Penalized Likelihood in Bioinformatics

Xiaoqian Liu

About me

- Lecturer for today's class
- Postdoc at Dr. Wenyi Wang's Lab
- Research interests: Statistical/Machine Learning, Computational Statistics & Optimization, Bioinformatics & Cancer Biology
- Enthusiastic about teaching and mentoring. Feel free to reach out to me if you have any questions about today's lecture or my research.
- Personal website: https://xiaoqian-liu.github.io/

Outline

- Motivation
- Penalization methods
- Applications in Bioinformatics
- Recap on penalization

Motivation

Q: What are the characteristics of gene expression data?

- High-dimensionality

Motivation

Q: What are the characteristics of gene expression data?

- High-dimensionality
- Sparsity

Motivation

Q: What are the characteristics of gene expression data?

- High-dimensionality
- Sparsity
- High-correlation

Motivation

Q: What are the characteristics of gene expression data?

- High-dimensionality
- Sparsity
- High-correlation
- Group structure

Motivation

Q: What are the characteristics of gene expression data?

- High-dimensionality
- Sparsity
- High-correlation
- Group structure
- Homogeneity

Motivation

Q: What are the characteristics of gene expression data?

- High-dimensionality
- Sparsity
- High-correlation
- Group structure
- Homogeneity
- ……

Motivation

It is crucial to incorporate the characteristics (structure) of the gene expression data into the analysis and modeling process.

- Structure recovery: a fundamental task in data science.

Motivation

It is crucial to incorporate the characteristics (structure) of the gene expression data into the analysis and modeling process.

- Structure recovery: a fundamental task in data science.
- Q: But how?

Motivation

It is crucial to incorporate the characteristics (structure) of the gene expression data into the analysis and modeling process.

- Structure recovery: a fundamental task in data science.
- Q: But how?
- A: Penalization — A powerful strategy for dealing with "structured" data analysis and modeling.

Penalization methods

- Penalization/regularization is achieved through a penalty function that promotes the desired structure.
- **Penalized/regularized likelihood models** are in general in the following form:

$$\min_{\beta} \ \ell(\beta) + \lambda\psi(\beta),$$

where $\ell(\beta)$ is the log-likelihood function, $\psi(\beta)$ is the penalty function, and λ is the regularization parameter balancing the trade-off between model fitting and model complexity.

Penalization methods

Penalty functions covered in today's lecture:

- Lasso

Penalization methods

Penalty functions covered in today's lecture:

- Lasso
- SCAD

Penalization methods

Penalty functions covered in today's lecture:

- Lasso
- SCAD
- MCP

Penalization methods

Penalty functions covered in today's lecture:

- Lasso
- SCAD
- MCP
- Elastic Net

Penalization methods

Penalty functions covered in today's lecture:

- Lasso
- SCAD
- MCP
- Elastic Net
- Group Lasso/SCAD/MCP

Penalization methods

Penalty functions covered in today's lecture:

- Lasso
- SCAD
- MCP
- Elastic Net
- Group Lasso/SCAD/MCP
- Distance penalization

Penalization methods — Lasso

The Lasso (least absolute shrinkage and selection operator) penalty (Tibshirani 1996) is defined as

$$\psi(\beta) = \|\beta\|_1 = \sum_{i}^{p} |\beta_i| \tag{1}$$

- L_1 norm as the penalty function
- The pioneering work of sparsity learning in statistics and machine learning.

Penalization methods – Lasso

Consider a linear regression problem

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix containing p covariate variables, and $\epsilon \in \mathbb{R}^n$ is the Gaussian noise with mean 0 and variance σ^2.

- The maximum likelihood estimator (MLE) of β is

$$\beta_{MLE} \; = \; \min_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 \qquad (2)$$

Penalization methods – Lasso

Consider a linear regression problem

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where $y \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix containing p covariate variables, and $\epsilon \in \mathbb{R}^n$ is the Gaussian noise with mean 0 and variance σ^2.

