高等统计选讲
— 高维统计分析入门

贾金柱
北京大学数学科学院

Contents

Preface

本书是为高年级研究生而写。适用于高年级统计学专业的研究生。

Chapter 1

高维统计学概况

高维统计学是目前统计学界热门的研究课题。究其原因，无外乎两个：1. 实际的科学问题需要高维统计学；2. 高维统计学的理论尚不健全，传统统计方法（大数定律、中心极限定理）不再适用。

一些实际的例子：1. 文本分类。2. 图像注释。3. 基因选择。4. 计算机程序的运行时间预测。等等。

本课程将通过对目前高位统计学前沿知识的介绍，引导同学和研究者进入高维统计学领域。本书内容包括：Convex Optimization, Concentration Inequality, Lasso 的理论分析（L2 consistency and sign consistency）, Lasso 的求解方法，模型评价方法和模型选择准则以及高维统计方法的最新发展，包括稀疏图模型，Dictionary Learning，Sparse PCA，Sparse Non-linear models 和Boosting.

课程要求：本课程要求学生学习过高等数学，概率论，和统计学。熟悉编程语言，C，matlab，和R。

作业：一般一个Topic 一次作业。两周的时间完成作业，不能以任何理由晚交作业。作业很重要，能够帮助理解上课所学知识。

期末考核：作业+ 期末论文。作业30%, 期末论文70%。

Chapter 2

Convex Optimization

通过本章的学习，学生能够掌握：

- Convex optimization 最优解满足的条件（KKT conditions）
- 如何求解一般的Convex Optimization
- Lasso 作为一般Convex Optimization 的一些性质

2.1 Convex Optimization

Optimization problem 的标准形式：

$$\begin{aligned} &minimize && f_0(x) \\ &subject\ \ to && f_i(x) \le 0, i = 1, \dots, n, \\ & && h_j(x) = 0, j = 1, \dots, m. \end{aligned}$$

如果目标函数和不等式约束函数都是凸函数，等式约束函数是线性函数，该优化称为Convex Optimization.

$$f_i(\alpha x + \beta y) \le \alpha f_i(x) + \beta f_i(y),$$

如果 $\alpha + \beta = 1, \alpha \ge 0, \beta \ge 0$. 用 p^* 代表上述问题的最小值。

一些术语：feasible, optimal 和locally optimal。

feasible: x is feasible, if $x \in \mathbf{dom} f_0$ 并且它满足所有的约束。optimal: x is optimal, if x is feasible, 并且$f_0(x) = p^*$. locally optimal: x is locally optimal, if there is an $R > 0$, 使得x 在局部范围内：

$$\{z : \|z - x\|_2 \le R\},$$

是下面问题的最优解(optimal).

$$\begin{aligned} &minimize(over\ z) && f_0(z) \\ &subjectto && f_i(z) \le 0, i = 1, \dots, n, \\ & && h_j(z) = 0, j = 1, \dots, m, \\ & && \|z - x\|_2 \le R. \end{aligned}$$

定理1. *Convex problem* 的局部最优解就是全局最优解。*(作业)*

2.2 Duality

2.2.1 Lagrange dual problem

Lagrangian:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x) + \sum_{j=1}^{m} \nu_j h_j(x)$$

Lagrange dual function is defined as:

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu),$$

where $\mathcal{D}$ 是 $f_0(x)$ 的定义域，记为 $\mathbf{dom} f_0$.

性质：

- $g(\lambda, \nu)$ is concave;
- Lower bound property: If $\lambda \succeq 0$, then $g(\lambda, \nu) \leq p^*$.

Proof. Suppose $\bar{x}$ is a feasible point. Then

$$L(\bar{x}, \lambda, \nu) \leq f_0(\bar{x}).$$

Hence

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leq L(\bar{x}, \lambda, \nu) \leq f_(\bar{x}) \leq p^*$$

□

The (Lagrange) dual problem:

$$\begin{aligned} &maximize \quad g(\lambda, \nu) \\ &subject \ \ to \quad \lambda \geq 0 \end{aligned} \tag{2.1}$$

因为 $g(\lambda, \nu)$ 是一个 concave function, 这是一个 convex problem，不管原问题是不是 convex problem. 记该问题的最优值为d^*, 那么d^* 是最好（大）的 lower bound, in some sense. 该性质称为 weak duality.

We refer to a pair (λ, ν) with $\lambda \succ 0$ and $g(\lambda, \nu) > -\infty$ as dual feasible.

2.2.2 Weak and strong duality

Weak duality: $d^* \leq p^*$

- always holds
- can be used to find non-trivial lower bounds for difficult problems.

Example: Two-way Partitioning

$$\begin{aligned} &minimize && x^T W x \\ &subject\ \ to && x_i^2 = 1, i = 1, \ldots, n. \end{aligned} \tag{2.2}$$

This is a non-convex problem and can be interpreted as a partition problem. W_{ij} is the cost of assigning i, j to the same set; $-W_{ij}$ is the cost of assigning to different sets.

Dual function:

$$\begin{aligned} g(v) &= \inf_x x^T W x + \sum_i \nu_i (x_i^2 - 1) \\ &= \inf_x x^T \left[W + diag(v) \right] x - \mathbf{1}^T \nu \\ &= \begin{cases} -\mathbf{1}^T \nu, & W + diag(\nu) \succcurlyeq 0; \\ -\infty, & O.W. \end{cases} \end{aligned}$$

$g(\nu)$ is a lower bound of p^* for any ν. By taking $\nu = -\lambda_{\min}(W)\mathbf{1}$, we have $p^* \geq n\lambda_{\min}(W)$. A better lower bound is given by the following convex optimization problem (SDP):

$$\begin{aligned} &maximize && -\mathbf{1}^T \nu \\ &subject\ \ to && W + diag(\nu) \succcurlyeq 0. \end{aligned} \tag{2.3}$$

Strong duality: $d^* = p^*$.

- does not always holds
- (usually) holds for convex problems
- conditions that guarantee strong duality in convex problems are called **constraint qualifications.**

2.2.3 Slater's constraint qualification

strong duality holds for a convex problem

$$\begin{aligned} &minimize && f_0(x) \\ &subject\ \ to && f_i(x) \le 0, i = 1, \dots, n, \\ & && A_{m\times p}x = b \end{aligned} \tag{2.4}$$

if it is strictly feasible, *i.e.*,

$$\exists x \in \mathbf{int}\ \mathcal{D} : f_i(x) < 0, i = 1, \dots, n, Ax = b$$

定理2. *Slater's constraint qualification guarantees strong duality for a convex problem.*

定理3 (Separating hyperplane theorem)**.** *Suppose C and D are two convex sets that do not intersect. Then there exists $a \neq 0$ and b such that $a^T x \le b,\ \forall x \in C$ and $a^T x \ge b,\ \forall x \in D$.*

Proof. 此证明的关键点在于两点：1. 目标函数和所有的约束函数是凸函数；2. 存在内点，使得所有的约束函数严格成立。

We assume $rank(A) = m$ and p^* is finite ($p^* = -\infty$ is trivial).

Let $\mathcal{A} \subseteq \mathbb{R}^n \times \mathbb{R}^m \times R$, $\mathcal{A} = \{(u, v, t) | \exists x \in \mathcal{D}, f_i(x) \le u_i, i = 1, \dots, n; h_j(x) = v_j, j = 1, \dots, m, f_0(x) \le t\}$, where $u = (u_1, \dots, u_n)$ corresponds to the n inequality constraints, ν corresponds to the m equality constraints and $h_j(x) = [Ax - b]_j$.

Obviously, $\mathcal{A}$ is a convex set. Define another convex set

$$\mathcal{B} = \{(0, 0, s) \in \mathbb{R}^n \times \mathbb{R}^m \times R | s < p^*\}.$$

We conclude that

$$\mathcal{A} \cap \mathcal{B} = \emptyset.$$

To see this, suppose that $(u, v, t) \in \mathcal{A} \cap \mathcal{B}$. Since $(u, v, t) \in \mathcal{B}$, we have $u = 0, v = 0$, and $t < p^*$. Since $(u, v, t) \in \mathcal{A}$, there exists an x with $f_i(x) \le 0, i = 1, \dots, n; Ax - b = 0$, and $f_0(x) \le t < p^*$, which is not impossible since p^* is the optimal value of the primal problem.

By the separating hyperplane theorem, there exists $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$ and $\alpha\mathbb{R}$ such that

$$\tilde{\lambda}^T u + \tilde{\nu} v + \mu t \ge \alpha, \forall (u, v, t) \in \mathcal{A} (1)$$

and

$$\tilde{\lambda}^T u + \tilde{\nu} v + \mu t \le \alpha, \forall (u, v, t) \in \mathcal{B} (2)$$

From (1), we conclude that $\tilde{\lambda} \succeq 0$ and $\mu \succeq 0$. (2) means $\mu t \leq \alpha$ for all $t < p^*$. Hence $\mu p^* \leq \alpha$. Together with (1), we have for any $x \in \mathcal{D}$,

$$\sum_{i=1}^{n} \tilde{\lambda}_i f_i(x) + \tilde{\nu}^T(Ax - b) + \mu f_0(x) \geq \alpha \geq \mu p^*. (3)$$

Now assume that $\mu > 0$(later we will show that $\mu \neq 0$). We divide (3) by μ to obtain

$$L(x, \tilde{\lambda}/\mu, \tilde{\nu}/\mu) \geq p^*$$

for all $x \in \mathcal{D}$. By minimizing over x, we have $g(\lambda, \nu) \geq p^*$, where $\lambda = \tilde{\lambda}/\mu, \quad \nu = \tilde{\nu}/\mu$. So $d^* = maximizeg(\lambda, \nu) \geq p^*$, By weak duality, we have $d^* \leq p^*$, so $d^* = p^*$.

Now we show that $\mu \neq 0$ (we will use the condition that $\exists x \in \mathbf{int}\mathcal{D}$ which is strictly feasible). Suppose that $\mu = 0$. Then from (3), for all $x \in \mathcal{D}$,

$$\sum_{i=1}^{n} \tilde{\lambda}_i f_i(x) + \tilde{\nu}^T(Ax - b) \geq 0.$$

For $\tilde{x}$ which satisfies the Slater condition, we have

$$\sum_{i=1}^{n} \tilde{\lambda}_i f_i(x) \geq 0.$$

Since $f_i(\tilde{x}) < 0$ and $\tilde{\lambda}_i \geq 0$, we conclude that $\tilde{\lambda} = 0$. From $(\tilde{\lambda}, \tilde{n}u, \mu) \neq 0$, we have $\nu \neq 0$. (3) implies that $\tilde{v}^T(Ax - b) \geq 0$ for all $x \in \mathcal{D}$. Since $v^T(A\tilde{x} - b) = 0$ and $\tilde{x} \in \mathbf{int}\mathcal{D}$, there are points in $\mathcal{D}$ with $\tilde{v}^T(Ax - b) < 0$ unless $A^T\tilde{v} = 0$. This contradicts our assumption that $rank(A) = m$. $\square$

2.2.4 Complementary slackness

Suppose that the primal and dual optimal values are attained and equal. Let x^* be a primal optimal point and (λ^*, ν^*) be a dual optimal point.

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, v^*) \\ &= \inf\left(f_0(x) + \sum_{i=1}^{n} \lambda^* f_i x + \sum_{j=1} m\nu^* h_j(x)\right) \\ &\leq f_0(x^*) + \sum_{i=1}^{n} \lambda_i^* f_i(x^*) + \sum_{j=1}^{m} \nu_j^* h_j(x^*) \\ &\leq f_0(x^*), \end{aligned}$$

from which, we have

- x^* minimizes $L(x, \lambda^*, \nu^*)$;
- $\lambda_i \times f_i(x^*) = 0$ (Complementary Slackness).

2.2.5 Karush-Kuhn-Tucker (KKT) conditions

The following four conditions are called KKT conditions (for a problem with differentiable f_i,h_j):

- primal constraints: $f_i(x) \leq 0$ and $h_j(x) = 0$
- dual constraints: $\lambda \succeq 0$
- complementary slackness: $\lambda_i f_i(x) = 0$
- gradient of Lagrangian w.r.t. x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^{n} \lambda_i \nabla f_i(x) + \sum_{j=1}^{m} \nu_j \nabla h_j(x) = 0$$

.

定理4 (Necessary)**.** *For any optimization problem with differentiable objective and constraint functions for which strong duality obtains, any pair of primal and dual optimal points must satisfy KKT conditions.*

For a convex problem, KKT conditions are also sufficient.

定理5 (Sufficient)**.** *For any convex optimization problem with differentiable objective and constraint functions, any points that satisfy the KKT conditions are primal and dual optimal, and have zero duality gaps.*

Proof. Suppose $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ satisfy the KKT conditions, then $L(x, \tilde{\lambda}, \tilde{\nu})$ is convex in x and $\tilde{x}$ minimizes $L(x, \tilde{\lambda}, \tilde{\nu})$. So we have

$$\begin{aligned} g(\tilde{\lambda}, \tilde{\nu}) &= L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}) \\ &= f_0(\tilde{x}) + \sum_{i=1}^{n} \tilde{\lambda}_i f_i(\tilde{x}) + \sum_{j=1}^{m} \nu_j h_j(\tilde{x}) \\ &= f_0(\tilde{x}). \end{aligned}$$

Hence

$$f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{\nu}) \leq d^* \leq p^* \leq f_0(\tilde{x}),$$

$\Longrightarrow$

$$d^* = p^* \leq f_0(\tilde{x}).$$

□

定理6 (Necessary and Sufficient.)**.** *If a convex optimization problem with differentiable objective and constraint functions satisfies Slater's condition, then KKT conditions provide necessary and sufficient conditions for optimality: x is optimal if and only if there are (λ, ν) that, together with x, satisfy the KKT conditions.*

2.2.6 Subgradient

Recall basic inequality for convex differentiable function f:

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

What if f is not differentiable?

定义1. *g is a subgradient of f (not necessarily convex) at x, if*

$$f(y) \geq f(x) + g^T (y - x) \text{ for all } y$$

.

- if f is convex, it has at least one subgradient at very point.
- if f is convex and differentiable, $\nabla f(x)$ is a subgradient of f at x.

定义2. *set of all subgradients of f at x is called the subdifferential of f at x, written as $\partial f(x)$.*

Recall for f convex, differentiable,

$$f(x^*) = \inf_x f(x) \Longleftrightarrow 0 = \nabla f(x),$$

generalization to nondifferentiable convex f:

$$f(x^*) = \inf_x f(x) \Longleftrightarrow 0 \in \partial f(x^*),$$

Proof.

$$f(x^*) \leq f(y) \Longleftrightarrow f(y) \geq f(x^*) + 0^T(y - x^*) \Longleftrightarrow 0 \in \partial f(x^*).$$

□

KKT conditions: (1)(2)(3) the same; (4) changes to

$$0 \in \nabla f_0(x) + \sum_{i=1}^{n} \lambda_i \nabla f_i(x) + \sum_{j=1}^{m} \nu_j \nabla h_j(x)$$

.

Some properties:

- $$\partial(\alpha f) = \alpha \partial f$$

- $$\partial(f_1 + f_2) = \partial f_1 + \partial f_2 \text{ (RHS is addition of sets) (作业)}$$

2.3 The Lasso and its duality

Let Y denote the vector of observed responses. $X = (X_1, \ldots, X_p)$ be the $n \times p$ matrix. X_j is the jth column of X. Let $\beta \in \mathbb{R}^p$. The Lasso:

$$min \ \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{2.5}$$

$\Leftrightarrow$

$$\begin{aligned} &minimize \quad \tfrac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n} x_{ij}\beta_j\right)^2 \\ &subject \ \ to \qquad \sum_{j=1}^{p}|\beta_j| \leq t, \end{aligned} \tag{2.6}$$

for some t.

The Lagrangian is

$$L(\beta, \lambda) = \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda(\sum_{j=1}^{p}|\beta_j| - t).$$

The dual function: $g(\lambda) = \inf_\beta L(\beta, \lambda) = \inf_\beta \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda(\sum_{j=1}^{p}|\beta_j| - t)$

If the strong duality holds, then we see that

$$d^* = \max_{\lambda} g(\lambda) = g(\lambda^*) = \inf_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda^*(\sum_{j=1}^{p} |\beta_j| - t),$$

we conclude that (2.5) and (2.6) have the same solution.

定理7 (Strong duality)**.** *Strong duality of the Lasso problem holds.*

Proof. We just need to show that Slater's conditions hold. Suppose $t \neq 0$. $0 \in \mathbf{int} f_0(\beta)$ and $0 = \sum_{j=1}^{p} |\beta_j| < t$ holds. When $t = 0$, the problem is trivial, because the solution of the prime problem is 0 and $p^* = \frac{1}{2}\|Y\|_2^2$. For the dual problem, by weak duality, we always have $d^* \leq p^*$. It is easy to see that when $\lambda > \max_j |X_j^T Y|$, $g(\lambda) = \frac{1}{2}\|Y\|_2^2 = p^*$. □

引理1. *If* $\lambda > \max_j |X_j^T Y|$*, then* 0 *is one minimizer of*

$$\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Proof. Let $\hat{\beta}$ is one minimizer of $\min_\beta \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$. Then

$$0 \in \partial[\frac{1}{2}\|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|]|_{\beta=\hat{\beta}}$$

.

$$\partial \frac{1}{2}\|Y - X\beta\|_2^2 = -X^T(Y - X\beta)$$

$$\partial[\lambda \sum_{j=1}^{p} |\beta_j|] = \lambda \mathbf{s} = \lambda(s(\beta_1), \ldots, s(\beta_j), s(\beta_p))^T,$$

where

$$s(x) = \begin{cases} 1, & x > 0; \\ \alpha \in [-1, 1], & \text{x} = 0; \\ -1, & x < 0. \end{cases}$$

So we have

$$X^T(Y - X\hat{\beta}) = \lambda \mathbf{s}.$$

Since $|X^T Y| \prec \lambda$, we see that $\hat{\beta} = 0$ satisfy the above equality by taking $s(\hat{\beta}_j) = \frac{X_j^T Y}{\lambda} \in [-1, 1]$. □

Now we see that KKT conditions are necessary and sufficient $\hat{\beta}$ to be the solution of the Lasso.

KKT conditions:

$$\begin{aligned} &1. \quad \textstyle\sum_{j=1}^{p} |\hat{\beta}_j| \le t \\ &2. \quad \lambda \ge 0 \\ &3. \quad \lambda[\textstyle\sum_{j=1}^{p} |\beta_j| - t] = 0 \\ &4. \quad X^T(Y - X\hat{\beta}) = \lambda \mathbf{s}. \end{aligned}$$

$\lambda = 0 \Longrightarrow \hat{\beta}$ is the least square estimate. $\lambda \neq 0 \Longrightarrow \|\hat{\beta}\|_1 = t$.

定理8. *The solution of the Lasso (2.6) always exists. If*

$$t < t_0 := min\{\|\beta\|_1 : X^T X\beta = X^T Y\},$$

then for any solution $\hat{\beta}$, we have $\|\hat{\beta}\|_1 = t$. If $t \ge t_0$, then some OLS estimate is the solution of the Lasso.

Proof. Because $t < t_0$, from 1 and 4, we have $\lambda \neq 0$. If $t \ge t_0$, some OLS estimate, especially

$$\hat{\beta} := \arg\min_{\beta}\{\|\beta\|_1 : X^T X\beta = X^T Y\}$$

together with $\lambda = 0$ satisfy all of these FOUR conditions. □

定理9 (Uniqueness)**.** *Suppose that β^* and β^+ are both the solution of the Lasso (2.5). Let S be the support of one solution β^*:*

$$S = \{j : \beta_j^* \neq 0\}$$

and

$$S^c = \{j : \beta_j^* = 0\}.$$

If

$$X_S^T X_S \text{ is invertible, and in KKT condition 4 } s_j < 1, \text{ for } j \in S^c$$

then the Lasso (2.5) has a unique solution:

$$\beta^+ = \beta^*.$$

Proof. Let $f_0(\beta) = \frac{1}{2}\|Y - X\beta\|_2^2$. Since both β^* and β^+ are the solution of the Lasso (2.5), we

have

$$f_0(\beta^*) + \lambda\|\beta^*\|_1 = f_0(\beta^+) + \lambda\|\beta^+\|_1.$$

By the definition of $\mathbf{s}$ which satisfy KKT condition 4, we have

$$\mathbf{s}^T\beta^* = \|\beta^*\|_1.$$

So,

$$f_0(\beta^*) + \lambda\ \mathbf{s}^T\beta^* = f_0(\beta^+) + \lambda\|\beta^+\|_1.$$

By subtracting $\lambda\ \mathbf{s}^T\beta^+$ from both sides, we have

$$f_0(\beta^*) + \lambda\ \mathbf{s}^T(\beta^* - \beta^+) = f_0(\beta^+) + \lambda(\|\beta^+\|_1 - \mathbf{s}^T\beta^+).$$

Note KKT condition 4 says that

$$\triangledown f_0(\beta^*) = -\lambda\mathbf{s}.$$

So we have

$$f_0(\beta^*) + \triangledown f_0(\beta^*)^T(\beta^+ - \beta^*) - f_0(\beta^+) = \lambda(\|\beta^+\|_1 - \mathbf{s}^T\beta^+).$$

By convexity of $f_0(\beta)$, we have

$$f_0(\beta^+) \geq f_0(\beta^*) + \triangledown f_0(\beta^*)^T(\beta^+ - \beta^*).$$

So the left hand side is less than 0, and hence

$$\|\beta^+\|_1 \leq \mathbf{s}^T\beta^+ \leq \|\beta^+\|.$$

So

$$\|\beta^+\|_1 = \mathbf{s}^T\beta^+,$$

from which we know that when $|s_j| < 1$, $\beta_j^+ = 0$.

Summarize, we have the following results:

$$\beta_j^+ = \beta_j^*, \text{ for } j \in S^c$$

and

$$f_0(\beta^*) + \triangledown f_0(\beta^*)^T(\beta^+ - \beta^*) - f_0(\beta^+) = 0.$$

Now let

$$\hat{\beta}_S := \|Y - X_S\beta_S\|_2^2 + \lambda\|\beta_S\|_1,$$

then $\beta^* = [\hat{\beta}_S, 0]$ and $\beta^+ = [\beta_S, 0]$. If $\hat{\beta}_S$ is unique, then $\beta^* = \beta^+$. In fact, $\hat{\beta}_S$ is unique, because the above problem is a strictly convex problem.

