Subgroup Analysis and Variable Selection

Xiaoli Li

USTC

April 16, 2023

Outline

1. Variable Selection
   - Lasso
   - Elastic net
   - Asymptotically unbiased concave penalties
2. Subpopulation structure identification
   - Fused Lasso
   - OSCAR
   - Grouping pursuit via truncated L_1-penalty for fusions
   - Concave pairwise fusion
   - MDSP
3. β-model of network
   - Monotonicity Lemma
   - Subgroup Identification for β-model

Table of Contents

1. Variable Selection
   - Lasso
   - Elastic net
   - Asymptotically unbiased concave penalties
2. Subpopulation structure identification
   - Fused Lasso
   - OSCAR
   - Grouping pursuit via truncated L_1-penalty for fusions
   - Concave pairwise fusion
   - MDSP
3. β-model of network
   - Monotonicity Lemma
   - Subgroup Identification for β-model

Variable Selection

Consider the usual regression problem: we have data $(\boldsymbol{x}_i, y_i), i = 1, 2, ..., N$. The OLS estimates often exhibit both poor prediction accuracy, i.e. low bias but large variance, and poor interpretation. Subset selection provides interpretable models but can be extremely variable due to its discrete quality. The ridge regression provides better prediction accuracy via a bias-variance trade-off but still cannot give an easily interpretable model. Thus, the aim of variable selection is to achieve both prediction accuracy and interpretation.

Many panelized variable selection methods have been developed including non-negative garotte, Lasso, SCAD, elastic net, adaptive lasso, group lasso, MCP and TLP.

Lasso

Model: $y_i = \alpha + \sum_j \beta_j x_{ij}$

Definition: the lasso estimate $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ is defined by

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \sum_j |\beta_j| \leq t. \tag{1}$$

Due to the non-differential nature of l_1-penalty, it can shrink some coefficients exactly to 0. So the lasso does both continuous shrinkage and automatic variable selection simultaneously.

The LARS algorithm (Efron, Hastie, Johnstone, and Tibshirani 2004) can solve the entire solution path of the lasso effectively.

Although the lasso has shown success in many situations, it has some limitations. Consider the following three scenarios.

- In the $p > n$ case, the lasso selects at most n variables before it saturates, because of the convex optimization problem.
- If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected
- For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression.
- The lasso shrinkage produces biased estimates for the large coefficients, and thus it could be suboptimal in terms of estimation risk. Meinshausen and Bühlmann (2004) showed the conflict of optimal prediction and consistent variable selection in the lasso. And H.Zou (2006) proved that the underlying model must satisfy a nontrivial condition if the lasso variable selection is consistent.

Assume two conditions:

- $y_i = \boldsymbol{x}_i\boldsymbol{\beta}^* + \epsilon_i$, where $\epsilon_1, ..., \epsilon_n$ satisfy Gauss-Markov assumption.
- $\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} \to \boldsymbol{C}$, where $\boldsymbol{C}$ is a positive definite matrix.

Theorem

Suppose that $\lim_{n\to\infty} P(\mathcal{A}_n = \mathcal{A}) = 1$. *Then there exists some sign vector* $\boldsymbol{s} = (s_1, ..., s_{p0})^\top$, $s_j = 1$ *or* -1, *such that*

$$|\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{s}| \leq 1 \tag{2}$$

The foregoing inequality is understood componentwise.

Immediately, we conclude that if condition (2) fails, then the lasso variable selection is inconsistent. And H.Zou (2006) constructed an interesting example that condition (2) was not satisfied.

Adaptive Lasso

We define the adaptive lasso. Suppose that $\hat{\boldsymbol{\beta}}$ is a root-n-consistent estimator to $\boldsymbol{\beta}^*$; for example, we choose $\hat{\boldsymbol{\beta}}(\text{OLS})$. Pick a $\gamma > 0$, and define the weight vector $\hat{\boldsymbol{w}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$. The adaptive lasso estimaties $\hat{\boldsymbol{\beta}}^{*(n)}$ are given by

$$\hat{\boldsymbol{\beta}}^{*(n)} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j\|^2 + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|. \tag{3}$$

The adaptive lasso enjoys the oracle properties.

Theorem

Suppose that $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$. Then the adaptive lasso estimates must satisfy the following:

$$\lim_{n\to\infty} P(\mathcal{A}_n^* = \mathcal{A}) = 1$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{*(n)} - \boldsymbol{\beta}_{\mathcal{A}}^*) \to \mathrm{N}(\boldsymbol{0}, \sigma^2 \times \boldsymbol{C}_{11}^{-1}).$$

Summary

Advantages:

- The (adaptive) lasso achieves both prediction accuracy and interpretation. Some coefficients of estimators can be shrunk exactly to 0.
- The estimates given by adaptive lasso enjoys favorable oracle properties including selection consistency and asymptotic normality.
- The (adaptive) lasso is a continuous convex optimization problem and can be effectively solved by LARS algorithm.

Disadvantages:

- The (adaptive) lasso is based on a homogeneous model assumption and thus cannot identify underlying group structure.
- The lasso cannot handle high-dimensional problem, i.e. $p > n$ since it selects at most n variables.

Naive Elastic net

To address the problems that were highlighted above, Zou and Hastie (2005) proposed a penalized least squares method using a novel elastic net penalty defind as following,

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1.$$

The naive elastic net estimator $\hat{\boldsymbol{\beta}}$ is the minimizer of the above equation:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \{L(\lambda_1, \lambda_2, \boldsymbol{\beta})\}. \tag{4}$$

Lemma

Given data set $(\boldsymbol{y}, \boldsymbol{X})$ *and* (λ_1, λ_2), *define an artificial data set* $(\boldsymbol{y}^*, \boldsymbol{X}^*)$ *by*

$$X^*_{(n+p)\times p} = (1+\lambda_2)^{-\frac{1}{2}} \begin{pmatrix} \boldsymbol{X} \\ \sqrt{(\lambda_2)}\boldsymbol{I} \end{pmatrix}, \qquad \boldsymbol{y}^*_{(n+p)} = \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix}$$

Let $\gamma = \lambda_1/\sqrt{1+\lambda_2}$ *and* $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. *Then the naive elastic net criterion can be written as*

$$L(\gamma, \boldsymbol{\beta}) = L(\gamma, \boldsymbol{\beta}^*) = |\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*|^2 + \gamma|\boldsymbol{\beta}^*|_1.$$

Then

$$\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\lambda_2}} \arg\min_{\boldsymbol{\beta}^*} L(\gamma, \boldsymbol{\beta}^*).$$

This lemma says that we can transform the naive elastic net problem into an equivalent lasso problem on augmented data. Also note that the naive elastic net can potentially select all p predictors in all situations and has the ability to selecting grouped variables.