- The Lasso penalized likelihood model is

$$\beta_{Lasso} = \min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{3}$$

Penalization methods – Lasso

- Solution path of Lasso

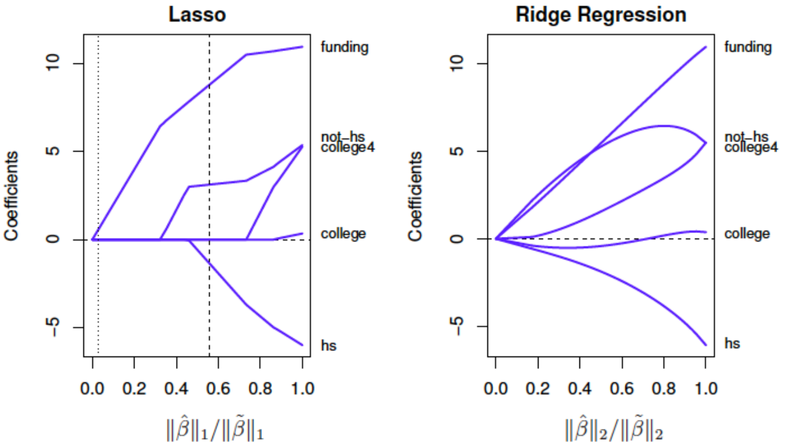

Figure 2.1 *Left: Coefficient path for the lasso, plotted versus the ℓ_1 norm of the coefficient vector, relative to the norm of the unrestricted least-squares estimate $\tilde{\beta}$. Right: Same for ridge regression, plotted against the relative ℓ_2 norm.*

Figure 1: Figure adopted from (Hastie, Tibshirani, and Wainwright 2015)

Penalization methods – Lasso

- Why can Lasso promote sparsity?

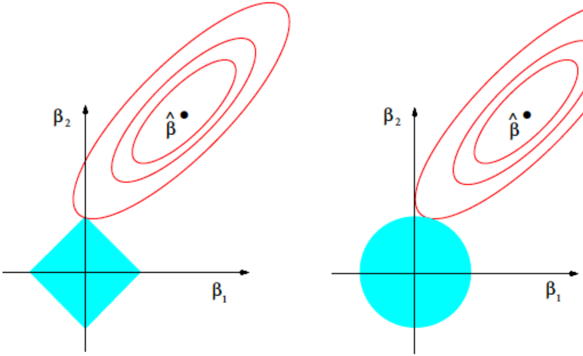

Figure 2.2 *Estimation picture for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions $|\beta_1|+|\beta_2| \leq t$ and $\beta_1^2+\beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the residual-sum-of-squares function. The point $\widehat{\beta}$ depicts the usual (unconstrained) least-squares estimate.*

Figure 2: Figure adopted from (Hastie, Tibshirani, and Wainwright 2015)

Penalization methods – Lasso

- How does Lasso promote sparsity?

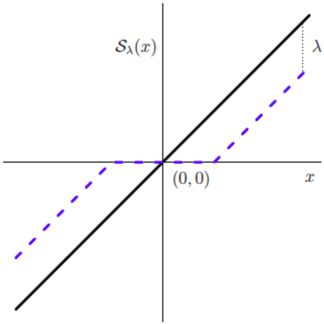

Figure 2.4 *Soft thresholding function* $\mathcal{S}_\lambda(x) = \text{sign}(x)\,(|x| - \lambda)_+$ *is shown in blue (broken lines), along with the* 45° *line in black.*

Figure 3: Figure adopted from (Hastie, Tibshirani, and Wainwright 2015)

Penalization methods – Lasso

- Advantages
  - Simplicity
  - Easy to compute
- Disadvantages
  - Underestimate large β_is, why?
  - Perform badly with correlated variables

Penalization methods – SCAD

To mitigate the underestimation of Lasso, one influential work by (Fan and Li 2001) is the smoothly clipped absolute deviations (SCAD) penalty:

$$\psi_\lambda(\beta) = \sum_{i=1}^{p} P(\beta_i; \lambda, \gamma)$$

Penalization methods – SCAD

The SCAD penalty:

$$\psi_\lambda(\beta) = \sum_{i=1}^{p} P(\beta_i; \lambda, \gamma)$$

where the univariate SCAD penalty is

$$P(x; \lambda, \gamma) = \begin{cases} \lambda|x|, & \text{if } |x| \le \lambda, \\ \frac{2\gamma\lambda|x| - x^2 - \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < |x| < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2}, & \text{if } |x| \ge \gamma\lambda, \end{cases} \tag{4}$$

for some $\gamma > 2$. Often, $\gamma = 3.7$ is used in practice.