□

Some facts: 1. Strictly convex:

$$f(\alpha x_1 + (1-\alpha)x_2) < \alpha f(x_1) + (1-\alpha)f(x_2)$$

2. For twice continuously differentiable function $f(x)$, if

$$\nabla^2 f(x) \succ 0,$$

then $f(x)$ is strictly convex.

3. If $f(x)$ is strictly convex and $g(x)$ is convex, then $f(x) + g(x)$ is strictly convex.

4. For a strictly convex function, there is at most one minimization point.

定理10 (作业)**.** *The solution of the Lasso (2.5) always exists. The solution β^* satisfy the following condition: (1) $X\beta^*$ is a constant; (2) $\|Y - X\beta\|_2^2$ is a constant. (3) $\lambda\|\beta\|_1$ is a constant.*

Proof. If $\lambda > 0$, (2.5) $\Longleftrightarrow$

$$\begin{aligned} min \quad & \tfrac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \\ s.t. \quad & \|\beta\|_1 \le \tfrac{1}{2}\|Y\|_2^2/\lambda. \end{aligned} \tag{2.7}$$

A continuous function on a closed set has a minimization point. Suppose β^* and β^+ are both the solutions. By the proof in Theorem 9, we have

$$f_0(\beta^*) + \nabla f_0(\beta^*)^T(\beta^+ - \beta^*) - f_0(\beta^+) = 0.$$

By symmetry of β^* and β^+, we also have

$$f_0(\beta^+) + \nabla f_0(\beta^+)^T(\beta^* - \beta^+) - f_0(\beta^*) = 0.$$

So we have

$$\nabla f_0(\beta^*)^T(\beta^+ - \beta^*) + \nabla f_0(\beta^+)^T(\beta^* - \beta^+) = 0.$$

$$[\nabla f_0(\beta^+) - \nabla f_0(\beta^*)]^T(\beta^* - \beta^+) = 0$$

$$[-X^T(Y - X\beta^*) + X^T(Y - X\beta^+)]^T(\beta^* - \beta^+) = 0$$

$$(\beta^* - \beta^+)X^TX(\beta^* - \beta^+) = 0$$

So we have

$$X\beta^* = X\beta^+.$$

Since

$$\|Y - X\beta^*\| + \lambda\|\beta^*\| = \|Y - X\beta^+\| + \lambda\|\beta^+\|,$$

we have

$$\|\beta^*\|_1 = \|\beta^+\|_1.$$

□

定理11. *The solution of the Lasso (2.5) always exists. If*

$$\lambda > 0,$$

then for any solution $\hat{\beta}$, we have $\|\hat{\beta}\|_1$ is a constant t^*. (2.5) has the same solution with the following problem:

$$\begin{aligned} &minimize \quad \tfrac{1}{2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{n} x_{ij}\beta_j\right)^2 \\ &subject \ \ to \qquad \sum_{j=1}^{p}|\beta_j| \le t^*, \end{aligned} \tag{2.8}$$

Proof. Suppose β^* is an arbitrary solution. From Theorem 10 we have $\sum_{j=1}^{p}|\beta_j^*|$ is a constant. let $t^* = \sum_{j=1}^{p}|\beta_j^*|$. Then β^* and λ satisfy KKT conditions.

(1) Any solution of (2.5) is the solution of (2.7). This is obvious. Because any solution of (2.5) together with λ and some $\mathbf{s}$ satisfy KKT conditions.

(2) Any solution of (2.7) is a solution of (2.5). We only need to prove that for any β^+ which is one solution of (2.7), it satisfies

$$\|Y - X\beta^+\|_2^2 + \lambda\|\beta^+\|_1 = \|Y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1 = p^*.$$

First $\|Y - X\beta^+\|_2^2 = \|Y - X\beta^*\|_2^2$ because both β^+ and β^* are solutions of (2.7). Now we prove

$$\|\beta^+\|_1 = \|\beta^*\|_1.$$

Suppose not, then

$$\|Y - X\beta^+\|_2^2 + \lambda\|\beta^+\|_1 < \|Y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1,$$

which contradicts our assumption that β^* is one minimizer of (2.5). □

2.4 Interior-point methods

The basic idea of Interior-point methods is to transform the inequality constrained optimization problem to equality constrained optimization problem. Then Newton method is applied.

2.4.1 Newton method

Second order approximation of $f(x)$:

$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v,$$

which achieves its minimum at $v = -\nabla^2 f(x)^{-1} \nabla f(x)$. The quantity

$$\Delta x_{nt} \equiv -\nabla^2 f(x)^{-1} \nabla f(x)$$

is called the Newton step (for f at x).

$\lambda(x) \equiv \left[2[\hat{f}(x) - \hat{f}(x + \Delta x_{nt})]\right]^{1/2} = \left[\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right]^{1/2} = \left[\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt}\right]^{1/2}$ is called the Newton decrement at x.

Algorithm 1 Newton method

Input: a start point x, tolerance $\epsilon > 0$.

1: **repeat**

2: compute the Newton step and decrement:

$$\Delta x_{nt} = -\nabla^2 f(x)^{-1} \nabla f(x);\ \lambda^2 = \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).$$

3: Update. $x = x + \Delta x_{nt}$

4: **until** $\lambda^2/2 \leq \epsilon$

2.4.2 Newton method for equality constrained problem

A convex optimization problem with equality constraints:

$$\begin{aligned} &minimize && f(x) \\ &subject\ \ to && Ax = b \end{aligned}$$

Suppose this problem has some good properties, say the problem achieves its minimum at some point x^* and the Slater's conditions hold. That is, there exists some $x \in \mathbf{int}(\mathbf{dom} f)$ such that $Ax = b$.

By KKT conditions, a point x^* is optimal if and only if there is a ν^*, such that

$$Ax^* = b, \quad \nabla f(x^*) + A^T\nu^* = 0.$$

The Newton method with equality constraints is almost the same as Newton's method without constraints, except for two differences: 1. the initial point must be feasible; 2. we make sure that the Newton step Δx_{nt} is a feasible direction: $A\Delta x_{nt} = 0$.

We derive the Newton step Δx_{nt} by second-order Taylor approximation near x:

$$\begin{aligned} &minimize\ (over\ z) && \hat{f}(x+v) := f(x) + \nabla f(x)^T v + \tfrac{1}{2} v^T \nabla^2 f(x) v \\ &subject\ \ to && A(x+v) = b \end{aligned}$$

By KTT conditions, we have

$$A\Delta x_{nt} = 0, \quad \nabla^2 f(x)\Delta x_{nt} + \nabla f(x) + A^T\nu^* = 0.$$

Write them in a matrix form,

$$\begin{bmatrix} \nabla^2 f(x) & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} \Delta x_{nt} \\ \nu^* \end{bmatrix} = \begin{bmatrix} -\nabla f(x) \\ 0 \end{bmatrix}.$$

The Newton decrement at x: $\lambda(x) \equiv \left[2[\hat{f}(x) - \hat{f}(x + \Delta x_{nt})]\right]^{1/2} = \left[\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{nt}\right]^{1/2}$.

Algorithm 2 Newton method

Input: a start point x, tolerance $\epsilon > 0$.

1: **repeat**
2: compute the Newton step and decrement: Δx_{nt} and λ^2
3: **until** $\lambda^2/2 \leq \epsilon$

2.4.3 The barrier method

Once we can transform inequality constrained minimization problems to equality constrained problem, then Newton methods described in previous subsection can be applied.

Note that

$$\begin{aligned} &minimize && f_0(x) \\ &subject\ \ to && f_i(x) \leq 0, i = 1, \ldots, n, \\ & && Ax = b. \end{aligned}$$

$\Longleftrightarrow$

$$\begin{aligned} &minimize && f_0(x) + \textstyle\sum_{i=1}^{n} I_-(f_i(x)) \\ &subject\ \ to && Ax = b, \end{aligned}$$

where $I_- : \mathbb{R} \to \mathbb{R}$ is the indicator function for the nonpositive reals:

$$I_-(u) = \begin{cases} 0 & u \leq 0 \\ \infty & u > 0. \end{cases}$$

By replacing the indicator function $I_-(u)$ with an approximated differentiable function:

$$\hat{I}_-(u) = -(1/t)log(-u),$$

we transform the original inequality constrained problem to a tractable equality constrained convex problem:

$$\begin{aligned} &minimize && f_0(x) + \textstyle\sum_{i=1}^{n} -(1/t)\log(-f_i(x)) \\ &subject\ \ to && Ax = b. \end{aligned} \tag{2.9}$$

$x^*(t)$ is called central path. The function

$$\phi(x) = -\sum_{i=1}^{n} \log(-f_i(x))$$

is called the log barrier for the original problem.

The gradient and Hessian of ϕ:

$$\begin{aligned}\nabla\phi(x) &= \sum_{i=1}^{n} -\frac{1}{f_i(x)}\nabla f_i(x) \\ \nabla^2\phi(x) &= \sum_{i=1}^{n} -\frac{1}{f_i(x)}\nabla^2 f_i(x) + \sum_{i=1}^{n} \frac{1}{f_i^2(x)}\nabla f_i(x)\nabla f_i(x)^T\end{aligned}$$

定理12. *Suppose the problem (2.11) has the following properties: 1. (2.11) can be solved via Newton's method 2. it has a unique solution for each $t > 0$.*

Suppose the original problem has the following properties: 1. it achieves its minimum at some point x^* and the minimum value $p^* > -\infty$. 2. Slater's conditions hold. Then

$$0 \le f_0(x^*(t)) - p^* \le n/t$$

Proof. Let $x^*(t)$ be the central path – the solution of problem (2.11), then it satisfies with KKT conditions:

$$Ax^*(t) = b, f_i(x^*(t)) < 0, i = 1, \ldots, m$$

and there exists a $\hat{\nu} \in \mathbb{R}^p$ such that

$$\begin{aligned}0 &= \nabla f_0(x^*(t)) + (\frac{1}{t})\nabla\phi(x^*(t)) + A^T\hat{\nu} \\ &= \nabla f_0(x^*(t)) + \frac{1}{t}\sum_{i=1}^{n} \frac{1}{-f_i(x^*(t))}\nabla f_i(x^*(t)) + A^T\hat{\nu}\end{aligned}$$

From the above equation, we see that: Every central point yields a dual feasible point, and hence a lower bound on the optimal value p^*. Define

$$\lambda_i^*(t) = -\frac{1}{-tf_i(x^*(t))}, i = 1, \ldots, m, \nu^*(t) = \nu.$$

We claim that the pair $\lambda^*(t), \nu^*(t)$ is dual feasible. It is because $\lambda^*(t) \succ 0$. Since

$$\nabla f_0(x^*(t)) + \sum_{i=1}^{n} \lambda^*(t)\nabla f_i(x^*(t)) + A^T\hat{\nu} = 0,$$

we see that $x^*(t)$ minimizes the Lagrangian:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{n} \lambda_i f_i(x) + \nu^T(Ax - b)$$

for $\lambda = \lambda^*(t)$ and $\nu = \nu^*(t)$, which means that $\lambda^*(t), \nu^*(t)$ is a dual feasible pair. The dual function is

$$g(\lambda^*, \nu^*) = f_0(x^*(t)) + \sum_{i=1}^{n} \lambda_i^*(t) f_i(x^*(t)) + \nu^*(t)^T (Ax^*(t) - b) = f_0(x^*(t)) - \frac{n}{t}$$

By weak duality, we have

$$f_0(x^*(t)) - \frac{n}{t} \leq p^* \Longrightarrow 0 \leq f_0(x^*(t)) - p^* \leq \frac{n}{t}$$

This confirms the intuition that $x^*(t)$ converges to an optimal point as $t \to \infty$. □

Algorithm 3 Barrier method

Input: strictly feasible point x, $t := t^{(0)} > 0, \mu > 1$, tolerance $\epsilon > 0$.

1: **repeat**
2: Centering step. Compute $x^*(t)$ by minimizing $tf_0 + \phi$, subject to $Ax = b$, starting at x.
3: Update. $x := x^*(t)$.
4: Increase t. $t := \mu t$.
5: **until** $n/t \leq \epsilon$

2.4.4 ℓ_1-norm approximation

Consider the ℓ_1-norm approximation problem

$$minimize \ \|Ax - b\|_1.$$

$\Longleftrightarrow$

$$\begin{array}{ll} minimize & \sum_{i=1}^{n} y_i \\ subject \ \ to & |Ax - b| \preceq y \end{array}$$

$\Longleftrightarrow$

$$\begin{array}{ll} minimize & \sum_{i=1}^{n} y_i \\ subject \ \ to & \begin{pmatrix} A & -I \\ -A & -I \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \preceq \begin{pmatrix} b \\ -b \end{pmatrix} \end{array}$$

This is a linear optimization problem. The centering step:

$$minimize \;\; t\mathbf{1}^T y + \phi.$$

Newton equation:

$$H \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} + g = 0,$$

where H is the Hessian, g the gradient of $t\mathbf{1}^T y + \phi$.

$$\begin{aligned} H &= \nabla^2\phi = \sum_{i=1}^{n} -\frac{1}{f_i(x)}\nabla^2 f_i(x) + \sum_{i=1}^{n} \frac{1}{f_i^2(x)}\nabla f_i(x)\nabla f_i(x)^T \\ &= \sum_{i=1}^{n} \frac{1}{f_i^2(x)}\nabla f_i(x)\nabla f_i(x)^T \\ &= \nabla f D \nabla f^T (in\ matrix\ form) \\ &= \begin{pmatrix} A^T & -A^T \\ -I & -I \end{pmatrix} \begin{pmatrix} D_1 & \\ & D_2 \end{pmatrix} \begin{pmatrix} A & -I \\ -A & -I \end{pmatrix}, \end{aligned}$$

where $D_1 = diag(\frac{1}{[Ax-y-b]_i^2})$,$D_2 = diag(\frac{1}{[-Ax-y+b]_i^2})$ and $\nabla f = [\nabla f_1, \ldots, \nabla f_n]_{p\times n}$.

$$\begin{aligned} g &= \nabla\phi + \begin{pmatrix} 0 \\ t\mathbf{1} \end{pmatrix} \\ &= \sum_{i=1}^{n} -\frac{1}{f_i(x)}\nabla f_i(x) + \begin{pmatrix} 0 \\ t\mathbf{1} \end{pmatrix} \\ &= \nabla f \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} + \begin{pmatrix} 0 \\ t\mathbf{1} \end{pmatrix} \\ &= \begin{pmatrix} A^T & -A^T \\ -I & -I \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} + \begin{pmatrix} 0 \\ t\mathbf{1} \end{pmatrix} \\ &= \begin{pmatrix} A^T(g_1 - g_2) \\ -g_1 - g_2 + t\mathbf{1} \end{pmatrix}, \end{aligned}$$

where g_1, g_2 are two column vector: $[g_1]_i = [-\frac{1}{Ax-y-b}]_i$ and $[g_2]_i = [-\frac{1}{-Ax-y+b}]_i$.

Finally, we have the Newton equation:

$$\begin{pmatrix} A^T & -A^T \\ -I & -I \end{pmatrix} \begin{pmatrix} D_1 & \\ & D_2 \end{pmatrix} \begin{pmatrix} A & -I \\ -A & -I \end{pmatrix} \begin{pmatrix} \Delta x_{nt} \\ \Delta y_{nt} \end{pmatrix} = - \begin{pmatrix} A^T(g_1 - g_2) \\ -g_1 - g_2 + t\mathbf{1} \end{pmatrix},$$

From which, under some simple regularity conditions, we have (homework)

$$A^T \tilde{D} A \Delta x_n t = -A^T \tilde{g}$$

and

$$\Delta y_{nt} = (D1 + D2)^{-1} \left[-\tilde{g}_2 + (D_1 - D_2) A \Delta x_{nt} \right],$$

where

$$\tilde{g}_1 = g1 - g2; \ \tilde{g}_2 = -g1 - g2 + t\mathbf{1}$$

$$\tilde{D} = 2 \left[diag(y)^2 + diag(b - Ax)^2 \right]^{-1}$$

$$\tilde{g} = \tilde{g}_1 + (D1 - D2)(D1 + D2)^{-1} \tilde{g}_2$$

.

2.5 Reading materials

An Interior-Point Method for Large-Scale ℓ_1 Regularized Least Squares, written by Kim, Koh, Lustig, Boyd and Gorinevsky.

An Interior-Point Method for Large-Scale ℓ_1 Regularized logistic regression, written by Kim, Koh, and Boyd.

2.6 Homework

1. Convex problem 的局部最优解就是全局最优解.

2. The solution of the Lasso (2.5) with $\lambda > 0$ always exists. The solution β^* satisfy the following condition: (1) $X\beta^*$ is a constant; (2) $\|Y - X\beta\|_2^2$ is a constant. (3) $\|\beta\|_1$ is a constant.

3. $D_1 = diag(\frac{1}{[Ax-y-b]_i^2})$,$D_2 = diag(\frac{1}{[-Ax-y+b]_i^2})$

$$\begin{pmatrix} A^T & -A^T \\ -I & -I \end{pmatrix} \begin{pmatrix} D_1 & \\ & D_2 \end{pmatrix} \begin{pmatrix} A & -I \\ -A & -I \end{pmatrix} \begin{pmatrix} \Delta x_{nt} \\ \Delta y_{nt} \end{pmatrix} = - \begin{pmatrix} A^T(g_1 - g_2) \\ -g_1 - g_2 + t\mathbf{1} \end{pmatrix},$$

From which, under some simple regularity conditions, we have (homework)

$$A^T \tilde{D} A \Delta x_n t = -A^T \tilde{g}$$

and

$$\Delta y_{nt} = (D1 + D2)^{-1} \left[-\tilde{g}_2 + (D_1 - D_2) A \Delta x_{nt} \right],$$

where

$$\tilde{g}_1 = g1 - g2; \ \tilde{g}_2 = -g1 - g2 + t\mathbf{1}$$

$$\tilde{D} = 2 \left[diag(y)^2 + diag(b - Ax)^2 \right]^{-1}$$

$$\tilde{g} = \tilde{g}_1 + (D1 - D2)(D1 + D2)^{-1} \tilde{g}_2$$

.

4.

$$\partial(f_1 + f_2) = \partial f_1 + \partial f_2 \text{ (RHS is addition of sets)}$$

5. Definition of Strictly convex:

$$f(\alpha x_1 + (1 - \alpha) x_2) < \alpha f(x_1) + (1 - \alpha) f(x_2)$$

(1) For twice continuously differentiable function $f(x)$, if

$$\nabla^2 f(x) \succ 0,$$

then $f(x)$ is strictly convex.

(2) If $f(x)$ is strictly convex and $g(x)$ is convex, then $f(x) + g(x)$ is strictly convex.

(3) For a strictly convex function, there is at most one minimization point.

2.7 Small project

ℓ_1 regularized robust regression.

References

Stephen Boyd and Lieven Vandenberghe. (2009) Convex optimization. Cambridge University Press.

www.stanford.edu/class/ee392o/subgrad.pdf

Michael R. Osborne; Brett Presnell; Berwin A. Turlach. (2000) On the LASSO and Its Dual, Journal of Computational and Graphical Statistics, Vol. 9, No. 2. (Jun., 2000), pp. 319-337.

Chapter 3

Applied problems

3.1 R package for the Lasso

1. The installation of LARS

   install.packages('lars')

2. The usage of LARS

   library(lars)

   ?lars

   A "lars" object is returned, for which print, plot, predict, coef and summary methods exist.

3.2 Simulations

$n = 1000, p = 3$ X_1, X_2, e *i.i.d.* $\sim N(0,1)$ $X_3 = \frac{2}{3}X_1 + \frac{2}{3}X_2 + \frac{1}{3}e_i$ $Y = X_1\beta_1 + X_2\beta_2 + \epsilon$ (a) $\beta_1 = 2, \beta_2 = 3$; (b)$\beta_1 = -2, \beta_2 = 3$.

$$C_{21}C_{11}^{-1} = (\frac{2}{3}, \frac{2}{3}).$$

(1) the differences between the two solution path:

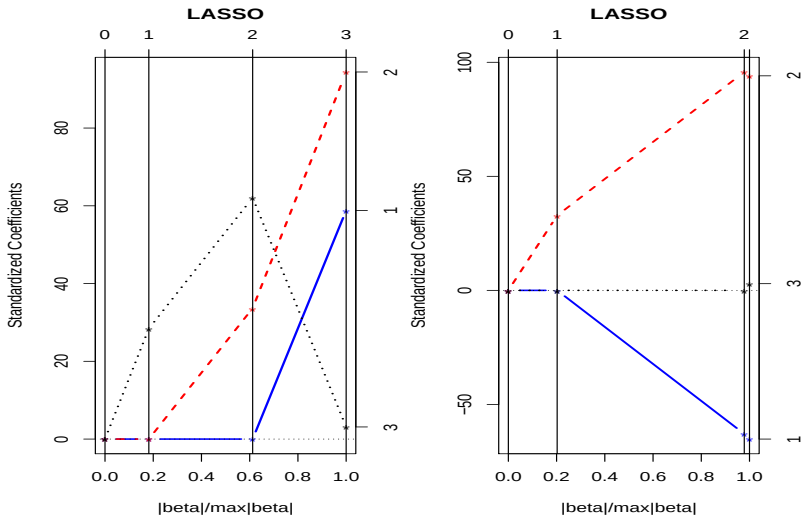

Figure 3.1: Two different Lasso solution paths. Left: $\beta = (2, 3, 0)$. Right: $\beta = (-2, 3, 0)$.

The key difference between the two settings is that the first setting ($\beta = (2, 3, 0)$) does not satisfy the **Irrepresentable Condition**, while the other one satisfy that condition.

定义3 (Irrepresentable Condition)**.**

$$\max \left| X_{S^c} X_S (X_S^T X_S)^{-1} sign(\beta) \right| < 1.$$

3.3 Applications–Microarray Classification

Data Description

Training set: 144 patients with 14 different types of cancer Test set: 54 patients. Gene expression measurements: 16,063 genes.

$$n = 144, \quad p = 16,063$$

Data visualization

Model Fitting

What model?

1. Regression.

$$\min \ \|Y - X\beta\|_2^2$$

Assumption:

1. $Y = X\beta + \epsilon$
2. $E(\epsilon) = 0$

In practice, it works quite well for binary Y.

