A Two-stage Subgroup Analysis

Chen Canyi

May 8, 2020

Table of contents

Introduction

- Biological heterogeneity is common in many diseases.
- Thus there have been many attempts to identify subgroups from a heterogenous population.
- One of the most popular methods for finding subgroups from a heterogenous population is the mixture model.

The mixture model

- need specify the underlying distribution for data
- the number of subgroups
- not suitable for the high dimensional setting $p >> n$, which is a common situation in precision medicine with high dimension genetic covariates p and relative small sample size n.

The model

- In this project, we consider the following model:

$$Y_i = \mu_i + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, i = 1, \ldots, n,$$

  where Y_i is the response for the i-th subject, μ_i's are unknown subject-specific intercepts.

- We assume the sparsity structure where only a subset $\mathcal{S}$ of $\boldsymbol{\beta}$ is nonzero. We assume $|\mathcal{S}| = s$.

Historical literature

- In the situation $p < n$, Ma and Huang (2015) considered optimizing the following objective fucntion

$$Q_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda) = 1/2 \sum_{i=1}^{n} (Y_i - \mu_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{1 \le i < j \le n} p(|\mu_i - \mu_j|, \lambda).$$

- alternating direction method of multipliers (ADMM) algorithm: a first order algorithm.

Historical literature (Cont.)

A nature extension to $p > n$

$$\widetilde{Q}_n(\boldsymbol{\mu}, \boldsymbol{\beta}; \lambda) = 1/2 \sum_{i=1}^{n} (Y_i - \mu_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \alpha_1 \sum_{1 \le i < j \le n} p(|\mu_i - \mu_j|, \lambda) + \alpha_2 p(|\boldsymbol{\beta}|, \omega).$$

ADMM has to iterate on at lease $\binom{n}{2} + p$ parameters sequentially.

The algorithm

- Instead we first use Lasso procedure to select related covariates from $\mathbf{x}_i$, then apply k-means algorithm to the adjusted $\widetilde{Y}_i$. For the Lasso procedure, we minize the following objective function

$$(\widehat{\mu}(\lambda), \widehat{\boldsymbol{\beta}}(\lambda)) = \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min}\ 1/2\sum_{i=1}^{n}(Y_i - \mu - \mathbf{x}_i^T\boldsymbol{\beta})^2 + p(\boldsymbol{\beta}, \lambda). \tag{1}$$

- $\widetilde{Y}_i = Y_i - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}$

Choice of concave penalty function

- ℓ_1-norm penalty (Tibshirani, 1996, LASSO), smoothly clipped absolute deviations penalty (Fan and Li, 2001, SCAD) or minimax concave penalty (Zhang, 2010, MCP).
- our choice: $p(\boldsymbol{\beta}, \lambda) = \lambda |\boldsymbol{\beta}|_1$.

Optimization methods

- The first order methods, such as Combettes and Pesquet (2011), Bach et al. (2012) and Tropp and Wright (2010), and Newton type algorithms, such as Fountoulakis et al. (2014) and Dassios et al. (2015), can be used to optimize (1).
- We suggest to use the primal dual active set method (Fan et al., 2014), which is in spirit a generalized version of Newton type method. It usually converges after one-step iteration if there is a very good initial value.

The remains: One-dimension cluster

- It remains to analysis the subgroups in $\widetilde{\mathbf{Y}} = (\widetilde{Y_1}, \ldots, \widetilde{Y_n})^T$.
- e.g., k-means, k-median, k-modes, even the deep learing algorithm.
- It is NP-hard in a general Euclidean space, even when the number of clusters k is 2 (Aloise et al., 2009).
- optimal one-dimensional clustering: $O(qknp)$ v.s. $(knlgn)$

Selection of the tuning parameters k and λ

- We select the tuning parameter λ by validation.
- Bayesian Information Criterion

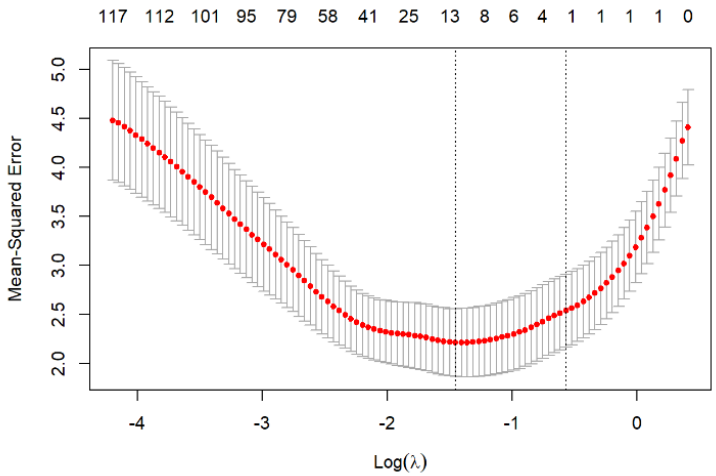
$$\mathrm{BIC}(k) = \log\left[\sum_{i=1}^{n}\{Y_i - \widehat{\mu}_i(k) - \mathbf{x}_i^T\widehat{\boldsymbol{\beta}}\}^2/n\right] + C_n \log n/n(k+s),$$