Penalization methods – SCAD

Structure of SCAD:

- Coincide with Lasso when $|x| \leq \lambda$
- Transition to a quadratic function with $\lambda < |x| < \gamma\lambda$
- Remain as a constant for all $|x| \geq \gamma\lambda$

Penalization methods – MCP

A second option to mitigate the underestimation of Lasso is the minimax concave penalty (MCP, (Zhang et al. 2010)):

$$\psi_\lambda(\beta) = \sum_{i=1}^{p} P(\beta_i; \lambda, \gamma)$$

where the univariate MCP is

$$P(x; \lambda, \gamma) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma}, & \text{if } |x| \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |x| > \gamma\lambda, \end{cases} \tag{5}$$

for some $\gamma > 1$. Often, $\gamma = 3$ is used in practice.

Penalization methods – MCP

Structure of MCP:

- A quadratic function with $|x| \leq \gamma\lambda$
- A constant for all $|x| > \gamma\lambda$

Penalization methods

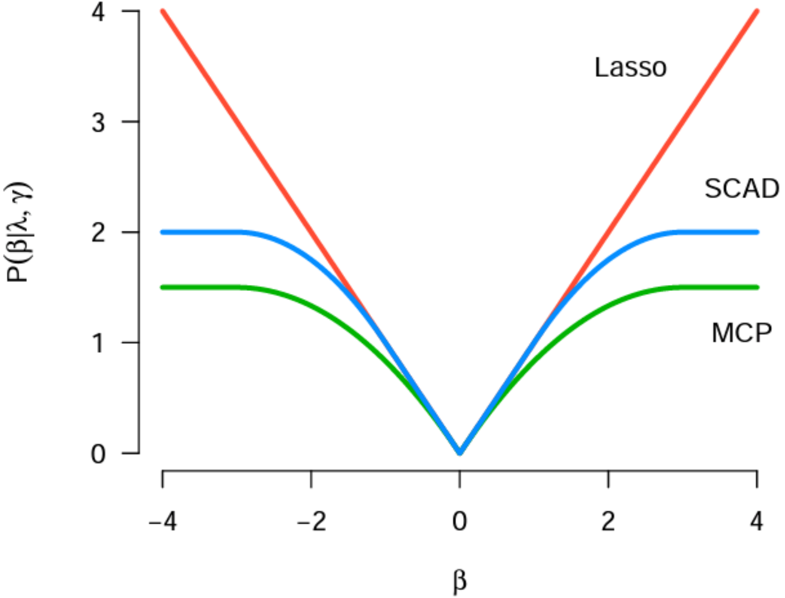

Figure 4: Visualization of Lasso, SCAD, and MCP (from Patrick

Penalization methods

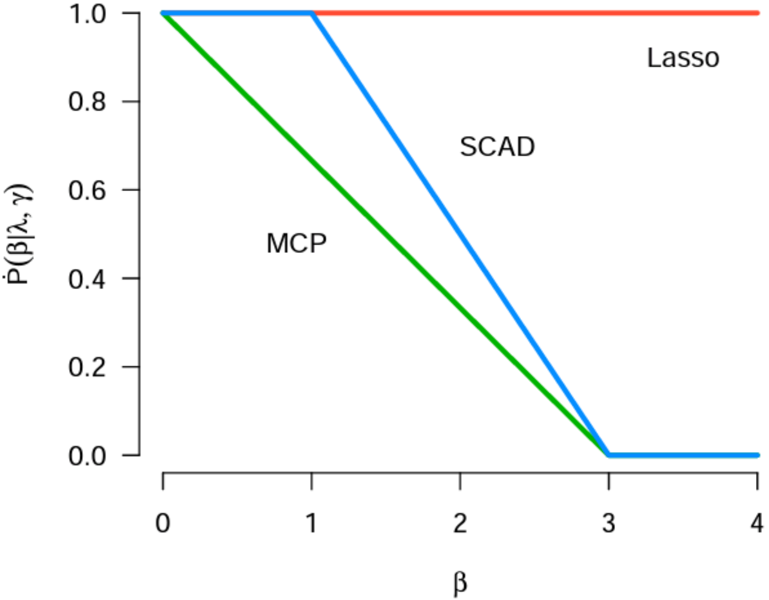

Figure 5: Visualization of derivatives of Lasso, SCAD, and MCP (from

Penalization methods – Elastic Net

- How to deal with correlated variables?

