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 $x_i = (x_{i1}, \cdots, x_{ip})$.
- **Regression model**:

\[
Y_i \equiv \mu(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \quad (1)
\]

where $\mu(x_i) \equiv x_i^T \beta$ and $\varepsilon_i \sim N(0, \sigma^2)$.

- **Goal**: Identify all potential groupings for optimal predication of $Y$, especially when $p \gg n$.

- **Grouping pursuit** amounts to estimating grouping $G^0 = (G_1^0, \ldots, G_K^0)^T$ as well as $\alpha^0 = (\alpha_1^0, \ldots, \alpha_K^0)^T$ given $G^0$ when true $\beta^0 = (\beta_1^0, \ldots, \beta_p^0)^T = \text{or} \approx (\alpha_1^0 1_{|G_1^0|}, \ldots, \alpha_K^0 1_{|G_K^0|})^T$ with $1_{|G_1^0|}$ denoting a vector of 1’s with length $|G_1^0|$.
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, · · ·

- **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 \( p \)th order Bell number:

\[
B_p = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^p}{k!} = \text{order } e^{e^{pa}} \text{ for some } 0 < a < 1.
\]
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](image)

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

- **Enumeration**
  - Partition \( \{1, \cdots, p\} \) into \( G = (G_1, \cdots, G_k) \). Given \( G \), compute OLS through regression of \( Y \) on grouped
  \[
  Z_{G_1} \equiv X_{G_1}1, \cdots, Z_{G_k} \equiv X_{G_k}1.
  \]
  - 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

\[ S(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - x_i^T \beta)^2 + \lambda_1 J(\beta); \quad J(\beta) = \sum_{j<j'} G'(\beta_j - \beta_{j'}) , \]

(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 $S(\beta)$ at $\beta$ is the set of all subgradients at $\beta$.

- **Subgradient** of $|\beta_j - \beta_{j'}|$ wrt $\beta_j$ at $\beta = \hat{\beta}(\lambda)$ is $b_{jj'}(\lambda)$.

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

- Due to overcompleteness of the penalty, $b_{jj'}(\lambda)$ can not be estimated.

- **Subgradient of $j$ wrt group $G_k(\lambda)$**: $B_j(\lambda) \equiv \sum_{j' \in G_k(\lambda) \setminus \{j\}} b_{jj'}(\lambda)$, with $\sum_{j \in G_k(\lambda)} B_j(\lambda) = 0$, because $b_{jj'} = -b_{j'j}$ for $j \neq j'$.

- **Subgradient of subset $A$ wrt group $G_k(\lambda)$**:
  
  $$
  B_A(\lambda) \equiv \sum_{j \in A} B_j(\lambda) = \sum_{(j,j') \in A \times (G_k(\lambda) \setminus A)} b_{jj'}(\lambda), \text{ with }
  $$

  $$
  |B_A(\lambda)| \leq |A|(|G_k(\lambda)| - |A|)
  $$
  .
Solution surface via DC programming

- **Decompose** $S(\beta)$ in (2) into a difference of two convex functions

\[
S_1(\beta) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^T \beta)^2 + \lambda_1 \sum_{j < j'} |\beta_j - \beta_{j'}|
\]

and

\[
S_2(\beta) = \lambda_1 \sum_{j < j'} G_2(\beta_j - \beta_{j'})
\]

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

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

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

- 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 \( \hat{\beta}^{(m)}(\lambda) \) of (3).
  - **Optimality** through subdifferentials: \( \nabla S^{(m)}(\beta)|_{\beta=\hat{\beta}^{(m)}(\lambda)} = 0 \).

- **Major challenges:** (1) (Discontinuity) \( \hat{\beta}^{(m)}(\lambda) \) may contain jumps in \((Y, \lambda_2)\); (2) (Overcompleteness) computing \( B_j^{(m)}(\lambda) \) via enumerations over \( \{b_{jj'}\} \) is infeasible, (Bell number).

- Homotopy (1) piecewise linear and continuous in \( \lambda_1 \) given \((Y, \lambda_2)\) with piecewise linear penalty and designed support point. Transition conditions: (1) Combining groups \( G_k(\lambda) \) with \( G_k(\lambda) \): \( \alpha_k(\lambda) = \alpha_l(\lambda) \); (2) Splitting group \( G_k(\lambda) \): \( |B_A(\lambda)| \leq |A| (|G_k(\lambda)| - |A|) \).

- Overcompleteness: Use piecewise linear property of \( B_j^{(m)}(\lambda) \) for searching (order: \( O(p^2 \log p) \)).