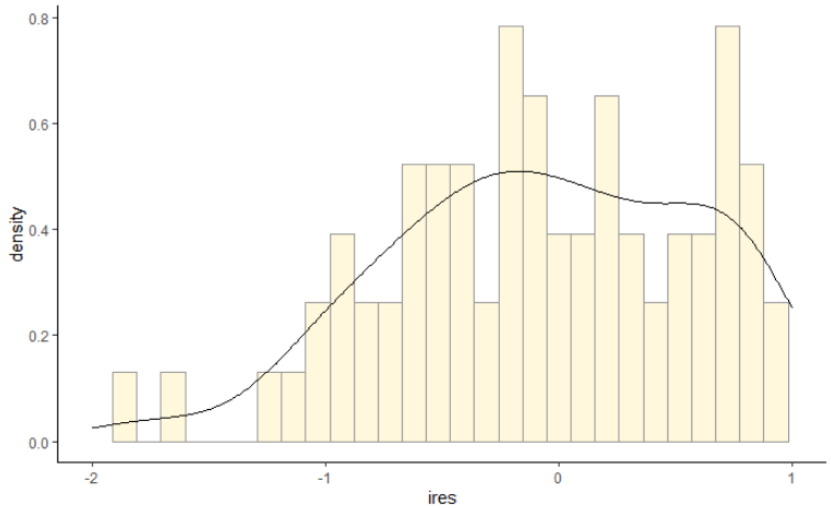
where $C_n = c\log\{\log(n+s)\}$ with some positive constant c. We take $c = 10$ by defaults.

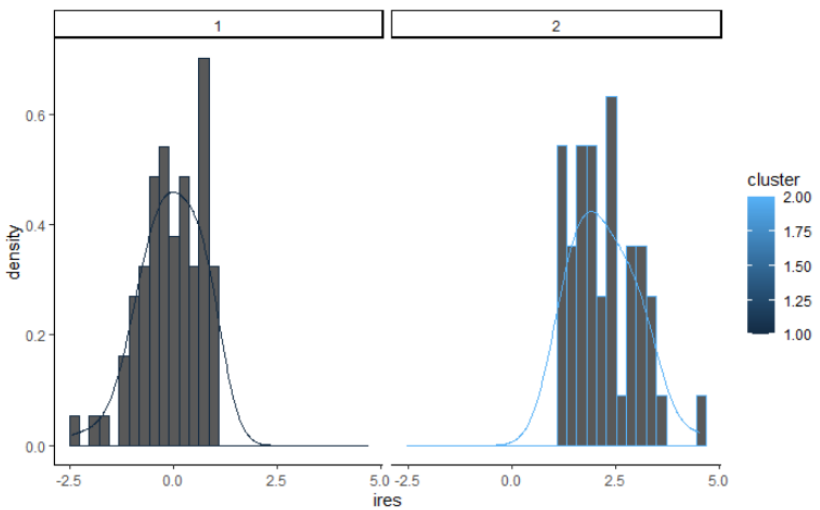
Chemores Data Example

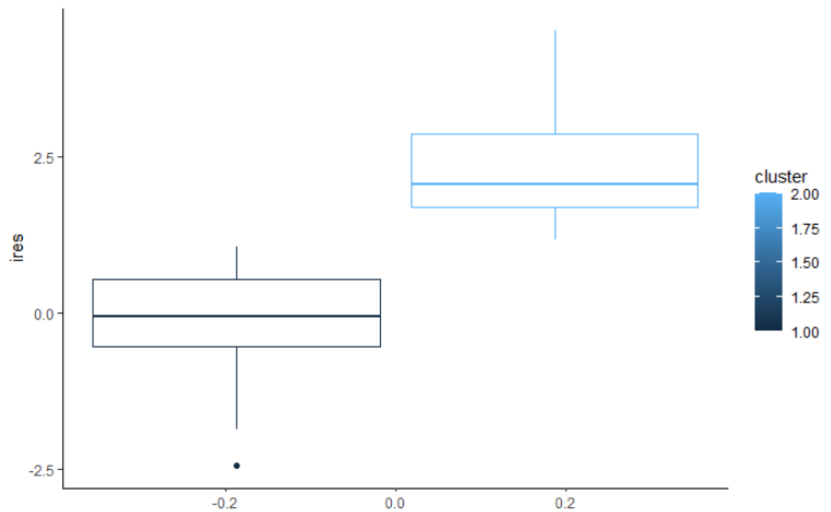
- Lung cancer is one of the most prevalent and deadiest cancers, which can be classified into two major subtypes, small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC). NSCLC accouting 80lung cancers, is a known heterogeneous group and its prognosis is generally poor (Tsuboi et al., 2007).
- Publicly available lung cancer genomic data from the Chemores Cohort Study. This data is part of an integrated study of mRNA, miRNA and clinical variables to characterize the molecular distinctions between squamous cell carcinoma (SCC) and adenocarcinoma (AC) in Non Small Cell Lung Cancer (NSCLC) aside large cell lung carcinoma (LCC).

Results


117 112 101 95 79 58 41 25 13 8 6 4 1 1 1 1 0
Mean-Squared Error
2.0 2.5 3.0 3.5 4.0 4.5 5.0
-4 -3 -2 -1 0
Log(λ)


density
0.0 0.2 0.4 0.6 0.8
-2 -1 0 1
ires


1
2
density
0.0 0.2 0.4 0.6
-2.5 0.0 2.5 5.0
ires
cluster
2.00
1.75
1.50
1.25
1.00


ires
-2.5 0.0 2.5
-0.2 0.0 0.2
cluster
2.00
1.75
1.50
1.25
1.00

Results