Groupping effect

Lemma

Given data set $(\boldsymbol{y}, \boldsymbol{X})$ *and* (λ_1, λ_2)*, the response* $\boldsymbol{y}$ *is centred and the predictors* $\boldsymbol{X}$ *are standardized. Let* $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ *be the naive elastic net estimate. Suppose that* $\hat{\beta}_i(\lambda_1, \lambda_2)\hat{\beta}_j(\lambda_1, \lambda_2) > 0$*. Define*

$$D_{\lambda_1,\lambda_2}(i,j) = \frac{1}{|\boldsymbol{y}|_1}|\hat{\beta}_i(\lambda_1, \lambda_2) - \hat{\beta}_j(\lambda_1, \lambda_2)|;$$

then

$$D_{\lambda_1,\lambda_2}(i,j) \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)},$$

where $\rho = \boldsymbol{x}_i^\top \boldsymbol{x}_j$*, the sample correlation.*

Elastic net

For the naive elastic net, double shrinkage does not help to reduce the variances much and introduces unnecessary extra bias, compared with pure lasso or ridge shrinkage. The author improve the prediction performance of the naive elastic net by correcting this double shrinkage as follows,

$$\hat{\boldsymbol{\beta}}_{EN} = (1+\lambda_2)\hat{\boldsymbol{\beta}}. \tag{5}$$

Theorem

Given data set $(\boldsymbol{y}, \boldsymbol{X})$ and (λ_1, λ_2), then the elastic net estimates $\hat{\boldsymbol{\beta}}$ are given by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \left(\frac{\boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \boldsymbol{I}}{1+\lambda_2} \right) \boldsymbol{\beta} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \lambda_1 |\boldsymbol{\beta}|_1. \tag{6}$$

Hence this theorem interprets the elastic net as a stabilized version of the lasso.

Summary

Advantages:

- The elastic net overcomes the problems that lasso has. For example, the elastic net can potentially select p variables and enjoys groupping effect.
- The LARS-EN algorithm computes the estimates given by the elastic net efficiently.

Disadvantages:

- The groupping effect is actually based on the correlation of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, which has little effect on finding underlying group structure.
- Although the author provides justification for choosing $1 + \lambda_2$ as the scaling factor, the reason why this factor matters is still not clear. Meanwhile, the author does not give the asymptotic properties and convergence rate about the estimator.

The concave penalties producing unbiased estimates

In practice, the lasso tends to over-shrink large coefficients and thus results in biasness. Therefore, to reduce the bias and obtain sparse solution, many scholars proposed concave penalties including SCAD (Fan and Li; 2001) and MCP (Zhang; 2010). These penalties are asymptotically unbiased and are more aggressive in enforcing a sparser solution.
The MCP has the form

$$p_\gamma(t,\lambda) = \lambda \int_0^t (1 - \frac{x}{\gamma\lambda})_+ dx, \gamma > 1,$$

and the SCAD penalty is

$$p_\gamma(t,\lambda) = \lambda \int_0^t \min\{1, (\gamma - x/\lambda)_+/(\gamma - 1)\} dx, \gamma > 2,$$

where γ is a parameter that controls the concavity of the penalty function. These concave penalties enjoy sparsity as the L_1 penalty and more importantly, the concave penalties have the unbiasedness property in that they do not shrink large estimated parameters, so that they remain unbiased in the iterations.

Statistical properties

Theorem

Let $\boldsymbol{V}_1, \ldots \boldsymbol{V}_n$ be independent and identically distributed, each with a density $f(\boldsymbol{V}, \beta)$ satisfying regularity conditions for the generalized linear model. Assume the penalty function satisfies

$$\liminf_{n\to\infty} \liminf_{\theta\to 0^+} p'_{\lambda_n}(\theta)/\lambda_n > 0.$$

If $\lambda_n > 0$ and $\sqrt{n}\lambda_n \to \infty$, then with probability tending to 1, the root-n consistent local minimizer $\hat{\boldsymbol{\beta}} = \left(\begin{array}{c} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{array}\right)$ must satisfy:

$$\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$$

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{10}) + \Sigma)\{\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (I_1(\boldsymbol{\beta}_{10}) + \Sigma)^{-1}\boldsymbol{b}\} \to N(\mathbf{0}, I_1(\boldsymbol{\beta}_{10}))$$

Here,

$$\Sigma = \mathrm{diag}\{p''_{\lambda_n}(|\beta_{10}|), ..., p''_{\lambda_n}(|\beta_{s0}|)\},$$

and

$$\boldsymbol{b} = (p'_{\lambda_n}(|\beta_{10}|)\mathrm{sgn}(\beta_{10}), ..., p'_{\lambda_n}(|\beta_{s0}|)\mathrm{sgn}(\beta_{s0})),$$

where s is the number of components of $\boldsymbol{\beta}_{10}$.
And $I_1(\boldsymbol{\beta}_{10})$ is the Fisher information knowing $\boldsymbol{\beta}_2 = \mathbf{0}$.

Table of Contents

1. Variable Selection
   - Lasso
   - Elastic net
   - Asymptotically unbiased concave penalties

2. Subpopulation structure identification
   - Fused Lasso
   - OSCAR
   - Grouping pursuit via truncated L_1-penalty for fusions
   - Concave pairwise fusion
   - MDSP

3. β-model of network
   - Monotonicity Lemma
   - Subgroup Identification for β-model

Subgroup Structure

Essential to high-dimensional data analysis is seeking a certain lower-dimensional structure in knowledge discovery, as in web mining. The central issue is automatic identification of homogenous subgroups in regression, which is called grouping pursuit.
First we consider the following linear model:

$$Y_i = \sum_{j=1}^{p} x_j \beta_j + \epsilon_j.$$

Our objective is to identify all possible homogenous subgroups of predictors, for optimal prediction of the outcome of $\boldsymbol{Y}$. Here homogeneity means that regression coefficients are of similar (same) values, that is, $\beta_{j_1} \approx \cdots \approx \beta_{j_K}$ within each group. Grouping pursuit estimates all distinct values of $\boldsymbol{\beta}$ as well as all corresponding subgroups of homogenous predictors.