- Homotopy Algorithm for computing \( \hat{\beta}(\lambda) \) as a function of \( \lambda \) simultaneously.
Homotopy method+DCP, continued

**Property**: terminate finitely and converge rapidly. Control at one $\lambda_0$ implies the entire surface.

Figure 3: Regularization solution path/surface.
Model selection for prediction

- Model selection:

\[
\hat{\text{GDF}}(\hat{\beta}(\lambda)) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \hat{\mu}(\lambda, x_i))^2 + \frac{1}{n} \sigma^2 \hat{df}(\lambda),
\]

(4)

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

- For piecewise smooth \(\hat{\beta}(\lambda)\) (\(m > 0\)),

\[
\hat{df}(\lambda) = \frac{\sigma^2}{\tau^2} \sum_{i=1}^{n} \text{Cov}^*(Y_i, \hat{\mu}^*(\lambda, x_i)) \text{ and Cov}^*(Y_i, \hat{\mu}^*(\lambda, x_i)),
\]

through data perturbation (GSURE, Shen & Ye, 2002).
Theory: Error analysis

- **Performance for grouping pursuit:**
  - **Error:** (Disagreement) \( P(\mathcal{G}(\lambda) \neq \mathcal{G}^0) \leq P(\hat{\beta}(\lambda) \neq \hat{\beta}^{(ols)}) \).
  - \( \mathcal{G}(\lambda), \mathcal{G}^0 \): estimated and true grouping (uniquely defined). \( \hat{\beta}(\lambda), \hat{\beta}^{(ols)} \): estimator defined by Algorithm 2 and OLS based on \( \mathcal{G}^0 \).

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

\[
\frac{K(K-1)}{2} \Phi\left(\frac{-n^{1/2}(\gamma_{\min} - \lambda_2)}{2\sigma c_{\min}^{-1/2}}\right) + p \Phi\left(\frac{-n\lambda_1}{\sigma \max_{1 \leq j \leq p} \|x_j\|}\right). \tag{5}
\]

- \( \Phi(z) \): CDF of \( N(0, 1) \).
- \( \|x_j\| \): \( L_2 \)-norm of \( x_j \).
- \( \gamma_{\min} \): \( \min \{ |\alpha_k^0 - \alpha_l^0| > 0 : 1 \leq k < l \leq K \} \).
- \( c_{\min} \): smallest eigenvalue of \( Z_{\mathcal{G}^0}^T Z_{\mathcal{G}^0} / n \).
- \( K \): \# of estimated groups, which is no larger than \( \min(n, p) \).
Theory: Error analysis, continued

If \( \max \left\{ \frac{nc_{\min}(\gamma_{\min} - \lambda_2)^2}{8\sigma^2}, \frac{n\lambda_1^2}{2\sigma^2 \max_{1 \leq j \leq p} \|x_j\|^2/n} \right\} - \log p \to \infty, \)

- **Grouping Consistency**

\[
P(G(\lambda) \neq G^0) \leq P(\hat{\beta}(\lambda) \neq \hat{\beta}^{(ols)}) \to 0, \quad p, n \to +\infty.
\]

- **Remarks:**
  - Roughly: \( p < \exp(O(n\lambda_1^2)), \lambda_1 \to 0, n^{1/2}\lambda_1 \to \infty, \)
    \( nc_{\min}(\gamma_{\min} - \lambda_2) \to \infty. \) \( (\max_{j:1 \leq j \leq p} \|x_j\|^2/n \) 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 \( \varepsilon_1 \).
Theory: Grouping

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

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

- Predictors with similar values of correlations are grouped together, as characterized by intervals $\bigcup_{k=1}^{K(\lambda)} \left( n\lambda_1 \delta_k(\lambda) - n\lambda_1 (|G_k(\lambda)| - 1), n\lambda_1 \delta_k(\lambda) + n\lambda_1 (|G_k(\lambda)| - 1) \right)$. 
Numerical examples

- **Ex1**: (Sparse grouping). In (2), $\varepsilon_i \sim N(0, \sigma^2)$ and $\sigma^2$ according to SNR; $\mathbf{x}_i \sim N(0, \Sigma_{p \times p})$ with $n = 50$, $p = 20$ and diagonal/off-diagonal elements $1/0.5$;

$$\beta = (0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2)^T.$$  

- **Ex2**: (Large $p$ but small $n$). In (2), $\varepsilon_i \sim N(0, \sigma^2)$, with $SNR = 10$; $\mathbf{x}_i \sim N(0, \Sigma)$ with $0.5|j - k|$ the $jk$-th element of $\Sigma$. Here

$$\beta = (3, \ldots, 3, -1.5, \ldots, -1.5, 1, \ldots, 1, 2, \ldots, 2, 0, \ldots, 0)^T.$$  

- **Mean squares error**: averaged over 100 replications.

- **Tuning**: $\lambda$ 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.