2. Binary Logistic regression. This is one special case of GLM - Generalized Linear Models.

$$Y_i \in \{0, 1\}.$$

$$\text{logit } P(Y_i = 1|x_i) = x_i^T\beta$$

$$\min \ \sum_{i=1}^{n} \left[\log\left(1 + e^{x_i^T\beta}\right) - y_i x_i^T\beta\right]$$

3. SVM

For separated case, the optimization problem is

$$\begin{aligned} &\max_{\beta,\beta_0,\|\beta\|_2=1} M \\ &s.t. \ \ y_i(x_i^T\beta + \beta_0) \geq M, i = 1, \ldots, n. \end{aligned} \tag{3.1}$$

which is equivalent to

$$\begin{aligned} &\min \|\beta\|_2 \\ &s.t. \ \ y_i(x_i^T\beta + \beta_0) \geq 1, i = 1, \ldots, n. \end{aligned} \tag{3.2}$$

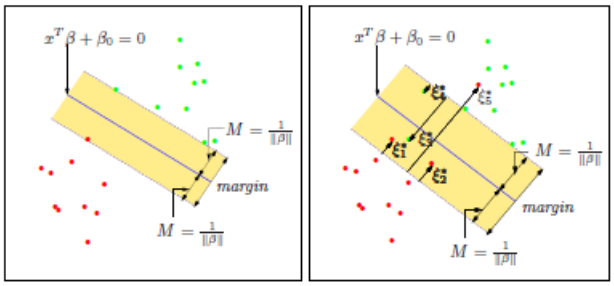

FIGURE 12.1. *Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width $2M = 2/\|\beta\|$. The right panel shows the nonseparable (overlap) case. The points labeled ξ_j^* are on the wrong side of their margin by an amount $\xi_j^* = M\xi_j$; points on the correct side have $\xi_j^* = 0$. The margin is maximized subject to a total budget $\sum \xi_i \leq$ constant. Hence $\sum \xi_j^*$ is the total distance of points on the wrong side of their margin.*

Figure 3.2: This figure is from "elements to statistical machine learning" P418.

Now consider the nonseperable case. The natura way to modify the constraint in (3.1) is by introducing the slack variable $\xi = (\xi_1, \ldots, \xi_n)$:

$$y_i(x_i^T\beta + \beta_0) \geq M(1 - \xi_i),$$

$\forall i, \xi_i \geq 0, \sum_i \xi_i \leq constant.$

Remark: $M\sum_i \xi_i$ measures the total amount distance of points on the wrong side of their margin.

An equivalent form of the no-separable SVM problem

$$\begin{aligned} &\min \|\beta\|_2^2 \\ s.t.\ \ &y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, i = 1, \ldots, n. \\ &\xi_i \geq 0, \sum_i \xi_i \leq constant \end{aligned} \tag{3.3}$$

Or equivalently,

$$\begin{aligned} &\min \tfrac{1}{2}\|\beta\|_2^2 + C\sum_i \xi_i \\ s.t.\ \ &y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, i = 1, \ldots, n. \\ &\xi_i \geq 0 \end{aligned} \tag{3.4}$$

Equivalently,

$$\min \sum_{i=1}^{n} [1 - y_i(x_i^T\beta + \beta_0)]_+ + \frac{\lambda}{2}\|\beta\|_2^2, \tag{3.5}$$

where x_+ indicates the positive part of x.

If $\lambda = C/2$, then (3.5) and (3.4) are equivalent. (Homework.)

4. k-NN This method requires no model to be fit. Given a query point x_0, we find the k training

points $x_{(r)}, r = 1, \ldots, k$ closest in distance to x_0,, and then classify using majority vote among the k neighbors.

5. Multi-category SVM. (Ref. Multicategory support vector machines. By Lee, Lin and Wahba. JASA 2004 67-81)

   Define $\mathbf{v}_j$ as a k-dim vector with the jth coordinate 1 and $-\frac{1}{k-1}$ elsewhere. If example i falls into class j, then we denote $\mathbf{y}_i = \mathbf{v}_j$. Now define k classifier $\mathbf{f}(X) = (f_1(X), \ldots, f_K(X))$ with sum-to-0 constraint

$$\sum_i f_i(X) = 0, \text{ for any } X \in \mathbb{R}^p.$$

   Let $\mathbf{L}(\cdot)$ be a function that maps a class label $\mathbf{y}_i$ to a 0-1 loss vector: if sample i falls into class j, then $\mathbf{L}(\mathbf{y_i})$ is a k-dim vector with 0 in the jth coordinate and 1 elsewhere. Then we can define the Multi-category SVM:

$$\min \frac{1}{n} \sum_{i=1}^{n} \mathbf{L}(\mathbf{y}_i)^T (\mathbf{f}(x_i) - \mathbf{y}_i)_+ + \frac{\lambda}{2} \sum_{i,j} \beta_{ij}^2.$$

   Classification rule:

$$\phi(x) = \arg\max_j f_j(x).$$

   When $K = 2$, the M-SVM reduces to the simple SVM. If $\mathbf{y}_i = (1, -1)$, then

$$\mathbf{L}(\mathbf{y}_i)^T (\mathbf{f}(x_i) - \mathbf{y}_i)_+ = (0, 1)^T (f_1(x_i) - 1)_+, f_2(x_i) + 1)_+) = (f_2(x_i) + 1)_+ = (1 - f_1(x_i))_+.$$

6. Multi-class logistic regression. It generalizes the binary response case to the multi-nominal response case.

$$P(Y_i = k|x_i) \propto \exp(x_i^T \beta_k + \beta_{k0})$$

   or,

$$P(Y_i = k|x_i) = \frac{\exp(x_i^T \beta_k + \beta_{k0})}{\sum_{j=1}^{K} \exp(x_i^T \beta_j + \beta_{j0})}$$

7. Sparse LDA.

   (1) LDA. Suppose that we model each class density as multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right].$$

Let π_k denote the prior probability of class k, then

$$\begin{aligned}\log\frac{P(G=k|X=x)}{P(G=l|X=x)} &= \log\frac{f_k(x)\pi_k}{f_l(x)\pi_l} \\ &= \left[\log(\pi_k)-\frac{1}{2}\log|\Sigma_k|-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right] \\ &\quad -\left[\log(\pi_l)-\frac{1}{2}\log|\Sigma_l|-\frac{1}{2}(x-\mu_l)^T\Sigma_l^{-1}(x-\mu_l)\right]\end{aligned}$$

Define

$$\delta_k(x)=\log(\pi_k)-\frac{1}{2}\log|\Sigma_k|-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k),$$

then

$$P(G=k|X=x)\geq P(G=l|X=x)\Longrightarrow\delta_k(x)\geq\delta_l(x).$$

So the decision rule can be described as:

$$G(x)=\arg\max_k\delta_k(x).$$

This is QDA. If $\Sigma_k=\Sigma$, for all k, then the QDA can be LDA with

$$\delta_k(x)=\log(\pi_k)+x^T\Sigma^{-1}\mu_k-\frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k.$$

In practice, parameters should be estimated from train data:

- $\hat{\pi}_k=N_k/N$
- $\hat{\mu}_k=\sum_{g_i=k}x_i/N_k$
- $\hat{\Sigma}=\sum_{k=1}^K\sum_{g_i=k}(x_i-\hat{\mu}_k)(x_i-\hat{\mu}_k)^T/(N-K)$

(2) FDA. Transform categorical response to continuous scores containing the information of predictors.

We use g_i to denote the class of sample i, then we solve

$$\arg\min\sum_i(\theta(g_i)-x_i^T\beta)^2.$$

To avoid a trivial solution, we use restrictions on θ - mean zero and unit variance.

For a two class separation problem, one direction is enough. But for a K-class separation problem, more directions are needed. Suppose we use $L\leq K-1$ directions, we can solve

the following problem:

$$\min \sum_l \sum_i (\theta_l(g_i) - x_i^T \beta_l)^2,$$

such that

$$\theta^T \theta = I_{L \times L}$$

Once the L orthogonal directions are obtained, we can use the new features to train the classifiers via LDA. It can be shown that this LDA is equivalent to the original LDA on the raw data.

The more powerful of this model is that

- introduce non-linear function
- introduce regularization (penalty)

8. Classification Tree.

$$f(X) = \sum_m c_m I\{X\} \in R_m$$

How to find the best partitions? Computationally infeasible. A greedy algorithm for

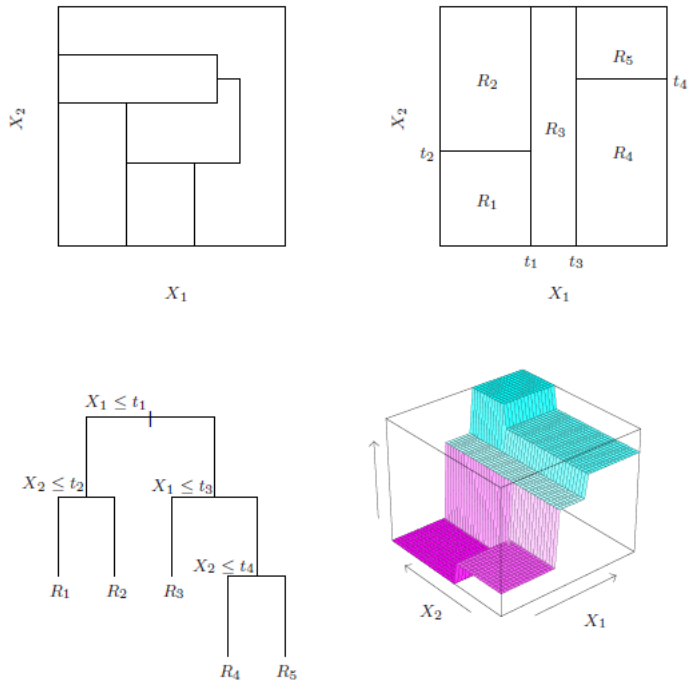

Figure 3.3: Partitions and CART

regression tree:

Consider a splitting variable j and split point s, define

$$R_1(j,s) = \{X|X_j \leq s\} \text{ and } R_2(j,s) = \{X|X_j > s\}.$$

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2 + \min_{c_2}\sum_{x_i\in R_1(j,s)}(y_i-c_2)^2\right]$$

Repeat the above procedure until some some criteria is reached.

For Classification tree, the only difference is that we won't use L2 loss function, instead we use 0-1 loss, that is

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}I\{y_i\neq c_i\} + \min_{c_2}\sum_{x_i\in R_1(j,s)}I\{y_i\neq c_2\}\right]$$

Selection of Tuning Parameter

(1) Cross validation

(2) Three splits

(3) AIC,BIC,Cp values

$$AIC = n\log(\|Y-\hat{Y}\|^2) + 2\ df$$

$$BIC = n\log(\|Y-\hat{Y}\|^2) + 2\log n\ df$$

$$Cp = \frac{\|Y-\hat{Y}\|^2}{\sigma^2} - n + 2\ df.$$

Results and Comparison

3.4 Homework

1. If $\lambda = C/2$, then (3.5) and (3.4) are equivalent.

2. Reproduce the simulations in Section 3.2 in "On model selection consistency of the Lasso" by Zhao and Yu. Report your findings. Hand in your report with your code.

3.5 Small Project

Working on handwritten digit recognition problem. Compare your method with popular methods you can think of.

Working on the microarray classification problem. Compare your method with popular methods you can think of.

Chapter 4

Concentration Inequalities

4.1 Introduction

Concentration inequalities deal with deviations of functions of independent random variables from their expectations.

Jenson Inequality: for convex f,

$$f(E(X)) \leq E[f(X)].$$

Proof. Since $f(x)$ is convex, we have

$$f(X) \geq f(E(X)) + a^T(X - E(X)),$$

for $a \in \partial f$ at $E(X)$. So we have

$$E[f(X)] \geq f(E(X)) + a^T E(X - E(X)) = f(E(X))$$

□

Applications of Jenson Inequality:

$$E(X^2) \geq [E(X)]^2$$

$$\frac{\sum_{i=1}^n x_i}{n} \geq (\Pi_{i=1}^n x_i)^{\frac{1}{n}}$$

KL-divergence is always non-negative: For two distributions P and Q of discrete random variables, their K – L divergence is defined as

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \geq 0 \text{ (homework)}$$

Markov inequality: If $X \geq 0$, then for all $t > 0$,

$$P[X \geq t] \leq \frac{E[X]}{t}.$$

Proof.

$$\begin{aligned} P[X \geq t] &= E[1_{\{X \geq t\}}] \\ &\leq E\left[\frac{X}{t} 1_{\{X \geq t\}}\right] \\ &\leq \frac{E[X 1_{\{X \geq t\}}]}{t} \\ &\leq \frac{E[X]}{t}. \end{aligned}$$

□

Chebyshev: for any $t > 0$,

$$P[|X - E[X]| \geq t] \leq \frac{Var(X)}{t^2}.$$

Chernoff: for all $t \in R$,

$$P[X \geq t] \leq \inf_{\lambda \geq 0} E[e^{\lambda(X-t)}].$$

Proof.

$$\begin{aligned} P[X \geq t] &= P(e^{\lambda X} \geq e^{\lambda t}) \\ &\leq \frac{E[e^{\lambda X}]}{e^{\lambda t}} \\ &= E[e^{\lambda(X-t)}] \end{aligned}$$

□

4.1.1 Asymptotic V.S. Non-asymptotic

Let $X_1, X_2, \ldots,$ be i.i.d. Bernoulli random variables: $P(X_i = 1) = P(X_i = -1) = 0.5$. Then by the Central Limit Theorem as $n \to \infty$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \to_{dist.} g,$$

where g is a $N(0, 1)$ standard normal random variable. Equivalently,

$$P\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i > t\right] \to P(g > t) = \frac{1}{\sqrt{2\pi}} \int_t^{\infty} e^{-x^2/2} dx.$$

Hoeffding's tail inequality tells us that, for any n and $t > 0$,

$$P\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i > t\right] \leq e^{-t^2/2}.$$

To prove Hoeffding's tail inequality, we can use Hoeffding's inequality:

引理2 (Heoffding's Inequality.)**.** *Let X be a random variable with $EX = 0$, $a \leq X \leq b$. Then for t*

$$E[e^{tX}] \leq e^{t^2(b-a)^2/8}.$$

Proof of Hoeffding's tail inequality:

Proof. By Chernoff Inequality,

$$\begin{aligned} P\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i > t\right] &\leq E(e^{\lambda(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i - t)}) \\ &= e^{-\lambda t}\Pi_{i=1}^{n} E e^{\frac{\lambda}{\sqrt{n}}X_i} \\ &\leq e^{-\lambda t}\Pi_{i=1}^{n} e^{[\frac{\lambda}{\sqrt{n}}]^2 2^2/8} (by\ Hoeffding's\ Inequality) \\ &= e^{\frac{1}{2}\lambda^2 - \lambda t} \end{aligned}$$

by taking $\lambda = t$, we have

$$P\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i > t\right] \leq e^{-t^2/2}.$$

□

Proof of Hoeffding's Inequality:

Proof. We only have to prove $t > 0$ case. For $t < 0$ case, we can write $tX = -t \times (-X)$. Since e^{tx} is convex, we have for any $0 \leq \alpha \leq 1$,

$$e^{\alpha ta + (1-\alpha)tb} \leq \alpha e^{ta} + (1-\alpha)e^{tb}.$$

By taking $\alpha = \frac{b-x}{b-a}$, we have

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

By taking expectation on both sides, we have

$$E(e^{tX}) \leq \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}$$

Finally it suffices to prove

$$\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} \leq e^{t^2(b-a)^2/8},$$

Let $f(t) = \log\left[\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb}\right]$, then we have

$$f(0) = 0$$

$$f'(t) = \frac{ab(e^{ta} - e^{tb})}{be^{ta} - ae^{tb}}$$

$$\begin{aligned} f''(t) &= \frac{ab(ae^{ta} - be^{tb})(be^{ta} - ae^{tb}) - [ab(e^{ta} - e^{tb})]^2}{[be^{ta} - ae^{tb}]^2} \\ &= \frac{ab[-b^2e^{t(a+b)} - a^2e^{t(a+b)} + 2abe^{t(a+b)}]}{[be^{ta} - ae^{tb}]^2} \\ &= \frac{-abe^{t(a+b)}(a-b)^2}{[be^{ta} - ae^{tb}]^2} \\ &\leq \frac{-abe^{t(a+b)}(a-b)^2}{-4abe^{t(a+b)}} \\ &= \frac{(a-b)^2}{4}, \end{aligned}$$

we used the condition that $a < 0$ and $b > 0$. By Taylor expansion,

$$\begin{aligned} f(t) &= f(0) + f'(0)t + \frac{1}{2}f''(\xi)t^2, \text{ for some } \xi \in [0,t] \\ &\leq \frac{t^2(b-a)^2}{8}, \end{aligned}$$

Equivalently,

$$\frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} \leq e^{t^2(b-a)^2/8}.$$

□

4.2 Bounded random variables

定理13 (Hoeffding)**.** *Let $X_1, \ldots, X_n$ be independent bounded random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one. Then for any $t > 0$ we have*

$$P\left[\sum X_i - E\left(\sum X_i\right) \geq t\right] \leq e^{-2t^2/\sum(b_i-a_i)^2}.$$

A disadvantage of this inequality is that it ignores information about the variance of the $X_i's$.

定理14 (Bennett's inequality.)**.** *Let $X_1, \ldots, X_n$ be independent real-valued random variables with zero mean and assum that $X_i \leq 1$. Let*

$$\sigma^2 = \frac{1}{n}\sum_i var(X_i).$$

Then for any $t > 0$,

$$P\left[\sum X_i > t\right] \leq \exp\left\{-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right\},$$

where $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.

Proof.

$$P\left[\sum X_i > t\right] \leq e^{-st}\Pi_i E(e^{sX_i})$$

Let

$$\psi(x) = \exp(x) - x - 1.$$

We see that

$$\psi(x) \leq \frac{x^2}{2}, \text{ for } x \leq 0.$$

$$\phi(sx) \leq x^2\psi(s), \text{ for } s \geq 0 \text{ and } x \in [0,1].$$

So,

$$\begin{aligned}
E[e^{sX_i}] &= 1 + sE(X_i) + E(\psi(sX_i)) \\
&= 1 + E(\psi(sX_i)) \\
&= 1 + E(\psi(sX_i)_+) + E(\psi(-(sX_i)_-) \\
&\leq 1 + E(\psi((sX_i)_+) + \frac{s^2}{2}E\left\{[(X_i)_-]^2\right\} \\
&\leq 1 + \psi(s)E\left\{[(X_i)_+]^2\right\} + \frac{s^2}{2}E\left\{[(X_i)_-]^2\right\} \\
&\leq 1 + \psi(s)E\left\{[(X_i)_+]^2\right\} + \psi(s)E\left\{[(X_i)_-]^2\right\} \\
&= 1 + \psi(s)E(X_i^2) \\
&\leq e^{\psi(s)\sigma_i^2}
\end{aligned}$$

So we have

$$\begin{aligned}
P\left[\sum X_i > t\right] &\leq e^{-st}\Pi_i E(e^{sX_i}) \\
&\leq \exp\left\{n\sigma^2\psi(s) - st\right\}
\end{aligned}$$

Note that

$$n\sigma^2\psi(s) - st = n\sigma^2(e^s - s - 1) - st$$

is minimized at

$$s = \log(1 + \frac{t}{n\sigma^2}).$$

Substituting this value in the upper bound, we obtain Bennett's inequality. □

定理15 (Bernstein's inequality.)**.** *Let $X_1, \ldots, X_n$ be independent real-valued random variables with zero mean and assume that $X_i \leq 1$. Let*

$$\sigma^2 = \frac{1}{n}\sum_i var(X_i).$$

Then for any $\epsilon > 0$,

$$P\left[\frac{1}{n}\sum X_i > \epsilon\right] \leq \exp\left\{\frac{-n\epsilon^2}{2(\sigma^2 + \epsilon/3)}\right\}.$$

Proof. This is because

$$h(u) \geq \frac{u^2}{2 + 2u/3}.$$

□

4.3 Gaussian Random Variables

引理3 (Bounds of Gaussian CDF)**.** *X is $N(0,1)$, for any $t > 0$,*

$$\frac{1}{\sqrt{2\pi}}e^{-t^2/2}\left(\frac{1}{t} - \frac{1}{t^3}\right) \leq P\left[X \geq t\right] \leq \frac{e^{-t^2/2}}{\sqrt{2\pi}t}.$$

Proof.

$$\begin{aligned}
P[X \geq t] &= \int_t^\infty \frac{1}{\sqrt{2\pi}}(\frac{1}{x})xe^{-\frac{x^2}{2}}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_t^\infty -\frac{1}{x}de^{-\frac{x^2}{2}} \\
&= \frac{1}{\sqrt{2\pi}}\left[\frac{e^{\frac{-t^2}{2}}}{t} - \int_t^\infty e^{-\frac{x^2}{2}}\frac{1}{x^2}dx\right] \quad \left(\leq \frac{1}{\sqrt{2\pi}}\frac{e^{\frac{-t^2}{2}}}{t}\right) \\
&= \frac{1}{\sqrt{2\pi}}\left[\frac{e^{\frac{-t^2}{2}}}{t} + \int_t^\infty \frac{1}{x^3}de^{-\frac{x^2}{2}}\right] \\
&= \frac{1}{\sqrt{2\pi}}\left[e^{\frac{-t^2}{2}}\left(\frac{1}{t} - \frac{1}{t^3}\right) + \int_t^\infty \frac{3}{x^4}e^{-\frac{x^2}{2}}dx\right] \quad \left(\geq \frac{1}{\sqrt{2\pi}}\left[e^{\frac{-t^2}{2}}\left(\frac{1}{t} - \frac{1}{t^3}\right)\right]\right)
\end{aligned}$$

□

定理16. *$X \sim N(0, \sigma^2)$, we have*

$$P(|X| \geq t) \leq e^{-t^2/(2\sigma^2)}$$

Proof. We only need to prove

$$P(|X| \geq t) \leq e^{-t^2/2},$$

for a standard normal r.v. X.

$$\begin{aligned} P(|X| > t) &= 2P(X > t) \\ &\leq \frac{2e^{-t^2/2}}{\sqrt{2\pi}t} \\ &\leq e^{-t^2/2} \quad (\text{if } t \geq \sqrt{\tfrac{2}{\pi}}) \end{aligned}$$

For $0 \leq t < \sqrt{\frac{2}{\pi}}$, let $f(x) \equiv 2P(X \geq t) - e^{-t^2/2}$. we have

$$f(0) = 0,$$

$$f'(x) = -\frac{2}{\sqrt{2\pi}}e^{-x^2/2} + xe^{-x^2/2} = (x - \sqrt{\frac{2}{\pi}})e^{-x^2/2} < 0,$$

so we have $f(x) \leq f(0) = 0$, for $0 \leq t < \sqrt{\frac{2}{\pi}}$.