Fused Lasso

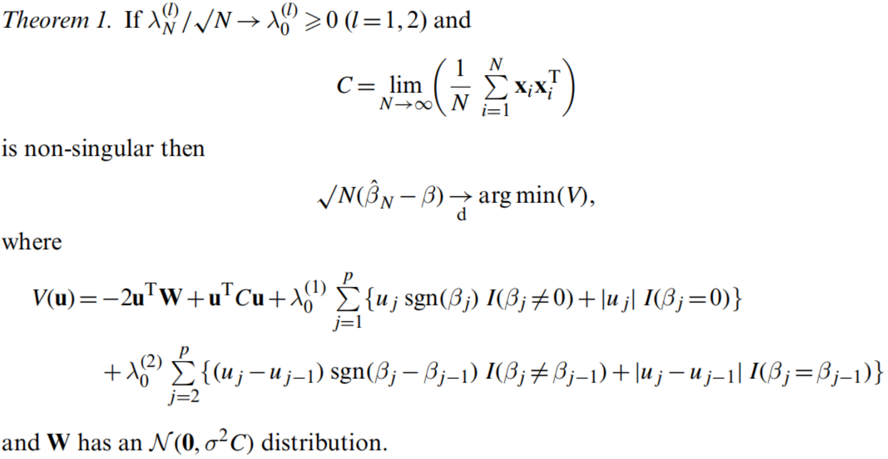
One drawback of the lasso in the present context is the fact that it ignores ordering of the features. For this purpose, Tibshirani (2005) proposed a fused Lasso by adding an L_1-penalty to the pair of adjacent coefficients. The fused lasso is defined by

$$\hat{\beta} = \arg\min \left\{ \sum_i (y_i - \sum_j x_i j \beta_j)^2 \right\} \quad \text{subject to } \|\beta\|_1 \leq s_1 \text{ and } \sum_{j=2}^{p} |\beta_j - \beta_{j-1}| \leq s_2. \tag{7}$$

The first constraint encourages sparsity in the coefficients; the second encourages sparsity in their differences.

Asymptotic properties

Here, the dimension p is fixed with $N \to \infty$.



Theorem 1. If $\lambda_N^{(l)}/\sqrt{N} \to \lambda_0^{(l)} \geqslant 0$ $(l = 1, 2)$ and

$$C = \lim_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)$$

is non-singular then

$$\sqrt{N}(\hat{\beta}_N - \beta) \underset{\mathrm{d}}{\to} \arg\min(V),$$

where

$$V(\mathbf{u}) = -2\mathbf{u}^{\mathrm{T}}\mathbf{W} + \mathbf{u}^{\mathrm{T}} C\mathbf{u} + \lambda_0^{(1)} \sum_{j=1}^{p} \{u_j \operatorname{sgn}(\beta_j)\, I(\beta_j \neq 0) + |u_j|\, I(\beta_j = 0)\}$$
$$+ \lambda_0^{(2)} \sum_{j=2}^{p} \{(u_j - u_{j-1}) \operatorname{sgn}(\beta_j - \beta_{j-1})\, I(\beta_j \neq \beta_{j-1}) + |u_j - u_{j-1}|\, I(\beta_j = \beta_{j-1})\}$$

and $\mathbf{W}$ has an $\mathcal{N}(\mathbf{0}, \sigma^2 C)$ distribution.

Sparsity of fused lasso solutions

Theorem

Set $\beta_0 = 0$*. Let* $n_{seq}(\beta) = \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq \beta_{j-1}\}$*. Then, under'non-redundancy' conditions on the design matrix* $\boldsymbol{X}$*, the fused lasso problem (7) has a unique solution* $\hat{\beta}$ *with* $n_{seq}(\hat{\beta}) \leq N$*.*

The non-redundancy conditions mentioned can be qualitatively summarized as follows.

- No N columns of the design matrix $\boldsymbol{X}$ are linearly dependent.
- None of a finite set of $N + 1$ linear equations in N variables (the coefficients of which depend on the specific problem) has a solution.

Summary

Advantages:

- The fused lasso seems a promising method for regression and classification, in settings where the features have a natural order.

Disadvantages:

- One difficulty in using the fused lasso is computational speed. When the numbers of parameter p and samples N are large, speed could become a practical limitation.
- The fused lasso assume parameter homogeneity over individuals and target on grouping similar-effect covariates.

OSCAR

The constrained least-squares optimization problem for the OSCAR is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j\|^2 \\ &\text{subject to } \sum_{j=1}^{p} |\beta_j| + c \sum_{j<k} \max\{|\beta_j|, |\beta_k|\} \leq t, \end{aligned} \tag{8}$$

where $c \geq 0$ and $t > 0$ are tuning constants with c controlling the relative weighting of the norms and t controlling the magnitude. The L_1 norm encourages sparseness, while the pairwise L_∞ norm encourages equality of coefficients.

As the contours are more likely to hit at a vertex, the nondifferentiability of the LASSO and Elastic Net at the axes encourage sparsity, with the LASSO doing so to a larger degree due to the linear boundary. Meanwhile, if two variables were highly correlated, the Elastic Net would more often include both into the model, as opposed to including only one of the two.

The shape of the constraint region in two dimensions is exactly an octagon. With vertices on the diagonals along with the axes, the OSCAR encourages both sparsity and equality of coefficients to varying degrees, depending on the strength of correlation, the value of c, and the location of the OLS solution.

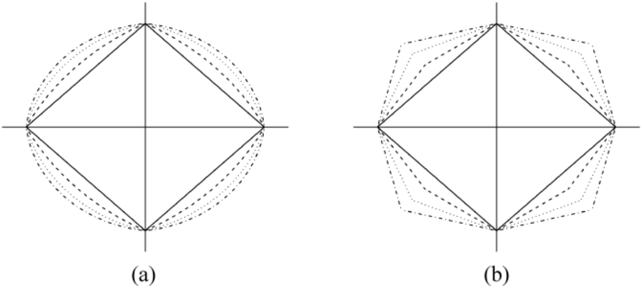


Figure 1. Graphical representation of the constraint region in the (β_1, β_2) plane for the LASSO, Elastic Net, and OSCAR. Note that all are nondifferentiable at the axes. (a) Constraint region for the Lasso (solid line), along with three choices of tuning parameter for the Elastic Net. (b) Constraint region for the OSCAR for four values of c. The solid line represents $c = 0$, the LASSO.

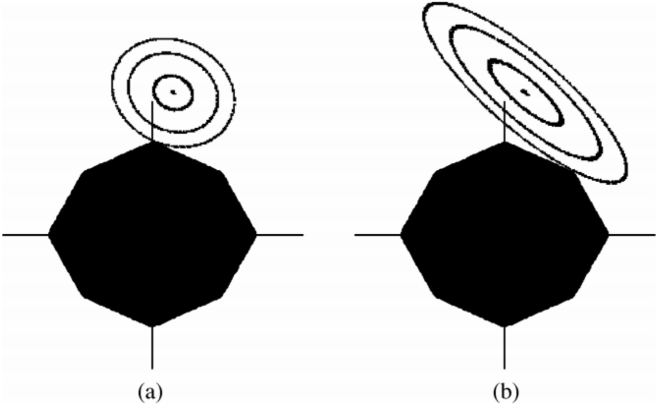
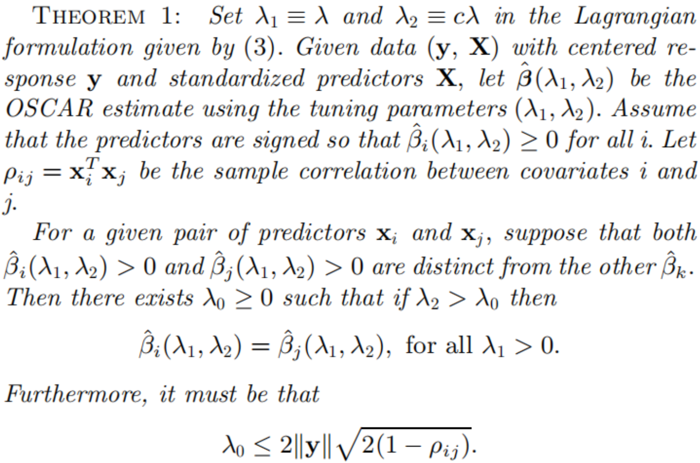


Figure 2. Graphical representation in the (β_1, β_2) plane. The OSCAR solution is the first time the contours of the sum-of-squares function hits the octagonal constraint region. (a) Contours centered at OLS estimate, low correlation ($\rho = 0.15$). Solution occurs at $\hat{\beta}_1 = 0$. (b) Contours centered at OLS estimate, high correlation ($\rho = 0.85$). Solution occurs at $\hat{\beta}_1 = \hat{\beta}_2$.

Exact Grouping Property



THEOREM 1: *Set $\lambda_1 \equiv \lambda$ and $\lambda_2 \equiv c\lambda$ in the Lagrangian formulation given by* (3). *Given data* ($\mathbf{y}$, $\mathbf{X}$) *with centered response* $\mathbf{y}$ *and standardized predictors* $\mathbf{X}$, *let* $\hat{\boldsymbol{\beta}}(\lambda_1, \lambda_2)$ *be the OSCAR estimate using the tuning parameters* (λ_1, λ_2). *Assume that the predictors are signed so that* $\hat{\beta}_i(\lambda_1, \lambda_2) \geq 0$ *for all i. Let* $\rho_{ij} = \mathbf{x}_i^T \mathbf{x}_j$ *be the sample correlation between covariates i and j.*

For a given pair of predictors $\mathbf{x}_i$ *and* $\mathbf{x}_j$, *suppose that both* $\hat{\beta}_i(\lambda_1, \lambda_2) > 0$ *and* $\hat{\beta}_j(\lambda_1, \lambda_2) > 0$ *are distinct from the other* $\hat{\beta}_k$. *Then there exists* $\lambda_0 \geq 0$ *such that if* $\lambda_2 > \lambda_0$ *then*

$$\hat{\beta}_i(\lambda_1, \lambda_2) = \hat{\beta}_j(\lambda_1, \lambda_2), \text{ for all } \lambda_1 > 0.$$

Furthermore, it must be that

$$\lambda_0 \leq 2\|\mathbf{y}\|\sqrt{2(1-\rho_{ij})}.$$

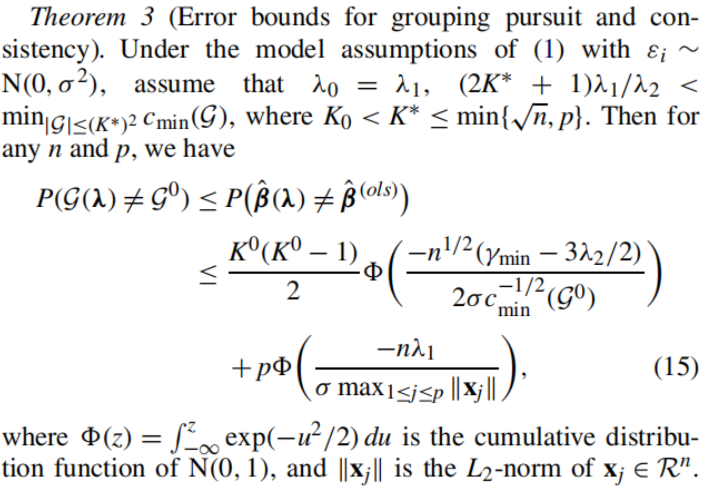
Grouping pursuit via truncated L_1-penalty for fusions

A natural extension of the fused lasso is to apply the L_1-penalty pairwisely, i.e. with penalty $\sum_{j<k} |\beta_j - \beta_k|$. However, this convex penalty is not desirable for predictive performance, because it is not adaptive for discriminating large from small pairwise differences. As a result, overpenalizing large differences due to shrinking small differences towards zero impedes predictive performance. Thus, Shen and Huang (2013) developed a grouping pursuit algorithm utilizing the truncated L_1-penalty for fusions.