The elastic net penalty (Zou and Hastie 2005) is defined as

$$P_\lambda(\beta) = \lambda(\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2), \tag{6}$$

which is a combination of the L_1-penalty (Lasso) and the squared L_2-penalty (ridge).

Penalization methods – Elastic Net

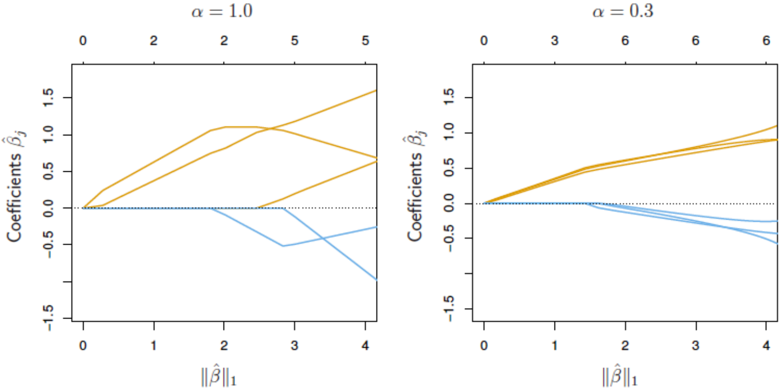

Figure 4.1 *Six variables, highly correlated in groups of three. The lasso estimates ($\alpha = 1$), as shown in the left panel, exhibit somewhat erratic behavior as the regularization parameter λ is varied. In the right panel, the elastic net with ($\alpha = 0.3$) includes all the variables, and the correlated groups are pulled together.*

Figure 6: An illustrative comparison of Lasso and Elastic Net on correlated features. Figure adopted from (Hastie, Tibshirani, and Wainwright 2015)

Penalization methods – Group Lasso

Consider a linear regression problem

$$\mathbf{y} = \mathrm{X}\beta + \epsilon$$

- Covariate variables in X have natural group structures

  e.g. categorical variables
- Aim: select (or not) a whole group of variables

Penalization methods – Group Lasso

Group Lasso (Yuan and Lin 2006) extends the Lasso penalty to the group selection (group sparsity) scenario. The group Lasso penalty is defined as

$$\psi(\beta) = \sum_{j=1}^{J} K_j \|\beta_j\|_2, \tag{7}$$

- $\beta = (\beta_1^T, ..., \beta_J^T)^T \in \mathbb{R}^p$ with $\beta_j \in \mathbb{R}^{p_j}$
- K_j: adjust for the group sizes, e.g. $K_j = \sqrt{p_j}$

Penalization methods – Group Lasso

Why group Lasso can promote sparsity at the group level?

- It applies Lasso to the L_2 norm of each subvector of each group

  Lasso-type penalization at the group level;

  Ridge-type penalization at the individual level.

- Want the sparsity at the individual level as well?

  It is called bi-level variable selection (See Homework).

Penalization methods – Group SCAD/MCP

- Can SCAD and MCP be extended to the group selection scenario?

  Yes!

- A more general class of group selection penalties:

$$\psi(\beta) = \sum_{j=1}^{J} P(\|\beta_j\|_2; K_j\lambda, \gamma), \tag{8}$$

where P is the univariate SCAD or MCP penalty.

Penalization methods – Distance penalization

Consider a very general setting

$$\min_{\beta} \ \ell(\beta) \quad \text{subject to} \ \ \beta \in C, \tag{9}$$

where $\ell(\beta)$ is the negative log-likelihood, and C is the constraint set that specifies the required structure on β.

- Very general in the sense that the structure of β is coded as a constraint on β.
- Sparsity case: $C = \{\beta : \|\beta\|_0 \leq k\}$ with k as an positive integer controlling the sparsity of β.