□

定理17 (Gaussian concentration inequality for Lipschitz functions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function which is Lipschitz with constant 1 (i.e., $|f(x) - f(y)| \leq ||x - y||_2$ for all $x, y \in \mathbb{R}^d$). Then for any t, we have*

$$P\left[f(X) - E(f(X)) \geq t\right] \leq \exp(-ct^2)$$

Proof. We only need to prove

$$P(f(X) \geq t) \leq \exp(-ct^2),$$

for $E(f(X)) = 0$. In fact,

$$P(f(X) \geq t) \leq E[\exp(s(f(X) - t))].$$

Let Y be an independent copy of X. By Jenson's inequality, we have

$$E[\exp(-sf(Y))] \geq \exp(-sEf(Y)) = 1,$$

So

$$\begin{aligned} E[\exp(sf(X))] &\leq E[\exp(sf(X))\exp(-sEf(Y))] \\ &= E[\exp(s[f(X) - f(Y)])] \\ &\quad \{ \text{ if X is one-dimensional variable, then we have the following conclusion...} \} \\ &\leq E[\exp(s|X - Y|)] \leq e^{2s^2} \\ &\quad \{ \text{ ... and the theorem is proved under a more mild condition: subgaussian.} \} \end{aligned}$$

Now we prove for a general dimension d. Suppose f is a differentiable function. Since it is a Lipshitz function, we have

$$\|\nabla f(x)\|_2 \leq 1.$$

Note that

$$f(X) - f(Y) = \int_0^1 \frac{d}{d\theta} f((1-\theta)Y + \theta X) d\theta,$$

by Jensen's Inequality, we have

$$\exp(t(f(X) - f(Y))) \leq \int_0^1 \exp(t\frac{d}{d\theta} f(Y(1-\theta) + X\theta)) d\theta$$

By chain rule

$$E\exp(t(f(X) - f(Y))) \leq \int_0^1 E\exp(t\nabla f(X_\theta)(X - Y)) d\theta \leq \exp(Ct^2)$$

$$P(f(X) \geq t) \leq E[\exp(s(f(X) - t))] \leq \exp(Cst^2 - st).$$

Taking $s = \frac{1}{2C}$, we have

$$P(f(X) \geq t) \leq \exp(-\frac{1}{2C}t^2).$$

□

4.4 Sub-Gaussian Random Variables

Sub-gaussian r.v. is a generalization of Gaussian random variables.

定义4 (Subgaussian random variables)**.** *A r.v. is subgaussian if* $\exists$ *c, C such that*

$$P(|x| \geq t) \leq Ce^{-ct^2}, \forall t \geq 0.$$

引理4. *Let* X *be a zero mean r.v., then the following two statements are equivalent*

1. X *is subgaussian*

2. $\exists c_2, C_2$,
$$Ee^{c_2 X^2} \leq C_2$$

3. $\exists c_3, C_3$,
$$E(e^{tX}) \leq C_3 e^{c_3 t^2} \quad \forall t,$$
Equivalently, $\exists c_3, C_3$,
$$E(e^{t|X|}) \leq C_3 e^{c_3 t^2} \quad \forall t > 0,$$
by the fact that
$$e^{t|X|} \leq e^{tX} + e^{-tX}$$

4. $\exists c_3 > 0$
$$E(e^{tX}) \leq e^{c_3 t^2} \quad \forall t.$$

Proof. We prove this theorem by the following way: 1. $\Longrightarrow$ 2. $\Longrightarrow$ 3. $\Longrightarrow$ 1.

1. $\Longrightarrow$ 2.

$$\begin{aligned} E(e^{c_2 X^2}) &= \int_0^\infty 2c_2 u e^{c_2 u^2} P(|X| \geq u) du + 1 \\ &\leq C \int_0^\infty u e^{c_2 u^2} e^{-cu^2} du + 1 \\ &= C \int_0^\infty u e^{-cu^2/2} du + 1 \quad \text{(by taking } c_2 = c/2) \\ &= -C/c \int_0^\infty de^{-cu^2/2} + 1 \\ &= C/c + 1 \end{aligned}$$

2. $\Longrightarrow$ 3. It suffices to prove $\exists c_3, C_3$, such that

$$E(e^{tX-c_3t^2}) \leq C_3.$$

In fact we know that

$$\begin{aligned} E(e^{tX-c_3t^2}) &\leq E^{\frac{X^2}{4c_3}} \\ &\leq C_2 \quad (\text{by taking } c_3 = \tfrac{1}{4c_2}) \end{aligned}$$

3. $\Longrightarrow$ 1.

$$\begin{aligned} P(|X| \geq t) &\leq E \quad e^{\lambda(|X|-t)} \\ &\leq \quad C_3 e^{c_3\lambda^2-\lambda t} \\ &= \quad C_3 e^{-\frac{t^2}{4c_3}} \quad (\text{by taking } \lambda = \tfrac{1}{2c_3}) \end{aligned}$$

Finally (3) $\Longleftrightarrow$ (4).

(3) $\Longrightarrow$ (4). For $|t| \geq 1$, we have

$$Ee^{tX} \leq C_3 e^{c_3t^2} \leq e^{c_4t^2}.$$

For $0 \leq |t| < 1$, we have

$$\begin{aligned} Ee^{tX} &= 1 + E(\psi(tX)) \\ &\leq 1 + E(\psi|tX|) \\ &\leq 1 + t^2 E\psi(|X|) \\ &\leq e^{c_5t^2}. \end{aligned}$$

So,

$$Ee^{tX} \leq e^{\max\{c_4,c_5\}t^2}.$$

□

Examples about sub-gaussian random variable:

1. $N(0,1)$ $N(0,\sigma^2)$

$$E(e^{tX}) = e^{\frac{1}{2}\sigma^2 t^2}$$

2. Bounded R.V. with mean 0. By Hoeffding's inequality,

$$E[e^{tX}] \leq e^{t^2(b-a)^2/8}.$$

Sub-Gaussian distribution is a generalization of Gaussian distribution.

定理18. *Let $X_1, X_2, \ldots, X_n$ be i.i.d, mean-zero subgaussian random variables. Also let $a_1, a_2, \ldots, a_n \in \mathbb{R}$ be such that $\sum_k a_k^2 = 1$. Then $\sum a_k X_k$ is a subgaussian random variable.*

Proof.

$$\begin{aligned} E(e^{t\sum a_k X_k}) &= \Pi_k e^{t a_k X_k} \\ &\leq \Pi_k e^{c a_k^2 t^2} \\ &= e^{ct^2} \end{aligned}$$

□

推论1 (Hoeffding Tail Inequality)**.**

$$P\left[\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i\right| > t\right] \leq Ce^{-ct^2}.(C=2, c=\frac{1}{2})$$

4.5 Random Matrix

定理19 (Wigner's Semicircle Law)**.** *Consider and $N \times N$ matrix A with entries $A_{ij} \sim N(0,\sigma^2)$. Define*

$$A_n = \frac{1}{\sqrt{2n}}(A + A').$$

Then A_n is symmetric with

$$var(A_{ij}) = \begin{cases} \sigma^2/n, & \text{if } i \neq j; \\ 2\sigma^2/n, & \text{if } i = j. \end{cases}$$

The density of eigenvalues of A_n is given by

$$f_n(\lambda) := \lim_{\Delta t \to 0^+} \frac{1}{n} \#\{\lambda \leq \lambda_i \leq \lambda + \Delta t\}/\Delta t$$
$$\longrightarrow_{n\to\infty} \begin{cases} \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - \lambda^2}, & \text{if } |\lambda| \leq 2\sigma; \\ 0, & \text{if } i = j. \end{cases}$$

Below is the R code to demonstrate the semi-circle law:

```
n <- 5000;
m <- array(rnorm(n^2),c(n,n));
m2 <- (m+t(m))/sqrt(2*n);# Make m symmetric
lambda <- eigen(m2, symmetric=T, only.values = T);
e <- lambda$values;
hist(e,breaks=seq(-2.01,2.01,.01),
main=NA, xlab="Eigenvalues",freq=F)
x = seq(-2.01,2.01,.01)
y = sqrt(4-x^2)/(2*pi)
lines(x,y,col = 'red')
```

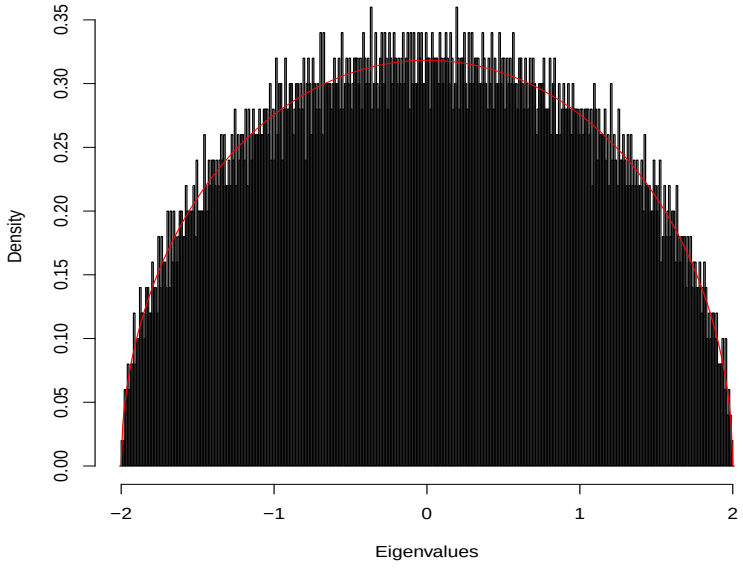

Figure 4.1: Demonstration of Semi-circle Law

Let $s_1 \geq s_2 \geq \ldots \geq s_p$ be the p singular values of $A_{n\times p}$. $(n \geq p)$ In this section, we bound s_1 and s_n. We mainly use the following result:

定理20 (Slepian's Inequality)**.** *Assume* $(X_t)_{t\in T}$,$(Y_t)_{t\in T}$ *are Gaussian centered process. If for all* $s,t\in T$,

$$E(X_t-X_s)^2\le E(Y_t-Y_s)^2,$$

then

$$E\left(\sup_{t\in T}X_t\right)\le E\left(\sup_{t\in T}Y_t\right).$$

Note: A Gaussian process is a stochastic process whose realizations consist of random values associated with every point in a range of times (or of space) such that each such random variable has a normal distribution. Moreover, every finite collection of those random variables has a multivariate normal distribution.

It is easy to show that

$$s_1=\sup_{\|u\|_2\le 1}\|Au\|_2=\sup_{\|u\|_2\le 1,\|v\|_2\le 1}<Au,v>$$

$$s_p=\inf_{\|u\|_2\le 1}\|Au\|_2=\inf_{\|u\|_2\le 1}\sup_{\|v\|_2\le 1}<Au,v>$$

定理21 (Upper Bound)**.** *Let* A *be an* $n\times p$ *Gaussian random matrix. Then*

$$E(s_1)\le\sqrt{n}+\sqrt{p}.$$

Proof. Let

$$X_{(u,v)}:=<Au,v>\ \text{ and } Y_{(u,v)}=<g,u>+<h,v>,$$

where $\|u\|_2=\|v\|_2=1$, $g=(g_1,g2,\dots g_p)$ is a gaussian random vector with g_i *i.i.d.* $\sim N(0,1)$, and $h=(h_1,h2,\dots h_n)$ is a gaussian random vector with h_i *i.i.d.* $\sim N(0,1)$. We can verify that for any (u,v) and (u',v')

$$E[X_{(u,v)}-X_{(u',v')}]^2\le E[Y_{(u,v)}-Y_{(u',v')}]^2.$$

The LHS:

$$E[X_{(u,v)}-X_{(u',v')}]^2=E(\sum_{ij}(A_{ij}u_iv_j-u'_iv'_j)^2)=\sum_{ij}(u_iv_j-u'_iv'_j)^2$$

The RHS:

$$E[Y_{(u,v)}-Y_{(u',v')}]^2=E[u^Tg+v^Th-u'^Tg-v'^Th]^2=\sum(u_i-u'_i)^2+\sum(v_j-v'_j)^2$$

Now we verify

$$\sum_{ij}(u_iv_j - u'_iv'_j)^2 \leq \sum(u_i - u'_i)^2 + \sum(v_j - v'_j)^2$$

$$\begin{aligned}\sum_{ij}(u_iv_j - u'_iv'_j)^2 &= \sum_i\sum_j(u_i^2u_j^2 + {u'_i}^2{v'_j}^2 - 2u_iu'_iv_jv'_j) \\ &= \sum_i\left[u_i^2 + {u'_i}^2 - 2u_iu'_i\left(\sum_j v_jv'_j\right)\right] \\ &= 2 - 2\left(\sum_i u_iu'_i\right)\left(\sum_j v_jv'_j\right)\end{aligned}$$

$$\sum(u_i - u'_i)^2 + \sum(v_j - v'_j)^2 = 4 - 2\sum_i u_iu'_i - 2\sum_j v_jv'_j$$

So

$$\begin{aligned}&\left[\sum(u_i - u'_i)^2 + \sum(v_j - v'_j)^2\right] - \sum_{ij}(u_iv_j - u'_iv'_j)^2 \\ = &\left[4 - 2\sum_i u_iu'_i - 2\sum_j v_jv'_j\right] - \left[2 - 2\left(\sum_i u_iu'_i\right)\left(\sum_j v_jv'_j\right)\right] \\ = &\ 2 + 2\left(\sum_i u_iu'_i\right)\left(\sum_j v_jv'_j\right) - 2\sum_i u_iu'_i - 2\sum_j v_jv'_j \\ = &\ 2\left(\sum_i u_iu'_i - 1\right)\left(\sum_j v_iv'_j - 1\right) \geq 0\end{aligned}$$

By Slepian's Inequality, we have

$$\begin{aligned}E(\sup X_{(u,v)}) &\leq E(\sup Y_{(u,v)}) \\ &= E(\sup < g, u > + < h, v >) \\ &= E(||g||_2) + E(||v||_2) \\ &< \sqrt{E(\|g\|_2^2)} + \sqrt{E(\|g\|_2^2)} \\ &= \sqrt{n} + \sqrt{p}\end{aligned}$$

□

定理22 (Gordon's Inequality)**.** *Let* $(X_{u,v})_{u\in U,V\in V}$ *and* $(Y_{u,v})_{u\in U,V\in V}$ *be centered Gaussain random process. If*

$$(1)\quad E(X_{u,v}-X_{u',v'})^2 \le E(Y_{u,v}-Y_{u',v'})^2 \text{ if } u\neq u' \text{ ;}$$
$$(2)\quad E(X_{u,v}-X_{u,v'})^2 = E(Y_{u,v}-Y_{u,v'})^2$$

Then

$$E\sup_{u\in U}\inf_{v\in V} X_{u,v} \le E\sup_{u\in U}\inf_{v\in V} Y_{u,v}.$$

Note that Gordon's inequality implies Slepian's inequality by taking the index set V to be a singleton set (i.e., $|V|=1$).

By applying Gordon's Inequality for $-X,-Y$, we have

$$E\inf_{u\in U}\sup_{v\in V} X_{u,v} \ge E\inf_{u\in U}\sup_{v\in V} Y_{u,v}.$$

定理23 (Lower Bound)**.** *Let* A *be an* $n\times p$ *Gaussian random matrix. Then*

$$E(s_n) \ge \sqrt{n}-\sqrt{p}.$$

Proof. We already show that $X_{u,v}$ and $Y_{u,v}$ constructed as before satisfy condition (1) in Theorem 22. Now we verify condition (2) in Theorem 22. That is,

$$\sum_{ij}(u_iv_j-u_iv'_j)^2 = \sum(u_i-u_i)^2+\sum(v_j-v'_j)^2,$$

which holds by the fact that $\sum u_i^2=1$. By Gordon's inequality we have

$$\begin{aligned}
E(s_n) &= E\inf_{u\in U}\sup_{v\in V} X_{u,v}\\
&\ge E\inf_{u\in U}\sup_{v\in V} Y_{u,v}\\
&= E\inf_{u\in U}\sup_{v\in V}(<g,u>+<h,v>)\\
&= E(\inf_{u\in U}<g,u>)+E(\|h\|_2)\\
&\ge -E(\|g\|_2)+E(\|h\|_2)\\
&\ge \sqrt{n}-\sqrt{p}.
\end{aligned}$$

The last inequality uses the fact that

$$f(n) := \sqrt{n}-E\sqrt{X_1^2+\ldots X_n^2}$$

is a decreasing function of n. □

Some facts about chi-square distribution:

1. density
$$f(x;n) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2} 1_{x\geq 0}$$

2. Gamma function
$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-x} dx$$

3. Expectation and Second moment.
$$E(X) = n \text{ and } E(X^2) = 2n + n^2$$

4. Squre Root.
$$E(\sqrt{X}) = \sqrt{2}\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})}$$

By the fact that
$$\Gamma(\frac{1}{2}) = \sqrt{\pi} \text{ and } \Gamma(1+x) = x\Gamma(x),$$

we have

5.
$$\mu(1) = \sqrt{\frac{2}{\pi}}$$
$$\mu(2) = \sqrt{\frac{\pi}{2}}$$
$$\mu(k+2) = (1+\frac{1}{k})\mu(k)$$

Finally we use some known concentration inequalities for the extreme eigenvalues of Gaussian random matrices [Davidson, K. R. and Szarek, S. J. (2001).] to bound the eigenvalues of a Gaussian random matrix. Although these results hold more generally, our interest here is on scalings (n, q) such that $q/n \to 0$.

定理24 (Davidson, K. R. and Szarek, S. J. 2001.)**.** *Let $\Gamma \in R^{n\times q}$ be a random matrix whose entries are i.i.d. from $N(0, 1/n)$, $q \leq n$. Let the singular values of Γ be $s_1(\Gamma) \geq \ldots \geq s_q(\Gamma)$. Then*
$$\max\left\{P\left[s_1(\Gamma) \geq 1 + \sqrt{\frac{q}{n}} + t\right], P\left[s_q(\Gamma) \leq 1 - \sqrt{\frac{q}{n}} - t\right]\right\} \leq \exp\{-nt^2/2\}.$$

Proof. Let $A = \sqrt{n}\Gamma$, then A satisfies the conditions in Theorems 21 and 23. We will show that both $s_1(A)$ and $s_q(A)$ are Lipschitz functions with constant 1. That is

$$|s_1(A) - s_1(B)| \leq \|A - B\|_2 \text{ and } |s_n(A) - s_n(B)| \leq \|A - B\|_2.$$

In fact

$$\begin{aligned}
\left| \max_{\|x\|_2=1} \|Ax\|_2 - \max_{\|x\|_2=1} \|Bx\|_2 \right| &\leq \max_{\|x\|_2=1} \|(A-B)x\|_2 \\
&= \sqrt{\Lambda_{\max}(A-B)^T(A-B)} \\
&\leq \sqrt{\sum \Lambda_i (A-B)^T(A-B)} \\
&= \sqrt{tr((A-B)^T(A-B))} \\
&= \|A-B\|_2
\end{aligned}$$

$$\begin{aligned}
\min_{\|x\|_2=1} \|Ax\|_2 &\leq \min_{\|x\|_2=1} (\|Bx\|_2 + \|(A-B)x\|_2) \\
&\leq \min_{\|x\|_2=1} \|Bx\|_2 + \max \|(A-B)x\|_2 \\
&\leq \min_{\|x\|_2=1} \|Bx\|_2 + \|A-B\|_2
\end{aligned}$$

Applying Gaussian concentration inequality for Lipschitz functions Theorem , we have

$$P(s_1(A) - Es_1(A) \geq t) \leq e^{-\frac{1}{2}t^2}$$

$$P(-s_n(A) + Es_n(A) \geq t) \leq e^{-\frac{1}{2}t^2}$$

$$P(s_1(A) \geq Es_1(A) + t) \leq e^{-\frac{1}{2}t^2}$$

$$P(s_n(A) \leq Es_n(A) - t) \leq e^{-\frac{1}{2}t^2}$$

(Note: here we used c in Theorem to be $1/2$, which is not shown in this Lecture.) By using Theorems 23 and 21:

$$Es_1(A) \leq \sqrt{n} + \sqrt{q} \text{ and } Es_n(A) \geq \sqrt{n} - \sqrt{q},$$

we have

$$P(s_1(A) \geq \sqrt{n} + \sqrt{q} + t) \leq e^{-\frac{1}{2}t^2}$$

$$P(s_n(A) \leq \sqrt{n} - \sqrt{q} - t) \leq e^{-\frac{1}{2}t^2}$$

Taking $A = \sqrt{n}\Gamma$ in the last inequality we have

$$P(\sqrt{n}s_1(\Gamma) \geq \sqrt{n} + \sqrt{q} + \sqrt{n}t) \leq e^{-\frac{n}{2}t^2}$$

$$P(\sqrt{n}s_n(\Gamma) \leq \sqrt{n} - \sqrt{q} - \sqrt{n}t) \leq e^{-\frac{n}{2}t^2}.$$

This is to say:

$$\max\left\{P\left[s_1(\Gamma) \geq 1 + \sqrt{\frac{q}{n}} + t\right], P\left[s_q(\Gamma) \leq 1 - \sqrt{\frac{q}{n}} - t\right]\right\} \leq \exp\{-nt^2/2\}.$$

□

Using Theorem 24, we now have some useful results.

定理25. *Let $U \in R^{n\times q}$ be a random matrix with elements from the standard normal distribution (i.e., $U_{ij} \sim N(0,1)$, i.i.d.) Assume that $q/n \to 0$. Let the eigenvalues of $\frac{1}{n}U^TU$ be $\Lambda_1(\frac{1}{n}U^TU) \geq \ldots \geq \Lambda_q(\frac{1}{n}U^TU)$. Then when n is big enough,*

$$P\left[\frac{1}{2} \leq \Lambda_i(\frac{1}{n}U^TU) \leq 2\right] \geq 1 - 2\exp(-0.03n). \tag{4.1}$$

推论2. *Let $X \in R^{n\times q}$ be a random matrix, of which, the rows are i.i.d. from the normal distribution with mean 0 and covariance Σ. Assume that $0 < \tilde{C}_{\min} \leq \Lambda_i(\Sigma) \leq \tilde{C}_{\max} < \infty$ and $q/n \to 0$, then when n is big enough,*

$$P\left[\frac{1}{2}\tilde{C}_{\min} \leq \Lambda_i(\frac{1}{n}X^TX) \leq 2\tilde{C}_{\max}\right] \geq 1 - 2\exp(-0.03n). \tag{4.2}$$

Proof. Let $U = X\Sigma^{-\frac{1}{2}}$, then U satisfies the condition in Lemma 25. Then

$$P\left[\frac{1}{2} \leq \Lambda_i(\frac{1}{n}U^TU) \leq 2\right] \geq 1 - 2\exp(-0.03n).$$

Since

$$\tilde{C}_{\min}\Lambda_1(\frac{1}{n}U^TU) \leq \Lambda_i(\frac{1}{n}X^TX) \leq \tilde{C}_{\max}\Lambda_q(\frac{1}{n}U^TU),$$

result (4.2) is obtained immediately. □

4.6 Stein's Method

4.6.1 James-Stein Estimator

Suppose $\theta \in \mathbb{R}$ is an unknow parameter vector. $y_1, y_2, \ldots, y_n$ are n independent observations normally distributed:

$$y_i \sim N(\theta, \sigma^2 I).$$

How to estimate θ?

Bias-variance trade-off.

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 \\ &= E(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 \\ &= var(\hat{\theta}) + bias^2 \end{aligned}$$

$$\hat{\theta}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\bar{y}\|^2}\right) \bar{y}$$

Property: James-Stein estimator has smaller MSE than LS estimtor $\bar{y}$ when $m \geq 3$.

4.6.2 Stein's Method for Concentration Inequalities

Stein's method was introduced by Charles Stein in the context of normal approximation for sums of dependent random variables. A general version of Stein' s method for concentration inequalities was introduced for the first time in the Ph.D. thesis of Sourav Chatterjee. We will introduce this powerful method here.

定理26. *Suppose (X, X') is an exchangeable pair of random variables, that is .*

$$(X, X') =_d (X', X).$$

Suppose $f(x)$ and $F(x, y)$ are square-integrable functions. F is antisymmetric:

$$F(X, X') = -F(X', X).$$

$$E\left[F(X, X') \mid X\right] = f(X).$$

Define

$$\Delta(X) = \frac{1}{2}E\left[|(f(X) - f(X'))F(X, X')| \mid X\right].$$

Then $E(f(X)) = 0$ *and the following concentration results hold for* $f(X)$*:*

1) If $E(\Delta(X)) < \infty$*, then* $Var(f(X)) = \frac{1}{2}E\left((f(X) - f(X'))F(X, X')\right)$.

2) Assume that

$$Ee^{\theta f(X)}|F(X, X')| < \infty$$

for all θ*. IF there exists non-negative constant* B *and* C *such that*

$$\Delta(X) \leq Bf(X) + C, a.s.,$$

then for any $t \geq 0$*,*

$$P(f(X) \geq t) \leq \exp\left(-\frac{t^2}{2C + 2Bt}\right) \text{ and } P(f(X) \leq -t) \leq \exp\left(-\frac{t^2}{2C}\right).$$

3) For any positive integer k*, we have the following inequality:*

$$E(f(X)^{2k}) \leq (2k - 1)^k E(\Delta(X)^k).$$

Proof. For any square-integrable function $h(x)$, we have

$$E(h(X)F(X, X')) = E(h(X')F(X', X)) = -E(h(X')F(X, X')).$$

So,

$$E(h(X)f(X)) = E(h(X)E(F(X, X')|X)) = E(h(X)F(X, X')) = \frac{1}{2}E([h(X) - h(X')]F(X, X')).$$

By taking $h(x) = 1$, we have $E(f(X)) = 0$. By taking $h(x) = f(x)$, we have $E(f^2(X)) = \frac{1}{2}E([f(X) - f(X')]F(X, X'))$.

Now we prove 2). Let $m(\theta) = E(e^{\theta f(X)})$.

$$m'(\theta) = E(e^{\theta f(X)}f(X)) = \frac{1}{2}E(e^{\theta f(X)} - e^{\theta f(X')})F(X, X').$$

$$\begin{aligned}\left|\frac{e^x - e^y}{x-y}\right| &= \int_0^1 e^{tx+(1-t)y}dt \\ &\leq \int_0^1 te^x + (1-t)e^y dt = \frac{1}{2}(e^x + e^y).\end{aligned}$$

So we must have

$$\begin{aligned}|m'(\theta)| &\leq \frac{1}{2}E(|e^{\theta f(X)} - e^{\theta f(X')}||F(X,X')|) \\ &\leq \frac{|\theta|}{4}E\left((e^{\theta f(X)} + e^{\theta f(X')})|f(X) - f(X')F(X,X')|\right) \\ &= \frac{|\theta|}{2}(E\left((e^{\theta f(X)}\Delta(X)\right) + E\left(e^{\theta f(X')}\Delta(X')\right) \\ &= |\theta|E\left((e^{\theta f(X)}\Delta(X)\right) \\ &\leq |\theta|E(e^{\theta f(X)}Bf(X) + C) \\ &= B|\theta|m'(\theta) + C|\theta|m(\theta)\end{aligned}$$

Since $m(\theta)$ is a convex function of θ and $m'(0) = 0$, so $m'(\theta)$ and θ has the same sign. So for $0 \leq \theta < 1/B$, the above inequality is equivalent to

$$\frac{d\log(m(\theta))}{d\theta} \leq \frac{C\theta}{1 - B\theta}$$

So

$$\log(m(\theta)) \leq \int_0^\theta \frac{Cu}{1-Bu}du \leq \frac{C\theta^2}{2(1-B\theta)}$$

$$\begin{aligned}P(f(X) \geq t) &\leq Ee^{\theta[f(X)-t]} \\ &= \exp(\log(m(\theta)) - \theta t) \\ &\leq \exp(\frac{C\theta^2}{2(1-B\theta)} - \theta t) \\ &= \exp(\frac{(C+2Bt)\theta^2 - 2\theta t}{2(1-B\theta)}) \\ &= \exp(-\frac{t^2}{2(C+Bt)})\text{by taking } \theta = \tfrac{t}{C+2Bt}\end{aligned}$$

For $\theta < 0$, we have

$$-m'(\theta) \leq B|\theta|m'(\theta) + C|\theta|m(\theta) < C|\theta|m(\theta) = -C\theta m(\theta)$$

$$\frac{d\log(m(\theta))}{d\theta} \geq C\theta$$

We have

$$\log m(\theta) \leq \frac{C\theta^2}{2}$$

$$\begin{aligned} P(-f(X) \geq t) &\leq Ee^{\theta[-f(X)-t]} \\ &= \exp(\log(m(-\theta)) - \theta t) \\ &\leq \exp(\frac{C\theta^2}{2} - \theta t) \\ &= \exp(-\frac{t^2}{2C}) \text{by taking } \theta = \frac{t}{C} \end{aligned}$$

Finally, we prove 3).

$$E(f(X)^{2k}) = \frac{1}{2}E((f(X)^{2k-1}) - f(X')^{2k-1})F(X, X').$$

By the fact that

$$|x^{2k-1} - y^{2k-1}| \leq \frac{2k-1}{2}(x^{2k-2} + y^{2k-2})|x - y|,$$

we have

$$E(f(X)^{2k}) \leq (2k-1)E(f(X)^{2k-2})\Delta(X) \leq (2k-1)(E(f(X)^{2k}))^{(k-1)/k}\Delta(X)E(\Delta(X)^k)^{1/K}$$

□

Example 1. Let $Y_i, i = 1, \ldots, n$ be independent random variables.

$$\mu_i = E(Y_i) \quad and \quad \sigma^2 = var(Y_i).$$

Let I be chosen uniformly random from $\{1, 2, \ldots, n\}$, and defining

$$X' = \sum_{j \neq I} Y_j + Y'_I,$$

where $Y_i, \ldots, Y_n$ are independent copies of $Y_1, \ldots, Y_n$. It can be verified that (X, X') is an exchangeable pair. Let $F(x, y) = n(x - y)$. Then F is an antisymmetric function.

$$\begin{aligned} E(F(X,X') \mid Y_1,\ldots,Y_n) &= E(n(Y_I - Y_I') \mid Y_1,\ldots,Y_n) \\ &= \sum_i E(n(Y_I - Y_I') \mid Y_1,\ldots,Y_n I = i) P(I = i | Y_1,\ldots,Y_n) \\ &= \sum_i (Y_i - \mu_i) = X - E(X) \end{aligned}$$

We have

$$f(X) = X - E(X).$$

$$\begin{aligned} \Delta(X) &= \frac{1}{2} E\left[|(f(X) - f(X'))F(X,X')| \mid X\right] \\ &= \frac{1}{2} E\left[|n(X - X')(X - X')| \mid X\right] \\ &= \frac{1}{2} E\left[n(Y_I - Y_I')^2 \mid X\right] \\ &= \frac{1}{2} \sum_i E\left[(Y_i - Y_i')^2 \mid X\right] \end{aligned}$$

From Theorem 1), we have

$$var(f(X)) = \frac{1}{2} \sum_i E\left[(Y_i - Y_i')^2\right] = \sum_i \sigma_i^2$$

If $|Y_i - \mu_i| \leq c_i$, for each i, then

$$\begin{aligned} E\left[(Y_i - Y_i')^2 \mid X\right] &= E\left[(Y_i - \mu_i)^2 \mid X\right] + E\left[(Y_i' - \mu_i)^2\right] \\ &\leq c_i^2 + \sigma_i^2. \end{aligned}$$

So

$$\Delta(X) \leq 0 * f(X) + \frac{1}{2} \sum_i (c_i^2 + \sigma_i^2)$$

By 2), we have

$$P(|f(X) - E(f(X))| \geq t) \leq 2e^{-\frac{t^2}{\sum_i (c_i^2 + \sigma_i^2)}}$$

4.7 Homework

1.
$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \geq 0$$

2. For a non-negative random variable (which can be continuous and discrete), show that
$$E(X) = \int_0^\infty P(X \geq t) dt$$

3. $X_i, i = 1, \ldots, n$ i.i.d. $\sim N(0,1)$. Show that
$$f(n) := \sqrt{n} - E\sqrt{X_1^2 + \ldots X_n^2}$$
is a decreasing function of n

4. Let $U \in R^{n \times q}$ be a random matrix with elements from the standard normal distribution (i.e., $U_{ij} \sim N(0,1)$, i.i.d.) Assume that $q/n \to 0$. Let the eigenvalues of $\frac{1}{n}U^T U$ be $\Lambda_1(\frac{1}{n}U^T U) \geq \ldots \geq \Lambda_q(\frac{1}{n}U^T U)$. Then when n is big enough,
$$P\left[\frac{1}{2} \leq \Lambda_i(\frac{1}{n}U^T U) \leq 2\right] \geq 1 - 2\exp(-0.03n). \tag{4.3}$$

5. Let $Y_i, i = 1, \ldots, n$ be independent random variables.
$$\mu_i = E(Y_i) \quad and \quad \sigma^2 = var(Y_i).$$
Let I be chosen uniformly random from $\{1, 2, \ldots, n\}$, and defining
$$X' = \sum_{j \neq I} Y_j + Y_I',$$
where $Y_i, \ldots, Y_n$ are independent copies of $Y_1, \ldots, Y_n$. Show that (X, X') is an exchangeable pair:
$$P(X \leq t_1, X' \leq t_2) = P(X' \leq t_1, X \leq t_2).$$

References

Boucheron, Lugosi and Bousquet. Concentration Inequalities.

Bousquet, Boucheron and Lugosi. Introduction to Statistical Learning.

Sourav Chatterjee. STEIN's Method for concentration inequalities.

www.math.ucdavis.edu/∼dneedell280.html

http://terrytao.wordpress.com/2009/06/09/talagrands-concentration-inequality/

Chapter 5

Properties of the Lasso

5.1 Sign consistency

To define the Lasso estimate, suppose the observed data are independent pairs $\{(x_i, Y_i)\} \in R^p \times R$ for $i = 1, 2, \ldots, n$ following the linear regression model

$$Y_i = x_i^T \beta^* + \epsilon_i, \tag{5.1}$$

where x_i^T is a row vector representing the predictors for the ith observation, Y_i is the corresponding ith response variable, ϵ_i's are independent and mean zero noise terms, and $\beta^* \in R^p$. Use $\mathbf{X} \in R^{n \times p}$ to denote the $n \times p$ design matrix with $x_k^T = (\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kp})$ as its kth row and with $X_j = (\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jn})^T$ as its jth column, then

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = (X_1, X_2, \ldots, X_p).$$

Let $Y = (Y_1, \ldots, Y_n)^T$ and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T \in R^n$. The Lasso estimate (**?**) is then defined as the solution to a penalized least squares problem (with regularization parameter λ):

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{2n} \|Y - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1, \tag{5.2}$$

where for some vector $x \in R^k$, $\|x\|_r = (\sum_{i=1}^k |x_i|^r)^{1/r}$.

Define

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Define $=_s$ such that $\hat{\beta}(\lambda) =_s \beta^*$ if and only if $\text{sign}(\hat{\beta}(\lambda)) = \text{sign}(\beta^*)$ elementwise.

We want to know if the Lasso can recover the sparsity pattern correctly. To be precisely, we introduce the term **Sign Consistency:**

定义5. *The Lasso is **sign consistent** if there exists a sequence λ_n such that,*

$$P\left(\hat{\beta}(\lambda_n) =_s \beta^*\right) \to 1, \ \textit{as } n \to \infty.$$

This section examines when the Lasso is sign consistent and when it is not sign consistent

under the sparse Poisson-like model for a nonrandom design matrix $\mathbf{X}$. First, some notation,

$$x_i(S) = e_i^T X(S),$$

where e_i is the unit vector with ith element one and the rest zero. Because $S = \{j : \beta_j^* \neq 0\}$ is the sparsity index set, $x_i(S)$ is a row vector of dimension q. Define

$$\beta^*(S) = (\beta_j^*)_{j \in S} \quad \text{and} \quad \overrightarrow{b} = \text{sign}(\beta^*(S)).$$

Suppose the Irrepresentable Condition holds. That is, for some constant $\eta \in (0, 1]$,

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \leq 1 - \eta. \tag{5.3}$$

The ℓ_∞ norm of a vector, $\| \cdot \|_\infty$, is defined as the vector's largest element in absolute value. In addition, assume that

$$\Lambda_{\min} \left(\frac{1}{n} X(S)^T X(S) \right) \geq C_{\min} > 0, \tag{5.4}$$

where $\Lambda_{\min}$ denotes the minimal eigenvalue and $C_{\min}$ is some positive constant. Condition (5.4) guarantees that matrix $X(S)^T X(S)$ is invertible. These conditions are also needed in **?** for sign consistency of the Lasso under the standard model. Define

$$\Psi(\mathbf{X}, \beta^*, \lambda) = \lambda \left[\eta \, (C_{\min})^{-1/2} + \left\| \left(\frac{1}{n} X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \right] \leq \lambda \left[\eta \, (C_{\min})^{-1/2} + \sqrt{q} C_{\min}^{-1} \right]$$

with which:

定理27. *Suppose that data* $(\mathbf{X}, Y)$ *follows linear model described by Equations* (5.1) *and each column of* $\mathbf{X}$ *is normalized to* l_2*-norm* $\sqrt{n}$*. Assume that (5.3) and (5.4) hold. If* λ *satisfies*

$$M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda),$$

then with probability greater than

$$1 - 2 \exp \left\{ -\frac{n\lambda^2 \eta^2}{2\sigma^2} + \log(p) \right\},$$

the Lasso has a unique solution $\hat{\beta}(\lambda)$ *with* $\hat{\beta}(\lambda) =_s \beta^*$*.*

Before proving this theorem, we first introduce a lemma about the solution of the Lasso.

引理5. *For linear model $Y = \mathbf{X}\beta^* + \epsilon$, assume that the matrix $X(S)^T X(S)$ is invertible. Then for any given $\lambda > 0$ and any noise term $\epsilon \in R^n$, there exists a Lasso estimate $\hat{\beta}(\lambda)$ which satisfies $\hat{\beta}(\lambda) =_s \beta^*$, if and only if the following two conditions hold*

$$\left| X(S^c)^T X(S)(X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] - \frac{1}{n} X(S^c)^T \epsilon \right| \leq \lambda, \tag{5.5}$$

$$sign\left(\beta^*(S) + (\frac{1}{n} X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] \right) = sign(\beta^*(S)), \tag{5.6}$$

where the vector inequality and equality are taken elementwise. Moreover, if (5.5) *holds strictly, then*

$$\hat{\beta} = (\hat{\beta}^{(1)}, 0)$$

is the unique optimal solution to the Lasso problem (5.2)*, where*

$$\hat{\beta}^{(1)} = \beta^*(S) + (\frac{1}{n} X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\beta^*) \right]. \tag{5.7}$$

Proof. The Lasso estimate satisfies the following condition:

$$\partial[\frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1]|_{\beta=\hat{\beta}\lambda} = 0,$$

that is,

$$\frac{1}{n} X^T (Y - X\hat{\beta}) + \lambda \overrightarrow{s} = 0,$$

where $\overrightarrow{s}_j = sign(\hat{\beta}_j(\lambda)), if \hat{\beta}_j(\lambda) \neq 0$ and $\overrightarrow{s}_j \in [-1, 1], if \hat{\beta}_j(\lambda) = 0$. $\hat{\beta}(\lambda) =_s \beta^*$ if and only if

$$\frac{1}{n} X(S)^T (Y - X(S)\hat{\beta}(S)) + \lambda sign(\beta^*(S)) = 0,$$

and

$$|\frac{1}{n} X(S^c)^T (Y - X(S)\hat{\beta}(S))| \leq \lambda.$$

By substituting Y with $X(S)\beta^*(S) + \epsilon$ and solving $\beta^*(S)$ we complete the proof.

□

Now we prove Theorem 27.

Proof. Define

$$\overrightarrow{b} = \text{sign}(\beta^*(S)),$$

and denote by e_i the vector with 1 in the ith position and zeroes elsewhere. Define

$$U_i = e_i^T (\frac{1}{n} X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda \overrightarrow{b} \right],$$

$$V_j = X_j^T \left\{ X(S)(X(S)^T X(S))^{-1} \lambda \overrightarrow{b} - \left[X(S)(X(S)^T X(S))^{-1} X(S)^T - I) \right] \frac{\epsilon}{n} \right\}.$$

By rearranging terms, it is easy to see that (5.5) holds strictly if and only if

$$\mathcal{M}(V) = \left\{ \max_{j \in S^c} |V_j| < \lambda \right\} \tag{5.8}$$

holds. If we define $M(\beta^*) = \min_{j \in S} |\beta_j^*|$ (recall that $S = \{j : \beta_j^* \neq 0\}$ is the sparsity index), then the event

$$\mathcal{M}(U) = \left\{ \max_{i \in S} |U_i| < M(\beta^*) \right\}, \tag{5.9}$$

is sufficient to guarantee that condition (5.6) holds. Finally, a proof of Theorem 27.

This proof is divided into two parts. First we analysis the asymptotic probability of event $\mathcal{M}(V)$, and then we analysis the event of $\mathcal{M}(U)$.

Analysis of $\mathcal{M}(V)$: Note from (5.8) that $\mathcal{M}(V)$ holds if and only if $\frac{\max_{j \in S^c} |V_j|}{\lambda} < 1$. Each random variable V_j is Gaussian with mean

$$\mu_j = \lambda X_j^T X(S)(X(S)^T X(S))^{-1} \overrightarrow{b}.$$

Define $\tilde{V}_j = X_j^T \left[I - X(S)(X(S)^T X(S))^{-1} X(S)^T \right] \frac{\epsilon}{n}$, then $V_j = \mu_j + \tilde{V}_j$. Using condition (5.3), we have $|\mu_j| \leq (1 - \eta)\lambda$ for all $j \in S^c$, from which we obtain that

$$\frac{1}{\lambda} \max_{j \in S^c} |\tilde{V}_j| < \eta \Rightarrow \frac{\max_{j \in S^c} |V_j|}{\lambda} < 1.$$

By the Gaussian comparison result, we have

$$P \left[\frac{1}{\lambda} \max_{j \in S^c} |\tilde{V}_j| \geq \eta \right] \leq 2(p - q) \exp\{ - \frac{\lambda^2 \eta^2}{2 \max_{j \in S^c} E(\tilde{V}_j^2)} \}.$$

Since

$$E(\tilde{V}_j^2) = \frac{1}{n^2} X_j^T H [VAR(\epsilon)] H X_j,$$

where $H = I - X(S)(X(S)^T X(S))^{-1} X(S)^T$ which has maximum eigenvalue equal to 1, and $VAR(\epsilon)$ is the variance-covariance matrix of ϵ, which is a diagonal matrix with the ith diagonal

element equal to σ^2.

$$E(\tilde{V}_j^2) \leq \frac{\sigma^2}{n^2}\|X_j\|_2^2 = \frac{\sigma^2}{n}.$$

Therefore,

$$P\left[\frac{1}{\lambda}\max_j|\tilde{V}_j| \geq \eta\right] \leq 2(p-q)\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2}\right\}.$$

So, we have

$$\begin{aligned} P\left[\frac{1}{\lambda}\max_j|V_j| < 1\right] &\geq 1 - P\left[\frac{1}{\lambda}\max_j|\tilde{V}_j| \geq \eta\right] \\ &\geq 1 - 2(p-q)\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2}\right\}. \end{aligned}$$

Analysis of $\mathcal{M}(U)$:

$$\max_i|U_i| \leq \|(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T\epsilon\|_\infty + \lambda\|(\frac{1}{n}X(S)^TX(S))^{-1}\overrightarrow{b}\|_\infty$$

Define $Z_i := e_i^T(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T\epsilon$. Each Z_i is a normal Gaussian with mean 0 and variance

$$\begin{aligned} var(Z_i) &= e_i^T(\frac{1}{n}X(S)^TX(S))^{-1}\frac{1}{n}X(S)^T[VAR(\epsilon)]\frac{1}{n}X(S)(\frac{1}{n}X(S)^TX(S))^{-1}e_i \\ &\leq \frac{\sigma^2}{nC_{\min}}. \end{aligned}$$

So, for any $t > 0$,

$$P(\max_{i\in S}|Z_i| \geq t) \leq 2q\exp\{-\frac{t^2nC_{\min}}{2\sigma^2}\},$$

by taking $t = \frac{\lambda\eta}{\sqrt{C_{\min}}}$, we have

$$P(\max_{i\in S}|Z_i| \geq \frac{\lambda\eta}{\sqrt{C_{\min}}}) \leq 2q\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2}\right\}.$$

Recall the definition of $\Psi(\mathbf{X},\beta^*,\lambda) = \lambda\left[\eta\ (C_{\min})^{-1/2} + \left\|\left(\frac{1}{n}X(S)^TX(S)\right)^{-1}\overrightarrow{b}\right\|_\infty\right]$, we have

$$P(\max_i|U_i| \geq \Psi(\mathbf{X},\beta^*,\lambda)) \leq 2q\exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2}\right\}.$$

By condition $M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda)$, we have

$$P(\max_i |U_i| < M(\beta^*)) \geq 1 - 2q \exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2}\right\}.$$

At last, we have

$$P\left[\mathcal{M}(V)\& \ \mathcal{M}(U)\right] \geq 1 - 2p \exp\left\{-\frac{n\lambda^2\eta^2}{2\sigma^2}\right\}$$

□

Theorem 27 gives a non-asymptotic result on the Lasso's sparsity pattern recovery property. The next corollary specifies a sequence of λ's that can asymptotically recover the true sparsity pattern. The essential requirements are that

$$(1)\ \frac{n\lambda^2}{\sigma^2 \log(p+1)} \to \infty \ \text{ and } (2)\ \ M(\beta^*) > \Psi(\mathbf{X}, \beta^*, \lambda).$$

Slightly stronger conditions:

$$(1)\ \frac{n\lambda^2}{\sigma^2 \log(p+1)} \to \infty \ \text{ and } (2)\ \ M(\beta^*) > \lambda(\eta\sqrt{C_{\min}} + \sqrt{q}/C_{\min}).$$

Define,

$$\Gamma(\mathbf{X}, \beta^*, \sigma^2) = \frac{n\eta^2[M(\beta^*)]^2}{8\sigma^2(\eta\ C_{\min}^{-1/2} + \sqrt{q}\ C_{\min}^{-1})^2 \log(p+1)}.$$

推论3. *As in Theorem 27, Suppose that data* $(\mathbf{X}, Y)$ *follows linear model described by Equations* (5.1) *and each column of* $\mathbf{X}$ *is normalized to* l_2*-norm* $\sqrt{n}$*. Assume that (5.3) and (5.4) hold. Take* λ *such that*

$$\lambda = \frac{M(\beta^*)}{2\left(\eta\ C_{\min}^{-1/2} + \sqrt{q}\ C_{\min}^{-1}\right)}, \tag{5.10}$$

then $\hat{\beta}(\lambda) =_s \beta^*$ *with probability greater than*

$$1 - 2\exp\left\{-\left(\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha) - 1\right)\log(p+1)\right\}.$$

If $\Gamma(\mathbf{X}, \beta^*, \sigma^2) \to \infty$*, then* $P[\hat{\beta}(\lambda) =_s \beta^*]$ *converges to one.*

This corollary gives a class of heteroscedastic models for which the Lasso gives a sign consistent estimate of β^*. This class requires that $\Gamma(\mathbf{X}, \beta^*, \sigma^2) \to \infty$ which means that

$$SNR := \frac{n[M(\beta^*)]^2}{\sigma^2} = \Omega\left(q \log(p+1)\right), \tag{5.11}$$

where $a_n = \Omega(b_n)$ means that a_n grows faster than b_n, that is, $a_n/b_n \to \infty$. In other words, this condition requires that SNR grows fast enough.

The next corollary addresses the classical setting, where p, q, and β^* are all fixed and n goes to infinity. While this is a straightforward result from Corollary 3, it removes some of the complexities and leads to good intuition. Since $M(\beta^*)$ and $\|\beta^*\|_2$ do not change with n, $\Gamma(\mathbf{X}, \beta^*, \sigma^2, \alpha) \to \infty$ in Corollary 3 when $n \to \infty$. Then:

推论4. *As in Theorem 27, Suppose that data* $(\mathbf{X}, Y)$ *follows linear model described by Equations* (5.1) *and each column of* $\mathbf{X}$ *is normalized to* l_2*-norm* $\sqrt{n}$*. Assume that (5.3) and (5.4) hold. In the classical case when* p, q *and* β^* *are fixed, by choosing* λ *as in equation* (5.10)*,*

$$P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \to 1$$

as $n \to \infty$*.*

A more beautiful result:

推论5. *As in Theorem 27, Suppose that data* $(\mathbf{X}, Y)$ *follows linear model described by Equations* (5.1) *and each column of* $\mathbf{X}$ *is normalized to* l_2*-norm* $\sqrt{n}$*. Assume that (5.3) and (5.4) hold. In the classical case when* p, q *and* β^* *are fixed, by choosing* λ *such that* $n\lambda^2 \to \infty$ *and* $\lambda \to 0$*, then*

$$P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \to 1 \text{ and } \|\hat{\beta}(\lambda) - \beta^*\|_\infty \to_P 0$$

as $n \to \infty$*.*

A suitable choice of λ is $\lambda = \log n/\sqrt{n}$. So far the results have given sufficient conditions for sign consistency of the Lasso. To further understand how the sign consistency of the Lasso might be sensitive to the heteroscedastic model, the next theorem gives necessary conditions on the ratio of β_j^* to the noise level.

定理28 (Necessary Conditions)**.** *Suppose that data* $(\mathbf{X}, Y)$ *follows linear model described by Equations* (5.1) *and each column of* $\mathbf{X}$ *is normalized to* l_2*-norm* $\sqrt{n}$*. Assume that (5.4) holds.*

(a) Consider $\frac{1}{n}X(S)^T X(S) = I_{q\times q}$*. For any* j*, define*

$$c_{n,j}^2 = \frac{n\beta_j^{*2}}{\sigma^2}. \tag{5.12}$$

Define $c_n = \min_j c_{n,j}$. *Then, for sign consistency, it is necessary that* $c_n \to \infty$. *Specifically,*

$$P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \leq 1 - \frac{\exp\left\{-c_n^2/2\right\}}{\sqrt{2\pi}(1+c_n)}.$$

(b) If the Irrepresentable Condition (5.3) *does not hold, specifically,*

$$\left\| X(S^c)^T X(S) \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} \right\|_\infty \geq 1, \tag{5.13}$$

then, the Lasso estimate is not sign consistent: $P\left[\hat{\beta}(\lambda) =_s \beta^*\right] \leq \frac{1}{2}$.

Proof. First prove (b). Without loss of generality, assume for some $j \in S^c$, $X_j^T X(S) \left(X(S)^T X(S) \right)^{-1} \overrightarrow{b} = 1 + \zeta$, then $V_j = \lambda(1+\zeta) + \tilde{V}_j$, where $\tilde{V}_j = -[X(S) \left(X(S)^T X(S) \right)^{-1} X(S)^T - I]\frac{\epsilon}{n}$ is a Gaussian random variable with mean 0, so $P(\tilde{V}_j > 0) = \frac{1}{2}$. So, $P(V_j > \lambda) \geq \frac{1}{2}$, which implies that for any λ, Condition (5.5) (a necessary condition) is violated with probability greater than 1/2.

For claim (a). Condition (5.6),

$$sign\left(\beta^*(S) + (\frac{1}{n} X(S)^T X(S))^{-1} \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] \right) = sign(\beta^*(S))$$

is also a necessary condition for sign consistency. Since $\frac{1}{n} X(S)^T X(S) = I_{q \times q}$, (5.6) becomes

$$sign\left(\beta^*(S) + \left[\frac{1}{n} X(S)^T \epsilon - \lambda sign(\beta^*(S)) \right] \right) = sign(\beta^*(S)),$$

which implies that

$$sign\left(\beta^*(S) + \frac{1}{n} X(S)^T \epsilon \right) = sign(\beta^*(S)). \tag{5.14}$$

Without loss of generality, assume for some $j \in S$, $\beta_j^* > 0$. Then (5.14) implies $\beta_j^* + Z_j > 0$, where $Z_j = e_j^T \frac{1}{n} X(S)^T \epsilon$ is a Gaussian random variable with mean 0, and variance

$$\begin{aligned} var(Z_j) &= e_j^T \frac{1}{n} X(S)^T VAR(\epsilon) \frac{1}{n} X(S) e_j \\ &= \frac{\sigma^2 e_j^T \left[X(S)^T X(S) \right] e_j}{n^2} \\ &= \frac{\sigma^2}{n}, \end{aligned}$$

where the last equality uses the definition of $c_{n,j}^2$ in Theorem 2. To summarize,

$$\begin{aligned}
P[\hat{\beta}(\lambda) =_s \beta^*] &\leq P[\beta_j^* + Z_j > 0] \\
&= P[Z_j > -\beta_j^*] \\
&= P[Z_j < \beta_j^*] \\
&= 1 - \int_{\beta_j^*}^{\infty} \frac{1}{\sqrt{2\pi var(Z_j)}} \exp\{-\frac{x^2}{2var(Z_j)}\}dx \\
&= 1 - \int_{\beta_j^*/\sqrt{var(Z_j)}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}dx \\
&\leq 1 - \frac{1}{\sqrt{2\pi}} \int_{\beta_j^*/\sqrt{var(Z_j)}}^{\infty} (\frac{x}{1+x} + \frac{1}{(1+x)^2}) \exp\{-\frac{x^2}{2}\}dx \\
&= 1 - \frac{\exp\left\{-\frac{\beta_j^{*2}}{2var(Z_j)}\right\}}{\sqrt{2\pi}(1 + \frac{\beta_j^*}{\sqrt{var(Z_j)}})} \\
&= 1 - \frac{\exp\left\{-\frac{c_{n,j}^2}{2}\right\}}{\sqrt{2\pi}(1 + c_{n,j})}.
\end{aligned}$$

□

Statement (a) would hold for the homoscedastic model by removing $diag(|\mathbf{X}\beta^*|)$ from the denominator in Equation (5.12). Equation (5.12) can be viewed as a comparison of the signal strength (β_j^{*2}) to the noise level ($var(X_j^T\epsilon)$). Theorem 28 shows that the signal strength needs to be large relative to the noise level.

Statement (b) says that the Irrepresentable Condition (5.3) is necessary for the Lasso's sign consistency. This necessary condition can also be found in both Zhao and Yu (2006) and Wainwright (2009). Zhao and Yu (2006) points out that the Irrepresentable Condition is almost necessary and sufficient for the Lasso to be sign consistent under the standard homosedastic model when p and q are fixed. Wainwright (2009) says that it is necessary for the Lasso's sign consistency under the standard model for any p and q.

5.2 Piecewise Linear Solution

定理29 (Piecewise Linear Solution)**.** *The Lasso solution is piecewise linear when λ varies from ∞ to 0.*

Proof. It is sufficient to prove

$$\frac{d\ \hat{\beta}(\lambda)}{d\ \lambda} \text{ is piecewise constant.}$$

For every value of λ we have a set of "active" variables

$$\mathcal{A} := \{j : \hat{\beta}_j(\lambda) \neq 0\},$$

such that

$$X_{\mathcal{A}}^T(Y - X_{\mathcal{A}}\beta_{\mathcal{A}}) - \lambda sign(\beta_{\mathcal{A}}) = 0$$

$$|X_{\mathcal{A}^c}^T(Y - X_{\mathcal{A}^c}\beta_{\mathcal{A}^c})| \leq \lambda$$

So on this set $\mathcal{A}$, we have

$$\beta_{\mathcal{A}} = [X_{\mathcal{A}}^T X_{\mathcal{A}}]^{-1}(X_{\mathcal{A}}^T Y - \lambda sign(\beta_{\mathcal{A}}))$$

So when $\mathcal{A}$ does not change, $\beta_{\mathcal{A}}$ will change linearly with λ. □

5.3 The Lasso and path Algorithms

5.3.1 LARS

By the piece-wise linear properties, we can have the following algorithm: 1. Initialize :

$$\beta = 0, \mathcal{A} = \arg\max_j |X_j^T Y|, \gamma_{\mathcal{A}} = -sign(X_{\mathcal{A}}^T Y), \gamma_{\mathcal{A}^c} = 0$$

2. While$(X^T(Y - X\beta) \neq 0)$: (a)

$$d_1 = \min\{d > 0 : |[X_j^T(Y - X(\beta + d\gamma))]| = |[X_{\mathcal{A}}^T(Y - X(\beta + d\gamma))]|, j \in \mathcal{A}^c\}$$

$$d_2 = \min\{d > 0 : (\beta + d\gamma)_j = 0, j \in \mathcal{A}\}$$

Find step length: $d = min(d_1, d_2)$.

(b) Take step: $\beta = \beta + d\gamma$

(c) Update the active set: If $d = d_1$, then add variable(s) attaining equality at d to $\mathcal{A}$. If $d = d_2$, then remove variable(s) attaining 0 at d from $\mathcal{A}$.

(d) Calculate new direction:

$$\gamma_{\mathcal{A}} = (X_{\mathcal{A}}^T X_{\mathcal{A}})^{-1} sign(\beta_{\mathcal{A}}) \text{ and } \gamma_{\mathcal{A}^c} = 0$$

5.3.2 Coordinate Descent

Coordinate descent for the Lasso

$$\min \frac{1}{2n}\|Y - X\beta - \gamma\| + \lambda\|\beta\|$$

$\gamma_k = mean(Y - X\beta)$ $\beta_j = S(\frac{1}{n}\sum x_{ij}(y_i - \sum_{\ell \neq j} x_{i\ell}\beta_\ell, \lambda)$

Coordinate descent for L1 Logistic regression

$$Pr(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + x^T\beta)}}$$

log-likelihood:

$$\ell(\beta_0, \beta) = \sum y_i(\beta_0 + x_i^T\beta) - \log(1 + e^{\beta_0 + x_i^T\beta})$$

Iteratively reweighted least squares: Suppose the current estimate is $(\tilde{\beta}_0, \tilde{\beta})$, then a quadratic approximation to the log-likelihood is:

$$\ell_Q(\beta_0, \beta) = \sum w_i(z_i - \beta_o - x_i^T\beta)^2 + C,$$

where

$$z_i = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}_i}{\tilde{p}_i(1 - \tilde{p}_i)}$$

$$w_i = \tilde{p}_i(1 - \tilde{p}_i)$$

5.4 Homework

1. In LARS algorithm described in Section 5.3.1, for $j \in \mathcal{A}^c$, if $0 < d \leq d_1$, where

$$d_1 = \min\{d > 0 : |[X_j^T(Y - X(\beta + d\gamma))]| = |[X_{\mathcal{A}}^T(Y - X(\beta + d\gamma))]|\},$$

then $\left|[X_j^T(Y - X(\beta + d\gamma))]\right| \leq \lambda - d$, where $\lambda = |[X_j^T(Y - X(\beta))]|, j \in \mathcal{A}$

2. Describe the detailed LARS algorithm for a linear model with three predictors:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

   (1) Write the detailed LARS algorithm.

   (2) Do simulations. Generate X_j from $N(0, 1)$ for $j = 1, 2, 3$. Generate ϵ from $N(0, 0.04)$. Set $\beta_1 = 1$, $\beta_2 = 2$ and $\beta_3 = 0$. Plot the solution path obtained from (1).

(3) Compare your solution path with the solution path obtained from LARS package and write up your findings.

3. Show that a quadratic approximation of the following expression

$$\ell(\beta_0, \beta) = \sum y_i(\beta_0 + x_i^T\beta) - \log(1 + e^{\beta_0 + x_i^T\beta})$$

at $(\tilde{\beta}_0, \tilde{\beta})$, is:

$$\ell_Q(\beta_0, \beta) = \sum w_i(z_i - \beta_o - x_i^T\beta)^2 + C,$$

where

$$z_i = \tilde{\beta}_0 + x_i^T\tilde{\beta} + \frac{y_i - \tilde{p}_i}{\tilde{p}_i(1 - \tilde{p}_i)}$$

$$w_i = \tilde{p}_i(1 - \tilde{p}_i),$$

and C is a constant which does not depend on unknown parameters.

Chapter 6

Model Assessment and Selection

6.1 Generalization Errors

6.1.1 Continuous Response

We have a target variable Y, a vector of inputs X, and a prediction model $\hat{f}(X)$ that has been estimated from a training set $\mathcal{T}$. The loss function for measuring errors between Y and $\hat{f}(X)$ is denoted by $L(Y, \hat{f}(X))$. A typical choice of $L(\cdot, \cdot)$ is

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2.$$

Test error, also referred to as generalization error, is the prediction error over an independent test sample:

$$Err_{\mathcal{T}} = E(L(Y, \hat{f}(X))|\mathcal{T}),$$

where both X and Y are drawn randomly from their joint distribution.

A related quantity is

$$Err = E(L(Y, \hat{f}(X))),$$

which averages over everything that is random, including the randomness in the training set that produced $\hat{f}$.

Training error is the average loss over the training sample:

$$\bar{err} = \frac{1}{N} \sum_i L(y_i, \hat{f}(x_i)).$$

Note that training error is not a good estimate of the test error.

Example: Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\bar{x}_1, \bar{y}_1), \ldots, \bar{x}_M, \bar{y}_M)$. Let $L_{train} = \frac{1}{N} \sum_i E(y_i - x_i^T \hat{\beta})^2$ and $L_{test} = \frac{1}{M} \sum_i E(\bar{y}_i - \bar{x}_i^T \hat{\beta})^2$. Then

$$L_{train} \leq L_{test}.$$

6.1.2 Categorical Response

Typical loss functions:

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}(X))$$

$$L(G, \hat{p}(X)) = -\sum I(G = k) \log \hat{p}_k(X).$$

Test error is $Err_{mathcalT} = E(L(G, \hat{G}(X))|\mathcal{T})$.

6.1.3 Splits of Data

We have two separate goals: 1) Model slection: estimating the performance of different models in order to choose the best one.

2) Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data.

If we have enough data, the best approach is to randomly divide the dataset into three parts: a training set, a validation set and a test set. The training set is used to fit the models; the test set is used for assessment of the generalization error for model selection; the test set is used for assessment of the generalization error of the final chosen model.

6.2 Bias-Variance Tradeoff

6.2.1 Bias-Variance Decomposition

$$Y = f(X) + \epsilon,$$

where $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$. Using squared-error loss, we can derive the expected prediction error at an input point $X = x_0$,

$$\begin{aligned} Err(x_0) &= E((Y - \hat{f}(x_0))^2) \\ &= E(f(x_0) - \hat{f}(x_0) + \epsilon)^2 \\ &= \sigma^2 + (E\hat{f}(x_0) - f(x_0))^2 + E(\hat{f}(x_0) - E\hat{f}(x_0))^2 \\ &= \sigma^2 + Bias^2 + Variance \end{aligned}$$

6.2.2 Bias-Variance tradeoff

We show the bias-variance tradeoff via a simple example.

$$Y = X\beta^* + \epsilon,$$

$$X'X = I, E(\epsilon = 0), E(\sigma^2) = \sigma^2.$$

Then the solution of ridge regression is

$$\hat{\beta} = \frac{X'Y}{1+\lambda}.$$

For a new test point x_0, the prediction error is

$$Err(x_0) = E(y - x_0'\hat{\beta})^2 = \sigma^2 + Bias^2 + Variance,$$

where

$$\begin{aligned} Bias^2 &= (x_0'\beta^* - Ex_0'^T\hat{\beta})^2 \\ &= (x_0'\beta^* - x_0'^T\beta^*/(1+\lambda))^2 \\ &= (1 - \frac{1}{1+\lambda})^2 (x_0'^T\beta^*)^2 (\text{ increasing with } \lambda) \end{aligned}$$

$$\begin{aligned} Variance &= E(x_0'\hat{\beta} - Ex_0'^T\hat{\beta})^2 \\ &= \frac{\sigma^2}{(1+\lambda)^2} x_0'x_0 (\text{ decreasing with } \lambda) \end{aligned}$$

6.3 Cross Validation

6.3.1 K-fold Cross Validation

We split the data into K roughly equal-sized parts. For the kth part, we fit the model to the other $K-1$ parts of the data, and calculate the prediction error of the fitted model when predicting the kth part of the data. We do this for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction error.

Question: Is the CV error a good estimate of generalization error?

6.4 Bootrstrap

We have $b = 1, \ldots, B$ bootstrap datasets, from each dataset we have the model $f^b(x)$ which can be used to give prediction error on x_i, $L(y_i, \hat{f}^b(x_i))$.

$$\hat{Err}_{boot} = \frac{1}{B}\frac{1}{M} L(y_i, \hat{f}^b(x_i)).$$

How to overcome the correlations between different samples? 0.632 bootstrap.

$$\hat{Err}_{boot} \leftarrow 0.368\hat{err} + 0.632\hat{Err}_{boot},$$

where $\hat{err}$ is the training error.

6.5 Homework

Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\bar{x}_1, \bar{y}_1), \ldots, \bar{x}_M, \bar{y}_M)$. Let $L_{train} = \frac{1}{N}\sum_i E(y_i - x_i^T\hat{\beta})^2$ and $L_{test} = \frac{1}{M}\sum_i E(\bar{y}_i - \bar{x}_i^T\hat{\beta})^2$. Then

$$L_{train} \leq L_{test}.$$

Chapter 7

Gaussian Graphical Models

7.1 Gaussian Graphical Models

Graphical model is used to describe the relationship between variables. A graph is denoted by (E, V), where E is the node (variable) set and $V \subset E \times E$ is the edge set. If X_i is independent of X_j given all of the other variables, then there is no edges between node X_i and X_j.

定理30. *$X = (X_1, \ldots, X_p)$ follows a joint normal distribution $N(0, \Sigma)$ with $\Sigma > 0$. The following three properties are equivalent:*

1. *There is no edge between X_i and X_j in the Gaussian Graphical model;*

2. $$\Sigma^{-1}(i, j) = 0;$$

3. $E(X_j | X_{V \setminus j}) = \sum_{k \neq j} \beta_{jk} X_k$, $\beta_{ji} = 0$

Before proving this theorem, we first give a result about partitioned matrices. Consider a general partitioned matrix

$$M = \begin{pmatrix} E & F \\ G & H \end{pmatrix},$$

where we assume that both E and H are invertible. Then we have

定理31.

$$\begin{aligned} \begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} &= \begin{pmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{pmatrix} \\ &= \begin{pmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{pmatrix} \end{aligned}$$

where

$$M/E = H - GE^{-1}F \text{ and } M/H = E - FH^{-1}G$$

Proof.

$$\begin{pmatrix} I & 0 \\ -GE^{-1} & I \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} \begin{pmatrix} I & -E^{-1}F \\ 0 & I \end{pmatrix} = \begin{pmatrix} E & 0 \\ 0 & H - GE^{-1}F \end{pmatrix}$$

So,

$$\begin{pmatrix} E & F \\ G & H \end{pmatrix}^{-1} = \begin{pmatrix} I & -E^{-1}F \\ 0 & I \end{pmatrix} \begin{pmatrix} E^{-1} & 0 \\ 0 & [H - GE^{-1}F]^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -GE^{-1} & I \end{pmatrix}$$
$$= \begin{pmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{pmatrix}$$

By

$$\begin{pmatrix} I & -FH^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} E & F \\ G & H \end{pmatrix} \begin{pmatrix} I & 0 \\ -H^{-1}G & I \end{pmatrix} = \begin{pmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{pmatrix},$$

we have another equation. □

Now we prove Theorem 30.

Proof. We partition Σ the following way:

$$\Sigma = \begin{pmatrix} \sigma_{ii} & \Sigma_{iV_2} \\ \Sigma_{V_2 i} & \Sigma_{V_2 V_2} \end{pmatrix},$$

where $V_2 = V \setminus \{i\}$ and $V = \{1, 2, \ldots, p\}$. Correspondingly, we write

$$\Sigma^{-1} = \begin{pmatrix} d_{ii} & D_{iV_2} \\ D_{V_2 i} & D_{V_2 V_2} \end{pmatrix},$$

From Theorem 31, we have

$$D_{iV_2} = -d_{ii}\Sigma_{iV_2}[\Sigma_{V_2,V_2}]^{-1}.$$

Now we partition Σ_{V_2,V_2} in the following way:

$$\Sigma_{V_2,V_2} = \begin{pmatrix} \sigma_{jj} & \Sigma_{jB} \\ \Sigma_{Bj} & \Sigma_{BB} \end{pmatrix},$$

where $B = V \setminus \{i, j\}$. Correspondingly, we write

$$[\Sigma_{V_2,V_2}]^{-1} = \begin{pmatrix} \tilde{d}_{jj} & \tilde{D}_{jB} \\ \tilde{D}_{Bj} & \tilde{D}_{BB} \end{pmatrix},$$

From Theorem 31, we have

$$\tilde{D}_{Bj} = -\tilde{d}_{jj}[\Sigma_{B,B}]^{-1}\Sigma_{Bj}.$$

We have

$$\begin{aligned}
\Sigma^{-1}(i,j) &= D_{iV_2}[j] \\
&= -d_{ii}[\Sigma_{ij}\tilde{d}_{jj} + \Sigma_{iB}\tilde{D}_{Bj}] \\
&= -d_{ii}[\sigma_{ij}\tilde{d}_{jj} - \tilde{d}_{jj}\Sigma_{iB}[\Sigma_{B,B}]^{-1}\Sigma_{Bj}] \\
&= -d_{ii}\tilde{d}_{jj}[\sigma_{ij} - \Sigma_{iB}[\Sigma_{B,B}]^{-1}\Sigma_{Bj}]
\end{aligned}$$

Not that $X_A|X_B$ is a joint Gaussian random variable and

$$Var(X_A|X_B) = \Sigma_{A,A} - \Sigma_{A,B}\Sigma_{B,B}^{-1}\Sigma_{B,A}(homework)$$

$X_i \perp x_j|otheres$ if and only of $[cov(X_A|X_B)]_{i,j} = 0$ that is

$$\sigma_{ij} - \Sigma_{iB}[\Sigma_{BB}]^{-1}\Sigma_{B,j} = 0,$$

which is equivalent to $\Sigma^{-1}(i,j) = 0$.

$E(X_i|X_{V\setminus i}) = \sum_{k\neq i}\beta_{ik}X_k \iff$

$$\beta_{ik} = arg\min E(X_i - \sum_{k\neq i}\beta_{ik}X_k)^2$$

So

$$\beta_{iV_2} = \Sigma_{iV_2}(\Sigma_{V_2V_2})^{-1} = -D_{iV_2}/d_{ii}.$$

So we have $\beta_{ji} = 0$ is equivalent to $\Sigma^{-1}[i,j] = 0$.

□

7.2 Neighborhood selection

For each j, let

$$\theta^{j,\lambda} = \arg\min_{\theta} \|X_j - X\theta\|_2^2 + \lambda\|\theta\|_1$$

$$ne(j,\lambda) = \{j : \theta^{j,\lambda} \neq 0\}$$

7.3 L1-loglikelihood

likelihood

$$f(\Sigma) = \Pi_{i=1}^{n} \frac{1}{(2\pi)^{p/2}|\Sigma^{1/2}|} \exp\{-\frac{1}{2} x_i^T (\Sigma)^{-1} x_i\} \tag{7.1}$$

$$\ell(\Sigma) = \log(|\Sigma^{-1}|) - tr(\Sigma^{-1} S,)$$

where

$$S = \frac{1}{n} \sum x_i x_i^T$$

$$\hat{\Sigma}^{-1} = \arg\min_{C} - \log|C| + tr(CS) + \lambda \sum |C_{ij}|$$

7.4 Graphical Lasso

We want to solve the following optimization problem:

$$\hat{\Theta} = \min_{\Theta} - \log|\Theta| + tr(\Theta S) + \lambda \sum |\Theta_{ij}|.$$

Let $W = \hat{\Theta}^{-1}$ and partition W as follows:

$$W = \begin{pmatrix} W_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}.$$

We also partition S accordingly:

$$S = \begin{pmatrix} S_{11} & s_{12} \\ s_{21} & s_{22} \end{pmatrix}.$$

We solve W column by column using block coordinate descent.

引理6.

$$\frac{\partial}{\partial \Theta} - \log|\Theta| + tr(\Theta S) = -\Theta^{-1} + S.$$

Proof.

$$\frac{\partial}{\partial \Theta_{ij}} tr(\Theta S) = \frac{\partial}{\partial \Theta_{ij}} \sum_k \sum_l \Theta_{kl} S_{lk} = S_{ji}$$

$$\frac{\partial}{\partial \Theta_{ij}} \log |\Theta| = \frac{1}{|\Theta|} \frac{\partial}{\partial \Theta_{ij}} |\Theta|.$$

Note that

$$|\Theta| = \sum_j (-1)^{i+j} \Theta_{ij} M_{ij},$$

where M_{ij} is the determinant of the matrix obtained by removing the ith row and jth column of Θ. So

$$\frac{\partial}{\partial \Theta_{ij}} |\Theta| = (-1)^{i+j} M_{ij}.$$

Denote the adjugate matrix $adj(A)$ with the (i, j) element $(-1)^{i+j} M_{ji}$, we have $A \times adj(A) = |A|I$, equivalently $A^{-1} = \frac{adj(A)}{|A|}$

So

$$\frac{1}{|\Theta|} \frac{\partial}{\partial \Theta} |\Theta| = \frac{1}{|\Theta|} adj(\Theta)^T = \Theta^{-1}.$$

□

定理32. *W_{11} fixed, we have $w_{12} = W_{11}\hat{\beta}$, and $w_{22} = s_{22} + \lambda$ where*

$$\hat{\beta} = \min_{\beta} \frac{1}{2} \|W_{11}^{1/2} \beta - W_{11}^{-1/2} S_{12}\|_2^2 + \lambda \|\beta\|_1. \tag{7.2}$$

Proof. Set Subgradient of (7.1) to be zero, by Lemma 6 : $W - S - \lambda \Gamma = 0$, where $\Gamma = sub(\Theta)$. Since Θ_{ii} is always greater than 0, we have $\Gamma_{22} = 1$. So $w_{22} - s_{22} - \lambda = 0$ and we have $w_{22} = s_{22} + \lambda$. W_{12} satisfies

$$W_{12} - S_{12} - \lambda \gamma_{12} = 0. \tag{7.3}$$

By setting subgradient of (7.2), we have

$$W_{11}^{1/2}(W_{11}^{1/2} \beta - W_{11}^{-1/2} S_{12}) - \lambda sign(\beta) = 0$$

β satisfies

$$W_{11}\beta - S_{12} + \lambda s = 0 \tag{7.4}$$

If $(\hat{\beta}, s)$ solves (7.4), then $W_{12} = W_{11}\hat{\beta}$ and $\gamma_{12} = -s$ solves (7.4). To show this, we only need to verify $-s$ is a subgradient of Θ_{12}. Expending

$$W\Theta = I,$$

we have

$$\begin{pmatrix} W_{11} & W_{12} \\ W_{21} & w_{22} \end{pmatrix} \begin{pmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \theta_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

$$W_{11}\Theta_{12} + w_{12}\theta_{22} = 0,$$

so

$$\Theta_{12} = -W_{11}^{-1} w_{12} \theta_{22},$$

$$sign(\hat{\Theta}_{12}) = -sign(\hat{\beta}).$$

So $-s$ is a subgradient of Θ.

□

while not converged **do**

Start with $W = S + \lambda I$. The diagonal elements of W will not change any more.

For each $j = 1, \ldots, p$ solve the Lasso problem (7.4) and obtain $\hat{\beta}$. Fill in the corresponding row and column of W with $w_{12} = W_{11}\hat{\beta}$

end while

7.4.1 Update of Θ

By

$$W_{11}\theta_{12} + w_{12}\theta_{22} = 0$$

and

$$w_{21}\theta_{12} + w_{22}\theta_{22} = 1$$

We have

$$\theta_{12} = -W_{11}^{-1} w_{12} \theta_{22} = -\hat{\beta}\theta_{22}$$

$$\theta_{22} = 1/[w_{22} - w_{21} W_{11}^{-1} w_{12}] = 1/[w_{22} - w_{21}\hat{\beta}]$$

Chapter 8

Dictionary Learning

8.1 Dictionary Learning

Consider a signal $x \in \mathcal{R}^n$ and a fixed dictionary $D = [d_1, \ldots, d_k] \in \mathcal{R}^{n \times k}$ (allowing $k > n$, making the dictionary overcomplete.) In this setting, sparse coding with an ℓ_1 regularization amounts to computing

$$\hat{\alpha} = \arg \min_{\alpha \in \mathcal{R}^k} \|x - D\alpha\| + \lambda_1 \|\alpha\|_1.$$

Now we consider a supervised setting, where the signal may belong to nay of p different classes. We model the signal x using a single shared dictionary D and a set of p decision functions $g_i(x, \alpha, \theta), i = 1, \ldots, p$ acting on x and its sparse code α over D. A possible choice of g_i is $g_i(x, \alpha, \theta) = w_i^T \alpha + b_i$.

Recall Multi-logistic regression:

$$P(Y_i = k|x_i) = \frac{\exp(g_i(x, \alpha, \theta))}{\sum_{j=1}^K \exp(g_j(x, \alpha, \theta))} = \frac{1}{\sum_{j=1}^K \exp\left[g_j(x, \alpha, \theta) - g_i(x, \alpha, \theta)\right]}$$

Let us assume that we are given p sets of training data $T_i, i = 1, 2, \ldots, p$ such that all samples in T_i belongs to class i. A direct method for learning unknown parameters is via maximum likelihood:

$$\max \sum_i \log P(y_i|X_i) = \max \sum_{i=1}^p \sum_{j \in T_i} -\log \sum_k \exp\left[g_k(x_j, \alpha, \theta) - g_i(x_j, \alpha, \theta)\right],$$

equivalently,

$$\min \sum_{i=1}^p \sum_{j \in T_i} \log \sum_k \exp\left[g_k(x_j, \alpha, \theta) - g_i(x_j, \alpha, \theta)\right],$$

where α is the sparse representation of x. A joint optimization problem – supervised dictionary learning is defined as

$$\min \sum_{i=1}^p \sum_{j \in T_i} \log \sum_k \exp\left[g_k(x_j, \alpha, \theta) - g_i(x_j, \alpha, \theta)\right] + \lambda_0 \|x_j - D\alpha_j\|_2^2 + \lambda_1 \|\alpha_j\|_1 + \lambda_2 \|\theta\|_2^2, \quad (8.1)$$

such that

$$\|D_j\|_2 \leq 1.$$

Define

$$S_i(\alpha, x, D, \theta) = \log \sum_k \exp\left[g_k(x_j, \alpha, \theta) - g_i(x_j, \alpha, \theta)\right] + \lambda_0 \|x_j - D\alpha_j\|_2^2 + \lambda_1 \|\alpha_j\|_1$$

. For observation x, it should be classifies into class i^* defined as:

$$i^*(\alpha, x, D, \theta) = \arg\min_k S_k(\alpha, x, D, \theta)$$

Equation (8.1) can be written as

$$\min \sum_{i=1}^{p} \sum_{j \in T_i} S_i(\alpha_j, x_j, D, \theta) + \lambda_2 \|\theta\|_2^2,$$

such that

$$\|D_j\|_2 \leq 1.$$

Note that we not only hope $S_i(x)$ to be small, we also hope that $S_i(x)$ is smaller than $S_j(x)$ for all j, if $x \in T_i$. Softmax function satisfies this purpose. The softmax function is defined as:

$$C_i(x_1, x_2, \ldots, x_p) = -\log \frac{e_i^x}{\sum_j x_j} = \log(\sum_j e^{x_j - x_i}).$$

So we can have another objective function:

$$\min \sum_{i=1}^{p} \sum_{j \in T_i} C_i(S_l(\alpha_j, x_j, D, \theta)_{l=1}^{p}) + \lambda_2 \|\theta\|_2^2,$$

such that

$$\|D_j\|_2 \leq 1.$$

A model combines the above two models is given by

$$\min \sum_{i=1}^{p} \sum_{j \in T_i} \mu C_i(S_l(\alpha_j, x_j, D, \theta)_{l=1}^{p}) + (1-\mu) S_i(\alpha_j, x_j, D, \theta) + \lambda_2 \|\theta\|_2^2,$$

such that

$$\|D_j\|_2 \leq 1.$$

8.2 Optimization Procedure

Algorithm 4 Sparse Dictionary Learnng

Input: p (number of classes); n (signal dimensions); $T_i, i = 1, \ldots, p$ (training signals); k (size of the dictionary); $\lambda_0, \lambda_1, \lambda_2$; μ

Output: $D \in \mathcal{R}^{n \times k}; \theta$

1: Initialization: Set D to be a random Gaussian matrix. Set θ to be zero.

2: Repeat until convergence

3: Supervised sparse coding: For all $l = 1, \ldots, p$, all $j \in T_i$, compute

$$\hat{\alpha}_{jl} = \arg\min_{\alpha} S_l(\alpha, x_j, D, \theta) \tag{8.2}$$

4: Dictionary update. Solve, under constraint $\|d_l\|_2 \leq 1$,

$$\min \sum_{i=1}^{p} \sum_{j \in T_i} \mu C_i(S_l(\hat{\alpha}_{lj}, x_j, D, \theta)_{l=1}^{p}) + (1-\mu) S_i(\hat{\alpha}_{ji}, x_j, D, \theta) + \lambda_2 \|\theta\|_2^2, \tag{8.3}$$

8.2.1 Supervised Sparse Coding

Now we solve Equation (8.2). Recall the definition of S_i.

$$\begin{aligned} S_i(\alpha, x, D, \theta) &= \log \sum_k \exp\left[g_k(x_j, \alpha_{ij}, \theta) - g_i(x_j, \alpha, \theta)\right] + \lambda_0 \|x_j - D\alpha_j\|_2^2 + \lambda_1 \|\alpha_j\|_1 \\ &= c_i(A^T \alpha_{ij} + b) + \lambda_0 \|x_j - D\alpha_j\|_2^2 + \lambda_1 \|\alpha_j\|_1 \end{aligned}$$

When D and θ are fixed, this is a Lasso problem. Coordinate descent can be used to solve this problem.

8.2.2 Dictionary Update

A local minimum can be obtained using projected gradient descent by taking partial derivatives to be zeros with respect to D and θ.

8.3 Application

Digits recognition

Chapter 9

Sparse PCA

9.1 PCA

PCA is an unsupervised method, which is widely used in data processing and dimensionality reduction. However, PCA suffers from the fact that each principal component is a linear combination of all the original variables, thus it is often difficult to interpret the results. We show that PCA can be formulated as a regression type optimization problem and thus the sparse loadings can be obtained via the Lasso.

PCA seeks the linear combination of the original variables such that the derived variables capture the maximal variance. PCA can be done via the SVD of the data matrix. In detail, let the data X be an $n \times p$ matrix, where n and p are the number of observations and the number of variables, respectively. Assume that the column means of X are all 0. Suppose we have the SVD of X,

$$\frac{1}{\sqrt{n}}X = UDV^T.$$

Then $UD = [U_1D_{11}, \ldots, U_pD_{pp}]$ are the principle components and the columns of V are the corresponding loadings of the principle components. The variance of the ith PC is D_{ii}^2. Usually the first $q(\ll p)$ components are used to represent the data and so dimensionality reduction is achieved.

Some comments to the above paragraph: The first component seeks the combination of X, such that it has the maximum variance:

$$\beta^{(1)} = \arg\max_{\beta} var(X\beta) \approx \arg\max_{\beta} \frac{1}{n}\beta X'X\beta$$

So $\beta^{(1)}$ is the first eigen vector of

$$\frac{1}{n}X'X = VD^2V^T.$$

$\beta^{(1)} = V[:, 1]$. The loadings of the other PCs are the columns of V. That is $\beta^{(i)} = V[:, i]$.

Note that

$$\frac{1}{\sqrt{n}}XV = UD,$$

so

$$\frac{1}{\sqrt{n}}XV[:, i] = U[:, i]D_{ii}$$

$$Var(XV[:, i]) = D_{ii}^2.$$

9.2 Direct Sparse Approximations

We first discuss a simple regression approach to PCA.

定理33. *Let*

$$X = UDV^T.$$

For all i, denote $Y_i = U_i D_{ii}$. Y_i is the ith PC. $\forall\ \lambda > 0$, suppose $\hat{\beta}_{ridge}$ is the ridge estimate given by

$$\hat{\beta}_{ridge} = \arg\min_{\beta} |Y_i - X\beta| + \lambda\|\beta\|_2^2.$$

Let $\hat{v} = \frac{\hat{\beta}_{ridge}}{\|\hat{\beta}_{ridge}\|_2}$, then

$$\hat{v} = V_i.$$

Proof.

$$\begin{aligned}
\hat{\beta}_{ridge} &= (X'X + \lambda I)^{-1} X' U_i D_{ii} \\
&= V[D^2 + \lambda I]^{-1} V^T X' X V_i \\
&= V[D^2 + \lambda I]^{-1} V^T V D^2 V^T V_i \\
&= V[\frac{D^2}{D^2 + \lambda I}] V^T V_i \\
&= V_i \frac{D_{ii}^2}{D_{ii}^2 + \lambda}
\end{aligned}$$

□

Now let us consider the following Elastic net problem by adding ℓ_1 penalty to the above regression problem, we have

$$\hat{\beta} = \arg\min_{\beta} |Y_i - X\beta| + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2.$$

We call

$$V_i = \frac{\hat{\beta}}{\|\hat{\beta}\|_2},$$

an approximation to V_i and XV_i ith approximate PC.

Based on the above results, we can have a two stage sparse PCA algorithm: (1) perform PCA (2) obtain the sparse approximation. We will give another "self-contained" Sparse PCA.

9.3 Self-contained Sparse PCA

定理34. *Let x_i^T denote the ith row vector of the matrix X. Let α and β are both $p \times 1$ matrices. For any $\lambda > 0$, let*

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_i \|x_i - \alpha\beta^T x_i\|_2^2 + \lambda\|\beta\|_2^2,$$

such that $\|\alpha\|_2^2 = 1$. Then $\hat{\beta} \propto V_1$.

Proof.

$$\begin{aligned}
\sum_i \|x_i - \alpha\beta^T x_i\|_2^2 + \lambda\|\beta\|_2^2 &= \sum_i x_i^T (I - \alpha\beta^T)^T (I - \alpha\beta^T) x_i \\
&= tr((I - \alpha\beta^T)^T (I - \alpha\beta^T) \sum (x_i x_i^T) \\
&= tr(I - \beta\alpha^T - \alpha\beta^T + \beta\alpha^T\alpha\beta^T) X^T X \\
&= tr(X'X) + tr(\beta^T X^T X \beta) - 2tr(\alpha^T X^T X \beta)
\end{aligned}$$

For a fixed α, the solution of $\hat{\beta}$ is

$$\hat{\beta} = (X^T X + \lambda)^{-1} X^T X \alpha$$

$$\hat{\alpha} = \arg\max_{\alpha} \alpha^T (X^T X)(X^T X + \lambda I)^{-1} X^T X \alpha$$

such that $\alpha^T\alpha = 1$. By $X = UDV^T$, we have

$$(X^T X)(X^T X + \lambda I)^{-1} X^T X \alpha = V[\frac{D^4}{D^2 + \lambda}]V^T,$$

so $\hat{\alpha} = sV_1$ and $\hat{\beta} = s\frac{D_{11}^2}{D_{11}^2 + \lambda} V_1$, where s can be 1 or -1. □

定理35. *Suppose we are considering the first k PCs. Let x_i denote the ith row vector of the matrix X. Let α and β are both $p \times k$ matrices. For any $\lambda > 0$, let*

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_i \|x_i^T - \alpha^T\beta x_i^T\|_2^2 + \lambda\|\beta\|_2^2,$$

such that $\|\alpha\|_2^2 = I_k$. Then $\hat{\beta}_j \propto V_j$, for $j = 1, 2, \ldots k$.

Proof.

$$\begin{aligned}
\sum_i \|x_i - \alpha\beta^T x_i\|_2^2 + \lambda\|\beta\|_2^2 &= \sum_i x_i^T (I - \alpha\beta^T)^T (I - \alpha\beta^T) x_i \\
&= tr((I - \alpha\beta^T)^T (I - \alpha\beta^T) \sum (x_i x_i^T) \\
&= tr(I - \beta\alpha^T - \alpha\beta^T + \beta\alpha^T\alpha\beta^T) X^T X \\
&= tr(X'X) + tr(\beta^T X^T X \beta) - 2tr(\alpha^T X^T X \beta) \\
&= tr(X'X) + \sum_j tr(\beta_j^T X^T X \beta_j) - 2tr(\alpha_j^T X^T X \beta_j)
\end{aligned}$$

For a fixed α, the solution of $\hat{\beta}$ is

$$\hat{\beta}_j = (X^T X + \lambda)^{-1} X^T X \alpha_j,$$

or equivalently

$$\hat{\beta} = (X^T X + \lambda)^{-1} X^T X \alpha.$$

$$\hat{\alpha} = \arg\max_{\alpha} \alpha^T (X^T X)(X^T X + \lambda I)^{-1} X^T X \alpha$$

such that $\alpha^T \alpha = I_k$. By $X = UDV^T$, we have

$$(X^T X)(X^T X + \lambda I)^{-1} X^T X \alpha = V[\frac{D^4}{D^2 + \lambda}]V^T,$$

so $\hat{\alpha} = sV[:, 1:k]$ and $\hat{\beta}_j = s_j \frac{D_{jj}^2}{D_{jj}^2 + \lambda} V_j$.

□

So a Sparse PCA can be obtained via:

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha,\beta} \sum_i \|x_i^T - \alpha^T \beta x_i^T\|_2^2 + \lambda\|\beta\|_2^2 + \sum_j \lambda_{1j}\|\beta_j\|_1.$$

9.4 Numerical Solution

For a fixed α, β is obtained via the following elastic net:

$$\min_{\beta_j} tr(X'X) + tr(\beta_j^T X^T X \beta_j) - 2tr(\alpha_j^T X^T X \beta_j) + \lambda_{1j}\|\beta_j\|_1$$

For fixed β, α is obtained via

$$\max tr(\alpha^T(X^TX)\beta),$$

such that

$$\alpha^T\alpha = I.$$

定理36. *Let A and B are two $m \times k$ matrices and B has rank k. Consider the following optimization problem:*

$$\hat{A} = \arg\max_A tr(A^TB), \text{ such that } A^TA = I_k.$$

Suppose the SVD of B is $B = UDV^T$, then $\hat{A} = UV^T$.

Proof. The constrain $A^TA = I_k$ is equivalent to $\frac{k(k+1)}{2}$ constraints:

$$A_i^TA_i = 1$$

$$A_i^TA_j = 0, i > j.$$

Using Lagrangian multiplier method, we define

$$L = \sum_i B_i^TA_i + \sum_i \frac{1}{2}\lambda_{ii}(A_i^TA_i - 1) + \sum_{i>j}\lambda_{ij}(A_i^TA_j).$$

Setting $\frac{\partial L}{\partial A_j} = 0$, we have

$$B_i = \lambda_{ii}\alpha_i + \sum_{j>i}\lambda_{ij}\alpha_j = 0.$$

So

$$B = A\Lambda.$$

$$A = B\Lambda^{-1}.$$

$$tr(A^TB) = tr\Lambda^{-T}B^TB = tr\Lambda^{-1}VD^2V^T = trV^T\Lambda^{-1}VD^2.$$

Note that

$$A^TA = \Lambda^{-T}B^TB\Lambda^{-1} = \Lambda^{-T}VD^2V^T\Lambda^{-1} = I_k$$

Let $C := V^T\Lambda^{-1}V$. Then

$$A^TA = C^TD^2C = I \Longrightarrow \sum C_{ij}^2D_{ii}^2 = 1.$$

$$tr(AB) = tr(CD^2) = \sum_i C_{ii} D_{ii}^2 \leq \sum_{ii} D_{ii}.$$

$$C_{ii} = 1/D_{ii} \text{ and } C_{ij} = 0.$$

So

$$\Lambda^{-1} = VCV^T = VD^{-1}V^T.$$

$$A = B\Lambda^{-1} = UDV^T VD^{-1}V^T = UV^T.$$

□

Chapter 10

Boosting

Boosting is a general method for improving the accuracy of any given learning algorithm. In this chapter we first introduce the boosting algorithm AdaBoost, and explains the underlying theory of boosting. Then we introduce the L2-boosting algorithm and a statistical view of boosting.

10.1 Adaboost

The adaboost is introduced first by Freund and Schapire, 1995. It takes as input a training set $(x_1, y_1), \ldots, (x_n, y_n)$, where each x_i belongs to some domain or instance space, and each label y_i is in some label set Y. Let's first assume $y_i \in \{-1, 1\}$. We define $F(x) = \sum_{m=1}^{M} c_m f_m(x)$ where each $f_m(x)$ is a classifier producing values 1 or -1 and c_m are constants; the corresponding prediction is $sign(F(x))$. The AdaBoost procedure trains the classifier $f_m(x)$ on weighted versions of the training sample, giving higher weight to cases that are currently misclassified. This is done for a sequence of weighted samples, and then the final classifier is defined to be a linear combination of the classifiers from each stage. The adaboost algorithm is described in Algorithm 5.

Algorithm 5 Adaboost Procedure

Input: $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathcal{R}^p, y_i \in \{-1, 1\}$.

1: Initialization: $w_i = 1/n$, for $i = 1, \ldots, n$.
2: **for** $m = 1, \ldots, M$ **do**
3: Estimate the classifier $f_m(x)$ from the training data with weights w_i.
4: Compute the error:

$$e_m = Pr_{i \sim w}[h_t(x_i) \neq y_i] = \sum_i w_i 1_{y_i \neq f_m(x_i)}.$$

5: Choose $c_m = \frac{1}{2} \log\left(\frac{1-e_m}{e_m}\right)$
6: Update:

$$w_i = \frac{w_i \exp(-c_m y_i f_m(x_i))}{Z_m}$$

where Z_m is a normalization factor,such that $\sum_i w_i = 1$.
7: **end for**
8: Output the final classifier:

$$sign(\sum_{m=1}^{M} c_m f_m(x)).$$

For a not bad "weak learner", $e_m \leq \frac{1}{2}$, so $c_m \geq 0$. So Step 6 increases the weight for misclassified samples.

Algorithm 6 Confidence Rated Adaboost Procedure

Input: $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathcal{R}^p, y_i \in \{-1, 1\}$.

1: Initialization: $w_i = 1/n$, for $i = 1, \ldots, n$.
2: **for** $m = 1, \ldots, M$ **do**
3: Estimate the "confidence rated" classifier $f_m(x)$ from the training data with weights w_i.
4: Compute the error:

$$e_m = Pr_{i\sim w}[h_t(x_i) \neq y_i] = \sum_i w_i 1_{y_i \neq f_m(x_i)}.$$

5: Choose $c_m = \frac{1}{2}\log\left(\frac{1-e_m}{e_m}\right)$
6: Update:

$$w_i = \frac{w_i \exp(-c_m y_i f_m(x_i))}{Z_m}$$

where Z_m is a normalization factor,such that $\sum_i w_i = 1$.
7: **end for**
8: Output the final classifier:

$$sign(\sum_{m=1}^{M} c_m f_m(x)).$$

10.2 Boosting — a Statistical View

In this section we show that the boosting algorithms are stage-wise estimation procedures for fitting an additive logistic regression model.

定理37. *The AdaBoost algorithm produces adaptive Newton updates for minimizing* $E(e^{-yF(x)})$.

Proof. Let $J(F) = E(e^{-yF(x)})$. Suppose we have a current estimate $F(x)$ and seek an improved estimate $F(x) + cf(x)$.

$$\begin{aligned}
\hat{c} &= \arg\min_c E(e^{-y[F(x)+cf(x)]}) \\
&= \arg\min_c E_w e^{-ycf(x)} \\
&= \frac{1}{2}\log\frac{e_m}{1-e_m}.
\end{aligned}$$

where $w = e^{-yF(x)}/Ee^{-yF(x)}$ and $e_m = E_w[1_{y\neq f(x)}]$.

$$\begin{aligned}
E_w e^{-ycf(x)} &= E_w[1_{y\neq f(x)}]e^c + E_w[1_{y=f(x)}]e^{-c} \\
&= e_m e^{-c} + (1-e_m)e^c.
\end{aligned}$$

By letting the derivative to be zero, we have $c = \frac{1}{2}\log\frac{e_m}{1-e_m}$.

In the next iteration, the new weight is proportional to $e^{-y[F(x)+\hat{c}f(x)]} = w \times e^{\hat{c}yf(x)}$ followed

by a normalization. □

10.3 Using the log-likelihood criteria

Let $y^* = (y+1)/2$ and

$$p(x) := P(y^* = 1|x) = \frac{e^{F(x)}}{e^{F(x)} + e^{-F(x)}}.$$

$$P(y^*|x) = \frac{e^{2y^* F(x)}}{1 + e^{2F(x)}}.$$

Then expected log-likelihood is

$$\ell(F) = E[2y^* F(x) - \log(1 + \exp(2F(x))].$$

Let

$$\hat{f} = \arg\min_f \ell(F+f) = \arg\min_f E[2y^*(F(x)+f(x)] - \log(1 + \exp(2F(x) + 2f(x))]$$

Note that

$$\ell(F+f) = \ell(F) + \ell'(F)f + \frac{1}{2}\ell''(F),$$

where

$$\ell'(F) = 2E(y^* - p(x)),$$

and

$$\ell''(F) = -4p(x)(1-p(x)).$$

So, the Newton update for f(x) is

$$\hat{f}(x) = -\frac{\ell'(F)}{\ell''(F)} = \frac{1}{2}\frac{E(y^* - p(x))}{p(x)(1-p(x))} = \frac{1}{2}E_w\frac{(y^* - p(x))}{p(x)(1-p(x))},$$

where $w = p(1-p)$ equivalently

$$\hat{f}(x) = \arg\min_f E_w(\frac{(y^* - p(x))}{p(x)(1-p(x))} - f(x))^2.$$

Finally we have the logitBoos algorithm as follows:

Algorithm 7 LogitBoost Procedure

Input: $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathcal{R}^p, y_i \in \{-1, 1\}$.

1: Initialization: $w_i = 1/n$, for $i = 1, \ldots, n$. $F = 0$ and probability estimates $p_i = 1/2$.
2: **for** $m = 1, \ldots, M$ **do**
3: Compute the working response and weights

$$z_i = \frac{y_i^* - p_i}{p_i(1 - p_i)}$$

$$w_i = p_i(1 - p_i).$$

4: Estimate $f_m(x)$ by weighted least-squares fitting of z to x.
5: Update:

$$F(x) = F(x) + \frac{1}{2} f_m(x) \text{ and } p(x)$$

6: **end for**
7: Output the final classifier:

$$sign(\sum_{m=1}^{M} c_m f_m(x)).$$

10.4 Boosting with the L2-Loss

From the above sections, we can see that the task of the Boosting algorithm is to estimate a function $F : \mathcal{R}^p \to \mathcal{R}$, minimizing an expected cost:

$$E(\ell(y, F)).$$

The most prominent examples for $\ell(\cdot, \cdot)$ are:

$\ell(y, F) = e^{-yF}$ with $y \in \{-1, 1\}$; AdaBoost

$\ell(y, F) = \log\{\exp(2yF)/(1 + \exp(2F))\}$ with $y \in \{0, 1\}$; LogitBoost

$\ell(y, F) = (y - F)^2/2$ with $y \in \mathcal{R}$ or $y \in \{-1, 1\}$; L2-boost.

For L2Boost, the update of f after $F(X)$ is known is

$$\hat{f} = \arg\min_f E(Y - F(X) - f|X)^2/2 = \arg\min_f E(Res - f|X).$$

So the L2Boost procedure can be described as follows:

L2boosting is nothing else than repeated least squares fitting of residuals. Generally speaking, we hope the weak learner to be simple.

Algorithm 8 L2Boost Procedure

Input: $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathcal{R}^p, y_i \in \{-1, 1\}$.

1: Initialization: $w_i = 1/n$, for $i = 1, \ldots, n$. $F = 0$.
2: **for** $m = 1, \ldots, M$ **do**
3: Estimate $f_m(x)$ by least-squares fitting of $Y - F(x)$ to x.
4: Update:

$$F(x) = F(x) + f_m(x)$$

5: **end for**
6: Output the final learner:

$$\sum_{m=1}^{M} f_m(x).$$

10.5 Path following algorithms using ϵ-Boosting

10.5.1 Gradient Descent View of Boosting

Given data $Z_i = (Y_i, X_i), i = 1, \ldots, n$. We want to learn a model in the following family:

$$\mathcal{F} = \{F : F(x) = \sum_j \beta_j h_j(x).\}$$

To find an estimate of β, we set up an empirical minimization problem:

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} L(Z_i, \beta),$$

where L is the loss function. Boosting is a progressive procedure that iteratively builds up the solution:

$$(\hat{j}, \hat{g}) = \arg\min_{j,g} \sum_i L(Z_i, \hat{\beta}^t + g1_j)$$

$$\hat{\beta}^{t+1} = \hat{\beta}^t + \hat{g}1_{\hat{j}}$$

$\epsilon-$Boosting has the similar procedure, but using a fixed step instead:

$$(\hat{j}, \hat{g}) = \arg\min_{j, s=\pm\epsilon} \sum_i L(Z_i, \hat{\beta}^t + s1_j)$$

$$\hat{\beta}^{t+1} = \hat{\beta}^t + \hat{s}1_{\hat{j}}$$

10.5.2 General Lasso

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \Gamma(\beta, \lambda) = \sum_i L(Z_i, \beta) + \lambda\|\beta\|_1$$

10.5.3 The Boosting Lasso Algorithm

Algorithm 9 Boosting Lasso Procedure

1: Step 1 (Initialization). Given Data $Z_i = (Y_i, X_i), i = 1, \ldots, n$ and a small stepsize constant $\epsilon > 0$ and a small tolerance parameter $\xi > 0$, take and initial forward step:

$$(\hat{j}, \hat{s}) = \arg \min_{j, s=\pm\epsilon} \sum_i L(Z_i, s1_j)$$

$$\hat{\beta}^0 = \hat{s}1_{\hat{j}}$$

Then calculate the initial regularization parameter

$$\lambda^0 = \frac{1}{\epsilon}(\sum_i L(Z_i, 0) - \sum_i L(Z_i, \hat{\beta}^0))$$

Set the active set index set $I_A^0 = \{\hat{j}\}$. Set $t = 0$.

2: Step 2 (Backward and Forward steps). Find the "backward" step that leads to the minimal empirical loss:

$$\hat{j} = \arg\min \sum_i L(Z_i, \hat{\beta}^t + s_j 1_j), \text{ where } s_j = -sign(\hat{\beta}^t)\epsilon$$

Take the step if it leads to a decrease of moderate size in the Lasso Loss, otherwise force a forward step and relax λ if necessary: If $\Gamma(\hat{\beta}^t + \hat{s}_{\hat{j}}1_j, \lambda^t) - \Gamma(\hat{\beta}^t, \lambda^t) \leq -\xi$, then

$$\hat{\beta}^{t+1} = \hat{\beta}^t + \hat{s}_{\hat{j}}1_j, \lambda^{t+1} = \lambda^t.$$

Otherwise,

$$(\hat{j}, \hat{s}) = \arg \min_{j, s=\pm\epsilon} \sum_i L(Z_i, \hat{\beta}^t + s1_j)$$

$$\lambda^{t+1} = \min\{\lambda^t, \frac{1}{\epsilon}(\sum_i L(Z_i, \hat{\beta}(t)) - \sum_i L(Z_i, \hat{\beta}^{t+1}) - \xi)$$

$$I_A^{t+1} = I_A^t \cup \{\hat{j}\}$$

3: Step 3 (Iteration). Increase t by one and repeat Step 2 and 3. Step when $\lambda^t \leq 0$.

引理7. *1. For any $\lambda \geq 0$, if there exists j and s with $|s| = \epsilon$ such that $\Gamma(s1_j, \lambda) \leq \Gamma(0, \lambda)$, we have $\lambda^0 \geq \lambda$.*

2. For any t, we have $\Gamma(\hat{\beta}^{t+1}, \lambda^{t+1}) \leq \Gamma(\hat{\beta}^t, \lambda^{t+1}) - \xi$.

3. For $\xi \geq 0$ and any t such that $\lambda^{t+1} < \lambda^t$, we have $\Gamma(\hat{\beta}^t \pm \epsilon 1_j, \lambda^t) > \Gamma(\hat{\beta}^t, \lambda^t) - \xi$ for every j and $\|\hat{\beta}^{t+1}\|_1 = \|\hat{\beta}^t\|_1 + \epsilon$.

Remark: Lemma (7) (1) guarantees that it is safe for BLasso to start with an initial λ^0 which is the largest λ such that would allow an ϵ step away from 0. Lemma (7) (2) says that for each value of λ, BLasso performs coordinate descent until there is no descent step. Then, by Lemma (7) (3), the value of λ is reduced and a forward step is forced.

Proof. 1. If there exists λ and j with $|s| = \epsilon$ such that

$$\Gamma(s1_j; \lambda) \leq \Gamma(0; \lambda),$$

then we have

$$\sum_i L(Z_i; s1_j) + \lambda\epsilon \leq \sum_i L(Z_i; 0).$$

Therefore

$$\begin{aligned}
\lambda &\leq \frac{1}{\epsilon}\{\sum_i L(Z_i; 0) - \sum_i L(Z_i, s1_j)\} \\
&\leq \frac{1}{\epsilon}\{\sum_i L(Z_i; 0) - \min_{j,|s|=\epsilon} \sum_i L(Z_i, s1_j)\} \\
&= \frac{1}{\epsilon}\{\sum_i L(Z_i; 0) - \sum_i L(Z_i, \hat{\beta}^0)\} \\
&= \lambda_0
\end{aligned}$$

2. Since a backward step is only taken when $\Gamma(\hat{\beta}^{t+1}; \lambda(t)) < \Gamma(\hat{\beta}^t - \xi)$ and $\lambda^{t+1} = \lambda^t$, we only need to consider forward steps. When a forward step is forced, if $\Gamma(\hat{\beta}^{t+1}; \lambda^{t+1}) > \Gamma(\hat{\beta}^t; \lambda^{t+1}) - \xi$, then

$$\sum_i L(Z_i, \hat{\beta}^{t+1}) + \lambda^{t+1}\|\hat{\beta}^{t+1}\|_1 > \sum_i L(Z_i, \hat{\beta}^t) + \lambda^{t+1}\|\hat{\beta}^t\|_1 - \xi$$

equivalently,

$$\lambda^{t+1}\|\hat{\beta}^t\|_1 - \lambda^{t+1}\|\hat{\beta}^{t+1}\|_1 > \sum_i L(Z_i, \hat{\beta}^t) - \sum_i L(Z_i, \hat{\beta}^{t+1}) - \xi$$

$$\lambda^{t+1}\epsilon > \sum_i L(Z_i, \hat{\beta}^t) - \sum_i L(Z_i, \hat{\beta}^{t+1}) - \xi$$

$$\lambda^{t+1} > \frac{1}{\epsilon}(\sum_i L(Z_i, \hat{\beta}^t) - \sum_i L(Z_i, \hat{\beta}^{t+1}) - \xi)$$

which contradicts the algorithm.

3. Since $\lambda^{t+1} < \lambda^t$ and λ cannot be relaxed by a backward step, we immediately have $\|\hat{\beta}^{t+1}\|_1 = \|\hat{\beta}^t\| + \epsilon$ (or else β can be obtained via a backward step). Then from

$$\lambda^{t+1} = \frac{1}{\epsilon}(\sum_i L(Z_i, \hat{\beta}^t) - \sum_i L(Z_i, \hat{\beta}^{t+1}) - \xi)$$

we have

$$\lambda^{t+1}(\|\beta^{t+1}\|_1 - \|\beta^t\|_1) = \sum_i L(Z_i, \hat{\beta}^t) - \sum_i L(Z_i, \hat{\beta}^{t+1}) - \xi$$

$$\sum_i L(Z_i, \hat{\beta}^{t+1}) + \lambda^{t+1}\|\beta^{t+1}\|_1 = \sum_i L(Z_i, \hat{\beta}^t) + \lambda^{t+1}\|\beta^t\|_1) - \xi$$

$$\begin{aligned}
\sum_i L(Z_i, \hat{\beta}^{t+1}) + \lambda^t\|\beta^{t+1}\|_1 &= \sum_i L(Z_i, \hat{\beta}^t) + \lambda^t\|\beta^{t+1}\|_1 + \lambda^{t+1}\|\beta^t\|_1 - \lambda^{t+1}\|\beta^{t+1}\|_1 - \xi \\
&= \sum_i L(Z_i, \hat{\beta}^t) + \lambda^t\|\beta^t\|_1 + \lambda^t\epsilon + \lambda^{t+1}\|\beta^t\|_1 - \lambda^{t+1}\|\beta^t\|_1 - \lambda^{t+1}\epsilon - \xi \\
&= \sum_i L(Z_i, \hat{\beta}^t) + \lambda^t\|\beta^t\|_1 + \lambda^t\epsilon - \lambda^{t+1}\epsilon - \xi \\
&> \sum_i L(Z_i, \hat{\beta}^t) + \lambda^t\|\beta^t\|_1 - \xi
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_i L(Z_i, \hat{\beta}^{t+1}) + \lambda^t\|\beta^{t+1}\|_1 &= \min_{j,|s|=\epsilon} \sum_i L(Z_i, \hat{\beta}^t + s1_j) + \lambda^t\|\beta^{t+1}\|_1 \\
&\leq \sum_i L(Z_i, \hat{\beta}^t \pm \epsilon 1_j) + \lambda^t\|\beta^{t+1}\|_1
\end{aligned}$$

So we have

$$\Gamma(\hat{\beta}^t \pm \epsilon 1_j, \lambda^t) > \Gamma(\hat{\beta}^t, \lambda^t) - \xi \text{ for every } j$$

□

定理38. *For a finite base learners and $\xi = o(\epsilon)$, if $\sum L(Z_i, \beta)$ is strongly convex with bounded second derivatives in β then as $\epsilon \to 0$, the Blasso path converges to the Lasso path uniformly.*

Note that strong convexity and bounded second derivatives imply that $M \geq m > 0$:

$$mI \preccurlyeq \Delta^2 \sum L \preccurlyeq MI$$

References

Y Freund, R Schapire (1999). A short introduction to boosting. JOURNAL-JAPANESE SOCIETY FOR Artificial Intelligence, 14(5):771-780.

Peter Buhlmann and Bin Yu (2003). Boosting with the L2 Loss: Regression and Classification. J. Amer. Statist. Assoc. 98, 324-340.

item J Friedman, T Hastie and R. Tibshirani (2000). Additive logistic regression: A statistical view of boosting- The annals of statistics.