The penalized least squares criterion for automatic grouping pursuit is defined by:

$$S(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda_1 \sum_{j<k} \min\{|\beta_j - \beta_k|, \lambda_2\} \tag{9}$$

Asymptotic theory

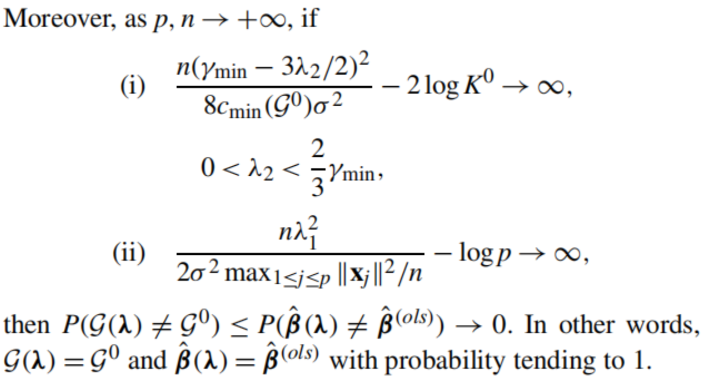


Theorem 3 (Error bounds for grouping pursuit and consistency). Under the model assumptions of (1) with $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, assume that $\lambda_0 = \lambda_1$, $(2K^* + 1)\lambda_1/\lambda_2 < \min_{|\mathcal{G}| \leq (K^*)^2} c_{\min}(\mathcal{G})$, where $K_0 < K^* \leq \min\{\sqrt{n}, p\}$. Then for any n and p, we have

$$P(\mathcal{G}(\boldsymbol{\lambda}) \neq \mathcal{G}^0) \leq P\big(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \neq \hat{\boldsymbol{\beta}}^{(ols)}\big) \\ \leq \frac{K^0(K^0 - 1)}{2} \Phi\left(\frac{-n^{1/2}(\gamma_{\min} - 3\lambda_2/2)}{2\sigma c_{\min}^{-1/2}(\mathcal{G}^0)}\right) \\ + p\Phi\left(\frac{-n\lambda_1}{\sigma \max_{1 \leq j \leq p} \|\mathbf{x}_j\|}\right), \qquad (15)$$

where $\Phi(z) = \int_{-\infty}^{z} \exp(-u^2/2)\, du$ is the cumulative distribution function of $\mathrm{N}(0, 1)$, and $\|\mathbf{x}_j\|$ is the L_2-norm of $\mathbf{x}_j \in \mathcal{R}^n$.

Asymptotic theory



Moreover, as $p, n \to +\infty$, if

$$\text{(i)} \quad \frac{n(\gamma_{\min} - 3\lambda_2/2)^2}{8c_{\min}(\mathcal{G}^0)\sigma^2} - 2\log K^0 \to \infty,$$

$$0 < \lambda_2 < \frac{2}{3}\gamma_{\min},$$

$$\text{(ii)} \quad \frac{n\lambda_1^2}{2\sigma^2 \max_{1 \le j \le p} \|\mathbf{x}_j\|^2/n} - \log p \to \infty,$$

then $P(\mathcal{G}(\boldsymbol{\lambda}) \neq \mathcal{G}^0) \le P(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \neq \hat{\boldsymbol{\beta}}^{(ols)}) \to 0$. In other words, $\mathcal{G}(\boldsymbol{\lambda}) = \mathcal{G}^0$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \hat{\boldsymbol{\beta}}^{(ols)}$ with probability tending to 1.

Concave pairwise fusion

Similar to the truncated L_1-penalty for fusions, Ma and Huang (2017) formulated clustering as a penalized regression problem by adopting a fusion-type penalty with either an L_p-shrinkage or a nonconvex penalty function. The model they investigated was

$$y_i = \mu_i + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i,$$

where $\mu_i' s$ are unknown subject-specific intercepts, $\boldsymbol{\beta}$ is the vector of unknown coefficients for the covariates $\boldsymbol{x}_i$.

The objective function is

$$Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda) = \frac{1}{2} \sum_{i=1}^{n} (y_i - \mu_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \sum_{1 \le i < j \le n} p_\gamma(|\mu_i - \mu_j|, \lambda), \quad (10)$$

where $p_\gamma(\dot{;}\lambda)$ is a concave penalty function with a tuning parameter $\lambda > 0$.

An important question is which penalty function should be used. Just like what we discussed above, the L_1 penalty applies the same thresholding to all pairs $|\mu_i - \mu_j|$. As a result, it leads to biased estimates and may not be able to correctly recover subgroups. This is similar to the situation in variable selection, where the Lasso tends to over-shrink large coefficients. Hence, the author used the concave penalties including SCAD (Fan and Li; 2001) and MCP (Zhang; 2010).

The concave penalties have the unbiasedness property in that they do not shrink large estimated parameters, so that they remain unbiased in the iterations. This property is particularly essential in the ADMM algorithms since the biases in the iterations may significantly affect the search for subgroups.

Theoretical properties

- Heterogeneous model

For the the heterogeneous model in which there are at least two subgroups, denote

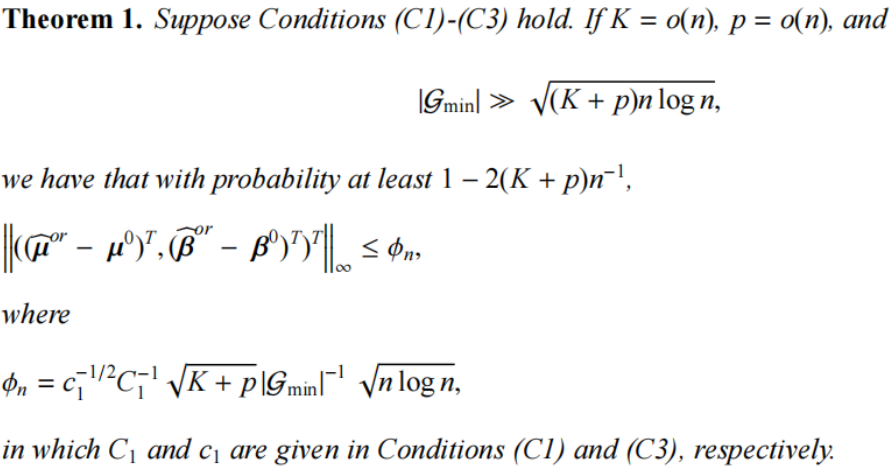
$$\mathcal{M}_{\mathcal{G}} = \{\boldsymbol{\mu} \in R^n : \mu_i = \mu_j, \text{ for any } i, j \in \mathcal{G}_k, 1 \leq k \leq K\}.$$

For each $\boldsymbol{\mu} \in \mathcal{M}_{\mathcal{G}}$, it can be written as $\boldsymbol{\mu} = \boldsymbol{Z}\boldsymbol{\alpha}$, where $\boldsymbol{Z} = \{z_{ik}\}$ is the $n \times K$ matrix with $z_{ik} = 1$ for $i \in \mathcal{G}_k$ and $z_{ik} = 0$ otherwise, and $\boldsymbol{\alpha}$ is a $K \times 1$ vector of parameters. We have $\boldsymbol{D} = \boldsymbol{Z}^\top \boldsymbol{Z} = \text{diag}(|\mathcal{G}_1|, ..., |\mathcal{G}_K|)$.

$$\rho(t) = \lambda^{-1} p_\gamma(t, \lambda) \text{ and } \bar{\rho}(t) = \rho'(|t|)\text{sgn}(t)$$

The author introduced the following conditions.