Penalization methods – Distance penalization

Distance penalization for constrained estimation

$$\min_{\beta} \ \ell(\beta) + \frac{\lambda}{2}\mathsf{dist}(\beta, C)^2. \tag{10}$$

where

$$\frac{1}{2}\mathsf{dist}(\beta, C)^2 = \min_{u \in C} \frac{1}{2}\|\beta - u\|_2^2. \tag{11}$$

Applications in Bioinformatics

Sparse logistic regression in cancer classification

- Data: leukemia patient samples
  - acute lymphoblast leukemia (ALL), 49 samples
  - acute myeloid leukemia (AML), 23 samples
  - each sample contains the profile of 7129 genes
  - available at https://search.r-project.org/CRAN/refmans/propOverlap/html/leukaemia.html
- Aim: leukemia subtype classification & gene selection

Applications in Bioinformatics

Sparse logistic regression in cancer classification

Consider a general binary classification problem. The data is given in the format $\{y_i, x_i\}_{i=1}^n$, where $y_i \in \{0, 1\}$ indicates the class label and $x_i \in \mathbb{R}^p$ contains the p covaraiate variables of the i-th sample.

The (linear) logistic regression model assumes the following conditional probability:

$$Pr(y = 1|x) = \frac{e^{x^T\beta}}{1 + e^{x^T\beta}}$$

Applications in Bioinformatics

Sparse logistic regression in cancer classification

The logistic model is fitted by minimizing the negative binomial log-likelihood of the data

$$\min_{\beta} \; -\ell(\beta) + \lambda \|\beta\|_1 \tag{12}$$

- $\ell(\beta) = \sum_{i=1}^{n} [y_i x_i^T \beta - \log(1 + x_i^T \beta)]$ is the negative log-likelihood
- $\|\beta\|_1$ is the penalty term for sparsity
- λ is the regularization parameter

Applications in Bioinformatics

Penalized likelihood for scRNA-seq data analysis

- UMI count data
  - For gene g in cell c, the UMI count is x_{gc}
- What's the distribution of x_{gc}?
  - Binomial distribution

$$x_{gc} \sim NB(\mu_{gc}, \theta_g),\ \ln \mu_{gc} = \beta_{g0} + \ln n_c$$

where θ_g is the gene-specific dispersion parameter, $n_c = \sum_g x_{gc}$ is the total sequencing depth and the variance of the NB distribution is $\mu_{gc} + \mu_{gc}^2/\theta_g$.

Applications in Bioinformatics

Penalized likelihood for scRNA-seq data analysis

- UMI count data
  - For gene g in cell c, the UMI count is x_{gc}
- What's the distribution of x_{gc}?
  - Zero-inflated mixture distribution

$$Pr(x_{gc} = x) = (1 - \pi_g)I(x = 0) + \pi_g I(x \neq 0)F(x|\mu_{gc}, \sigma_g^2)$$

Applications in Bioinformatics

Penalized likelihood for scRNA-seq data analysis

- Penalization in scRNA-seq data analysis?
  - clustering / cell cell subgroup detection
  - gene selection
  - Other tasks

Recap on Penalization

- Penalization is a strategy
  - not just for sparsity; not only for likelihood-based models
- A general penalization framework:

$$\min_{\beta} \text{loss}(\beta) + \text{penalty}(\beta) \qquad (13)$$

  - loss(β) is derivaed from the specific problem
  - penalty(β) is defined according to the structure of β
- Penalization in other applications:

  classification/clustering/PCA/CCA/matrix recovery

Recap on Penalization

Penalization in classification — L_1-regularized SVM

$$\min_{\beta} \ \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(x_i;\beta)]_+ + \lambda\|\beta\|_1 \tag{14}$$

- The first term is known as the hinge loss.
- If $f(x_i;\beta) = x_i^T\beta$, then it's a linear SVM.
- The second term is the penalty term promoting sparsity in β.

References

Fan, Jianqing, and Runze Li. 2001. "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association* 96 (456): 1348–60.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright. 2015. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–88.

Yuan, Ming, and Yi Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1): 49–67.

Zhang, Cun-Hui et al. 2010. "Nearly Unbiased Variable Selection Under Minimax Concave Penalty." *The Annals of Statistics* 38 (2): 894–942.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical*