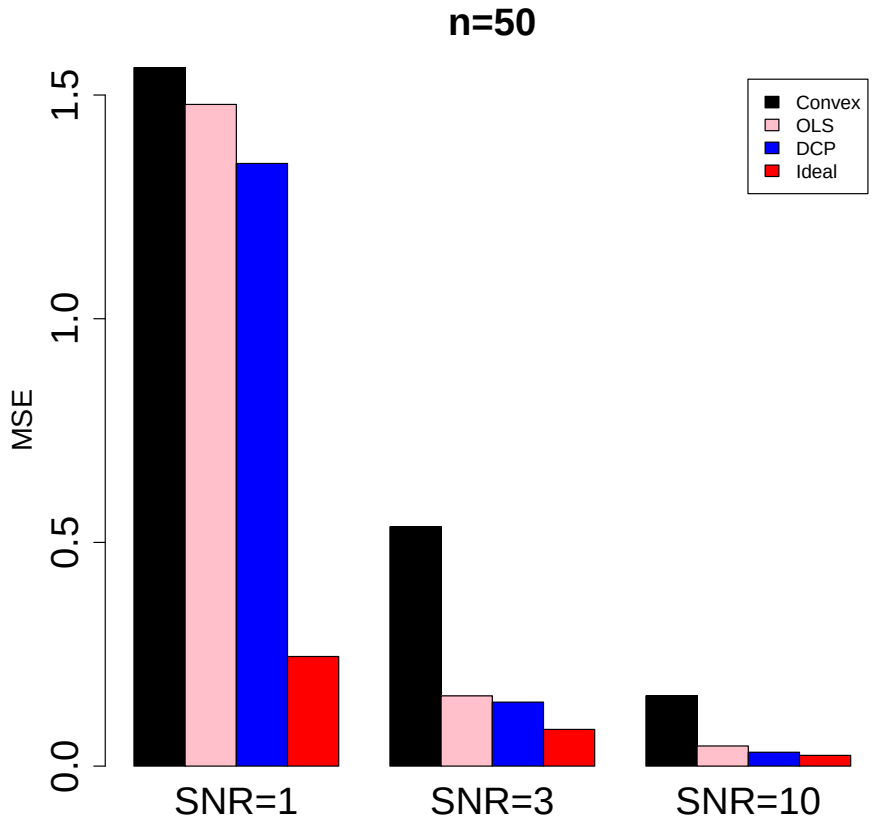
Theorem: (Grouping) For any $j = 1, \ldots, p$, $j \in \mathcal{G}_k(\boldsymbol{\lambda})$ if $|r_j(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})) - n\lambda_1\delta_k(\boldsymbol{\lambda})| \leq n\lambda_1(|\mathcal{G}_k(\boldsymbol{\lambda})| - 1); k = 1, \ldots, K(\boldsymbol{\lambda})$. Here $\delta_k(\boldsymbol{\lambda}) = \delta^{(m*)}(\lambda)$ and $\delta^{(m)}(\lambda)$ is defined in Theorem 1.

- Predictors with similar values of correlations are grouped together, as characterized by intervals $\cup_{k=1}^{K(\boldsymbol{\lambda})}\Big(n\lambda_1\delta_k(\boldsymbol{\lambda}) - n\lambda_1(|\mathcal{G}_k(\boldsymbol{\lambda})| - 1), n\lambda_1\delta_k(\boldsymbol{\lambda}) + n\lambda_1(|\mathcal{G}_k(\boldsymbol{\lambda})| - 1)\Big)$.