- (C1) Assume $\|\boldsymbol{X}_j\| = \sqrt{n}, \lambda_{min}[(\boldsymbol{Z}, \boldsymbol{X})^\top(\boldsymbol{Z}, \boldsymbol{X})] \geq C_1|\mathcal{G}_{\min}|$, and $\|\boldsymbol{X}\| \leq C_2 p$ for some constants $0 < C_1 \leq 1$ and $0 < C_2 < \infty$.
- (C2) $p_\gamma(t, \lambda)$ is a symmetric function of t, and it is non-decreasing and concave in t for $t \in [0, \infty)$. $\rho(t)$ is a constant for all $t \geq a\lambda$, and $\rho(0) = 0$. $\rho'(t)$ exists and is continuous except for a finite number of t and $\rho'(0+) = 1$.
- (C3) The noise vector $\boldsymbol{\epsilon}$ has sub-Gaussian tails such that $P(|\boldsymbol{a}^\top\boldsymbol{\epsilon}| > \|\boldsymbol{a}\|x) \leq 2\exp(-c_1 x^2)$ for any vector $\boldsymbol{a} \in R^n$ and $x > 0$, where $0 < c < \infty$.



Theorem 1. *Suppose Conditions (C1)-(C3) hold. If* $K = o(n)$, $p = o(n)$, *and*

$$|\mathcal{G}_{\min}| \gg \sqrt{(K+p)n\log n},$$

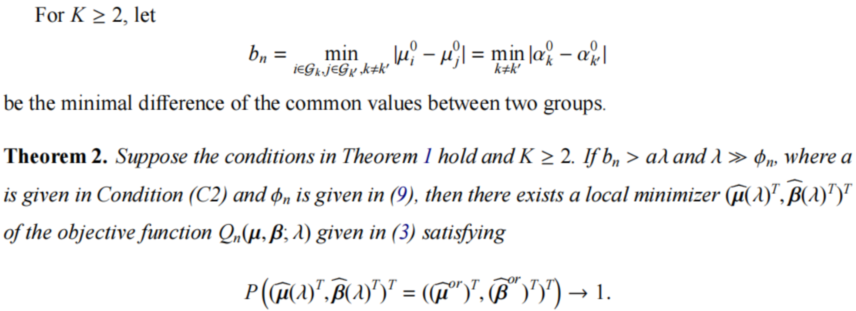
we have that with probability at least $1 - 2(K+p)n^{-1}$,

$$\left\|\left((\widehat{\boldsymbol{\mu}}^{or} - \boldsymbol{\mu}^0)^T, (\widehat{\boldsymbol{\beta}}^{or} - \boldsymbol{\beta}^0)^T\right)^T\right\|_\infty \le \phi_n,$$

where

$$\phi_n = c_1^{-1/2} C_1^{-1} \sqrt{K+p}\,|\mathcal{G}_{\min}|^{-1} \sqrt{n\log n},$$

in which C_1 *and* c_1 *are given in Conditions (C1) and (C3), respectively.*



For $K \geq 2$, let

$$b_n = \min_{i \in \mathcal{G}_k, j \in \mathcal{G}_{k'}, k \neq k'} |\mu_i^0 - \mu_j^0| = \min_{k \neq k'} |\alpha_k^0 - \alpha_{k'}^0|$$

be the minimal difference of the common values between two groups.

Theorem 2. *Suppose the conditions in Theorem 1 hold and $K \geq 2$. If $b_n > a\lambda$ and $\lambda \gg \phi_n$, where a is given in Condition (C2) and ϕ_n is given in (9), then there exists a local minimizer $(\widehat{\boldsymbol{\mu}}(\lambda)^T, \widehat{\boldsymbol{\beta}}(\lambda)^T)^T$ of the objective function $Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda)$ given in (3) satisfying*

$$P\left((\widehat{\boldsymbol{\mu}}(\lambda)^T, \widehat{\boldsymbol{\beta}}(\lambda)^T)^T = ((\widehat{\boldsymbol{\mu}}^{or})^T, (\widehat{\boldsymbol{\beta}}^{or})^T)^T\right) \to 1.$$

MDSP

The fusion-type of penalties emphasizes on subgrouping and feature selection is not incorporated. In addition, the pairwise fusion also leads to estimation bias due to pulling individuals together from different subgroups. Therefore, Annie Qu (2019) proposed an effective individualized model selection approach using multidirectional shrinkage to select unique relevant features for different individuals and identify subgroups based on heterogeneous covariates' effects simultaneously.

Consider the heterogeneous regression model:

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{Z}_i \boldsymbol{\alpha} + \boldsymbol{\epsilon}_i, \quad i = 1, ..., N,$$

where each individual is associated with a unique effect $\boldsymbol{\beta}_i = (\beta_{i1}, ..., \beta_{ip})^{\mathrm{T}}$ for some targeting variables $\boldsymbol{X}_i$, in addition to a homogeneous effect $\boldsymbol{\alpha} = (\alpha_i, ..., \alpha_q)^{\mathrm{T}}$ for some control variables $\boldsymbol{Z}_i$. The random error $\boldsymbol{\epsilon}_i = (\epsilon_{i,1}, ..., \epsilon_{i,m})^{\mathrm{T}}$ are independent over different individuals, while within an individual, $\epsilon'_{i,t}s$ have mean 0 and variance σ^2, and could be correlated.

Heterogeneous Regression Model

We could select and estimate the regression parameters $\boldsymbol{\beta}_i$'s and $\boldsymbol{\alpha}$ by minimizing the penalized objective function

$$\frac{1}{2}\sum_{i=1}^{N}\|\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_i-\boldsymbol{Z}_i\boldsymbol{\alpha}\|_2^2+\sum_{i=1}^{N}\sum_{k=1}^{p}h_{\lambda_{N,m}}(\beta_{ik}), \tag{11}$$

where $h_{\lambda_{N,m}}(\cdot)$ refers to a feature selection penalty function. Let$\boldsymbol{\beta}_{(N)}$ =vec$(\boldsymbol{\beta}_i)_{i=1}^{N}$, $\boldsymbol{Y}$ =vec$(\boldsymbol{y}_i)_{i=1}^{N}$, $\boldsymbol{X}$ =bdiag$(\boldsymbol{X}_i)_{i=1}^{N}$ and $\boldsymbol{Z}=[\boldsymbol{Z}_1^{\mathrm{T}}...\boldsymbol{Z}_N^{\mathrm{T}}]$. Without the penalty term in (1), the ordinary least squares (OLS) estimator is obtained as

$$\mathrm{vec}(\hat{\beta}_{(N)}^{OLS},\hat{\alpha}^{OLS})=([XZ]^{\mathrm{T}}[XZ])^{-1}[XZ]^{\mathrm{T}}Y$$

where the dimension of parameters $(Np+q)$ will diverge as sample size N increases. The model in (1) only utilize individual-specific information, which will lead to inefficient estimation and over-fitting of a model.

Multidirectional Separation Penalty

To achieve more efficient individualized modeling, it is crucial and beneficial to encourage grouping some individuals which share similar treatment (covariates) effects.

For the individualized coefficients $\boldsymbol{\beta}_{\cdot k} = (\beta_{1k}, ..., \beta_{Nk})^{\mathrm{T}}$ of the kth heterogeneous-effect predictor, we assume that there are B_k subgroups as

$$\beta_{ik} = \begin{cases} \gamma_k^{(l)}, \ \ \textit{if} \quad i \in \mathcal{G}_k^{(l)}, \quad l = 1, ..., B_k - 1 \\ 0, \ \ i \in \mathcal{G}_k^{(0)} \end{cases} \qquad \text{for } i = 1, ..., N \qquad (12)$$

where each $\gamma_k^{(l)}$ is an unknown nonzero sub-homogeneous effect shared by individuals within the lth subgroup, and the index partition sets $\{\mathcal{G}_k^{(l)}\}$ represent the corresponding subgroup memberships in terms of the heterogeneous effects of the kth predictor. This is different from conventional subgroup analysis approaches which assume a uniform subgroup structure on individuals over all covariates' effects.

Multidirectional Separation Penalty

Penalized objective function with the sub-homogeneous effect γ induced in a multidirectional separation penalty(MDSP) $s_\lambda(\cdot,\cdot)$ as

$$\begin{aligned} Q_{N,m}(\boldsymbol{\alpha}, \boldsymbol{\beta}_{(N)}, \gamma) = \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_i, \boldsymbol{\alpha}))^{\mathrm{T}} \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_i, \boldsymbol{\alpha})) \\ + \sum_{i=1}^{N} \sum_{k=1}^{p} h_{\lambda_{N,m}}(\boldsymbol{\beta}_{(N)}, \gamma) \end{aligned} \tag{13}$$

where$\boldsymbol{\mu}_i(\boldsymbol{\beta}_i, \boldsymbol{\alpha})) = \boldsymbol{X}_i \boldsymbol{\beta}_i + \boldsymbol{Z}_i \boldsymbol{\alpha}$. To obtain more efficient estimation, the within-individual serial correlations are utilized by a weighting matrix $\boldsymbol{V}_i = \boldsymbol{A}_i^{\frac{1}{2}} \boldsymbol{R}_i \boldsymbol{A}_i^{\frac{1}{2}}$, where $\boldsymbol{A}_i$ is a diagonal matrix of marginal variance of $\boldsymbol{y}_i$ and $\boldsymbol{R}_i$ is a working correlation matrix.

Multidirectional Separation Penalty

The key component of the proposed model is a designed multidirectional separation penalty (MDSP) function $s_{\lambda}(\beta_{ik}, \gamma_{ik})$, defined as

$$s_{\lambda}(\beta_{ik}, \gamma_{ik}) = \lambda_{N,m} \min\{|\beta_{ik}|, |\beta_{ik} - \gamma_{ik}|\} \tag{14}$$

taking a selection over multiple marginal penalizations on individualized coefficients, where $\lambda_{N,m}$ is a tuning parameter for penalization level.This MDSP term applies in (3) with a double summation over both individuals and covariates.

First, from an individual-wise point of view, given $\gamma_k' s$, the penalty term $\sum_{k=1}^{p} s_{\lambda}(\beta_{ik}, \gamma_{ik})$ carries feature selection on the ith individualized coefficients $\boldsymbol{\beta}_i$.

Table of Contents

1. Variable Selection
   - Lasso
   - Elastic net
   - Asymptotically unbiased concave penalties
2. Subpopulation structure identification
   - Fused Lasso
   - OSCAR
   - Grouping pursuit via truncated L_1-penalty for fusions
   - Concave pairwise fusion
   - MDSP
3. β-model of network
   - Monotonicity Lemma
   - Subgroup Identification for β-model

β-model

The β-model assumes that A_{ij}'s are generated as independent Bernoulli random variables with

$$P(A_{ij}=1)=p_{ij}=\frac{e^{\beta_i+\beta_j}}{1+e^{\beta_i+\beta_j}},$$

where $\boldsymbol{\beta}=(\beta_1,\beta_2,...,\beta_n)^\top$ is an unkown parameter. The resulting log-likelihood under the β-model is

$$\sum_{i=1}^{n}\beta_i d_i-\sum_{1\le i<j\le n}\log(1+e^{\beta_i+\beta_j})$$

and the degree sequence $\boldsymbol{d}=(d_1,...,d_n)^\top$ is a sufficient statistic,

Monotonicity Lemma

For the β-model, we have an important lemma:

Lemma

(Monotonicity Lemma) Given the degree sequence $\boldsymbol{d} = (d_1, ..., d_n)^\top$*, the maximum likelihood estimator must satisfy:*

- *If* $d_i = d_j$*, then* $\hat{\beta}_i = \hat{\beta}_j$.
- *If* $d_i < d_j$*, then* $\hat{\beta}_i \leq \hat{\beta}_j$.

This lemma implies that the estimator sequence must have the same order as the degree sequence, by which the computation can be significantly simplified.

Description of problem

Assume that in the β-model, there are two subgroups of $\boldsymbol{\beta}$, i.e. $\{1, 2, ..., n\} = S_1 \cup S_2$, $\beta_i = \gamma_1$ for any $i \in S_1$ and $\beta_j = \gamma_2$ for any $j \in S_2$. This can be described as the following optimization problem:

$$\boldsymbol{\beta} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^n} \ell_n(\boldsymbol{\beta}) \quad \text{s.t.} \quad \sum_{i=1}^{n} \min\{|\beta_i - \gamma_1|_0, |\beta_j - \gamma_2|_0\} \leq s$$

where $\ell_n(\boldsymbol{\beta}) = -\sum_{i=1}^{n} d_i \beta_i + \sum_{1 \leq i < j \leq n} \log(1 + e^{\beta_i + \beta_J})$ and s is a pre-specified parameter which equals to 0 in this problem especially. Our aim is to correctly identify the subgroup structure, classify each $i \in \{1, 2, ..., n\}$ to its subgroup and estimate the parameter γ_1 and γ_2.

My idea

Since in this problem, for any i, β_i is either γ_1 or γ_2, using the monotonicity lemma we can conclude that the subgroups are "continuous". First we rearrange the degree sequence as $\{d_{(i)}, i = 1, ..., n\}$ satisfying $d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(n)}$. Then we will obtain $\beta_{(1)}, ..., \beta_{(t)} = \gamma_1$ and $\beta_{(t+1)}, ..., \beta_{(n)} = \gamma_2$. Hence, we only need to consider $n+1$ situations, i.e. $t = 0, 1, ..., n$. For each situation, since we already know the subgroup structure, the problem becomes an unconstrained optimization problem and we only need to calculate the maximum likelihood estimation and compare the n different values of the likelihood function. The situation corresonding to the largest value of likelihood function is accepted.

Monotonicity lemma in this model

In our model, we can easily prove the following results:

Lemma

(Monotonicity Lemma) Given the degree sequence $\boldsymbol{d} = (d_1, ..., d_n)^\top$, the maximum likelihood estimator must satisfy:

- *If $d_i < d_j$, then $\hat{\beta}_i \leq \hat{\beta}_j$.*

Therefore, by indexing the degree sequence, only the notes between two different groups which share same degree cannot be distinguished. However, definitely we cannot classify notes with same degree into different groups because they have the same likelihood value and in practice, we will assume notes belonging to different groups must have different degree.

Simulation results

I conduct a series of simulations to ground this method. Specifically, I consider there are $n = 100$ notes and $s1$ notes share $\beta_i = \gamma_1 = -1/3 \log n$ while other $s2 = n - s1$ notes share $\beta_j = \gamma_2 = 1/5 \log n$. $s1$ ranges from 1 to $n - 1 = 99$.

I consider three criterion for evaluating the performance of this method, including

- Fre: Frequency, the possibility that we identify true subgroups and correctly classify notes.
- MSRE: Mean Square Root Error.
- RE: Relative Error.

The simulation results show that the method identifies two subgroups and classifies notes almost perfectly. Also the estimation of $\boldsymbol{\beta}$ is accurate.

References

[1] Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," Journal of the American Statistical Association, 96, 1348–1360.

[2] Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010), "Pairwise Variable Selection for High-Dimensional Model-Based Clustering," Biometrics, 66, 793–804.

[3] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive Mixtures of Local Experts," Neural Computation, 3, 79–87.

[4] Ma, S., and Huang, J. (2017), "A Concave Pairwise Fusion Approach to Subgroup Analysis," Journal of the American Statistical Association, 112, 410–423.

[5] Pan, W., and Shen, X. (2007), "Penalized Model-Based Clustering With Application to Variable Selection," Journal of Machine Learning Research, 8, 1145–1164.

References

[1] Pan, W., Shen, X., and Liu, B. (2013), “Cluster Analysis: Unsupervised Learning via Supervised Learning With a Non-Convex Penalty,” The Journal of Machine Learning Research, 14, 1865–1889.

[2] Raftery, A. E., and Dean, N. (2006), “Variable Selection for Model-Based Clustering,” Journal of the American Statistical Association, 101, 168–178.

[3] Rinaldo, A. (2009), “Properties and Refinements of the Fused Lasso,” The Annals of Statistics, 37, 2922–2952.

[4] Shen, X., and Huang, H.-C. (2010), “Grouping Pursuit Through a Regularization Solution Surface,” Journal of the American Statistical Association, 105, 727–739.

[5] Shen, X., Huang, H.-C., and Pan, W. (2012), “Simultaneous Supervised Clustering and Feature Selection Over a Graph,” Biometrika, 99, 899– 914.

References

[1] Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," Journal of the Royal Statistical Society, Series B, 58, 267–288.

[2] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," Journal of the Royal Statistical Society, Series B, 67, 91–108.

[3] Wang, H., Li, R., and Tsai, C.-L. (2007), "Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method," Biometrika, 94, 553– 568.

[4] Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," Journal of the Royal Statistical Society, Series B, 68, 49–67.

[5] Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," The Annals of Statistics, 38, 894–942.

References

[1] Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," Journal of the American Statistical Association, 101, 1418–1429.

[2] Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," Journal of the Royal Statistical Society, Series B, 67, 301– 320.

[3] Sourav Chatterjee, Persi Diaconis, and Allan Sly. Random graphs with a given degree sequence. The Annals of Applied Probability, 21(4):1400–1435, 2011.

[4] Mingli Chen, Kengo Kato, and Chenlei Leng. Analysis of networks via the sparse β-model. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83:887–910, 2021. doi: https://doi.org/10. 1111/rssb.12444.

[5] Stefan Stein and Chenlei Leng. A sparse β-model with covariates for networks. arXiv preprint arXiv:2010.13604, 2020.

References

[1] Meijia Shao, Yu Zhang, Qiuping Wang, Yuan Zhang, Jing Luo and Tian Yan. $L-2$ regularized maximum likelihood for β-model in large and sparse networks. arXiv preprint arXiv:2110.11856v3, 2023.

Thank you!