Numerical examples

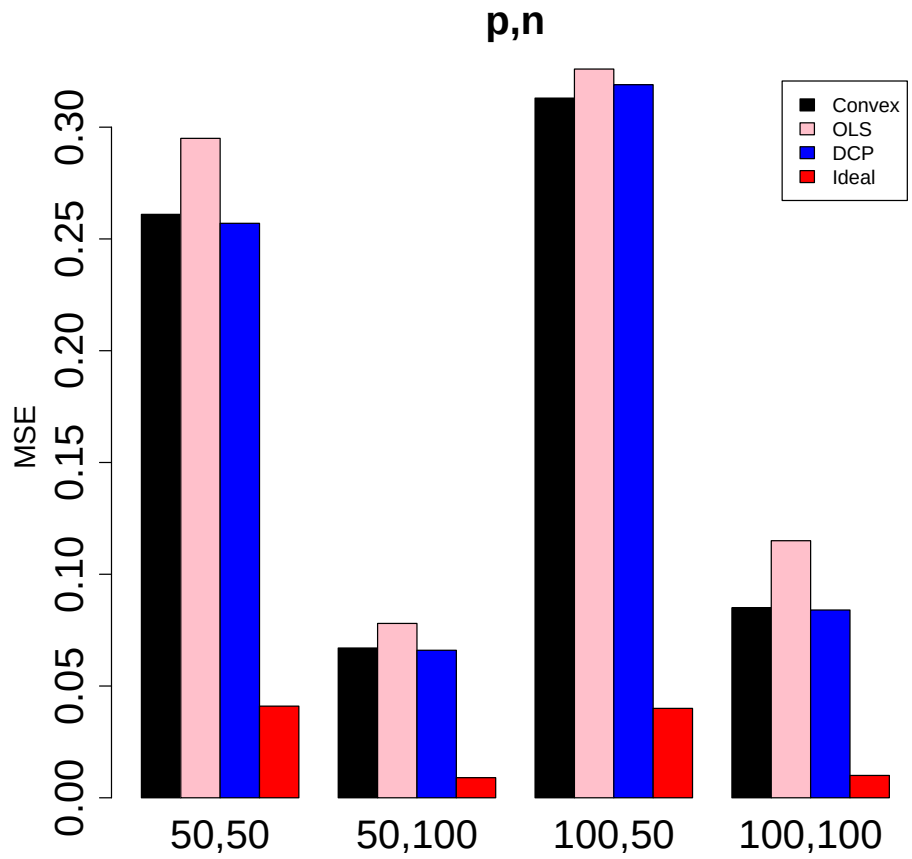
- Ex1: (Sparse grouping). In (2), $\varepsilon_i \sim N(0, \sigma^2)$ and σ^2 according to SNR; $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma}_{p\times p})$ with $n = 50$, $p = 20$ and diagonal/off-diagonal elements $1/0.5$;
  $\boldsymbol{\beta} = (\underbrace{0, \ldots, 0}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{5}, \underbrace{2, \ldots, 2}_{5})^T$.
- Ex2: (Large p but small n). In(2), $\varepsilon_i \sim N(0, \sigma^2)$, with $SNR = 10$; $\boldsymbol{x}_i \sim N(0, \boldsymbol{\Sigma})$ with $0.5^{|j-k|}$ the jk-th element of $\boldsymbol{\Sigma}$. Here
  $$\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{5}, \underbrace{-1.5, \ldots, -1.5}_{5}, \underbrace{1, \ldots, 1}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{180})^T.$$
- *Mean squares error:* averaged over 100 replications.
- *Tuning:* $\boldsymbol{\lambda}$ is estimated by minimizing GDF over grid points.
- *Comparison:* Convex ($\sum_{j<j'} |\beta_j - \beta_{j'}|$), OLS given estimated grouping.

Mean square error: Example 1



- DCP outperforms its convex counterpart and OLS based on estimated grouping.
- DCP is close to the ideal optimal performance when SNR is high.
- The average number of iterations is about 3-4.

Mean square error: Example 2



- DCP performs similarly as its convex counterpart and outperforms OLS based on estimated grouping.
- DCP is not too close to the ideal optimal performance.
- The average number of iterations is about 2.

Take Away Messages

- Grouping in regression analysis can reduce estimation variance while retaining the roughly the same amount of bias, leading to better predictive accuracy.
- Develop its graph version.
- Study other types of grouping, e.g., grouping coefficients of similar size not value, which involves the absolute values.