Notes from (Parametric) Statistical Theory

Cambridge Part III Mathematical Tripos 2012-2013

Lecturer: Richard Samworth

Vivak Patel

April 23, 2013

Contents

Part I
Background Theory

1 Introduction

1.1 Multivariate Normal Distribution

1. Definitions of Normal Distribution

   Definition 1.1. *A random variable X has:*

   (a) A univariate normal distribution if its density function is:

   $$f(x, \mu, \sigma^2) = \left(2\pi\sigma^2\right)^{-1/2} \exp\left[\frac{-1}{2\sigma^2}\left(x-\mu\right)^2\right] = N(\mu, \sigma^2)$$

   (b) A multivariate normal distribution if

   $$\forall t \in \mathbb{R},\ t^T X \sim N$$

2. Linear Transformations of Multivariate Normal

   Proposition 1.1. *Let $X \sim N_n(\mu, \Sigma)$*

   (a) If $A \in \mathbb{R}^{m\times n}$ then $AX \sim N_m(A\mu, A\Sigma A^T)$.
   (b) If Σ is positive definite then $(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_n$
   (c) If $\Sigma = \sigma^2 I$ then

   $$\|X\|^2 \sim \sigma^2\chi^2_n\left(\frac{\|\mu\|}{\sigma^2}\right)$$

   Proof. Let $X \sim N_n(\mu, \sigma^2)$.

   (a) First we prove that AX is normally distributed, then determine its mean and covariance.
      i. Let $t \in \mathbb{R}^m$. Then $t^T A$ is an arbitrary vector in $\mathbb{R}^n$. Therefore, $(t^T A)X$ is normally distributed since X is multivariate normally distributed.
      ii. For the mean and variance, we have:

      $$\mathbb{E}\left[AX\right] = A\mathbb{E}\left[X\right] = A\mu$$
      $$\mathbf{Cov}\left[AX\right] = A\mathbf{Cov}\left[X\right]A^T = A\Sigma A^T$$

   (b) Note that for a matrix $A \in \mathbb{R}^{m\times n}$ and vector $b \in \mathbb{R}^n$,

   $$A(X-b) \sim N_m(A(\mu-b), A\Sigma A^T)$$

   Therefore,

   $$Z = \Sigma^{-1/2}(X-\mu) \sim N_n(0, I)$$

   By defnition of χ^2,

   $$\|Z\|^2 \sim \chi^2_n$$

(c) By the definition of the non-central χ^2 distribution, $\|X\|^2$ is noncentral χ^2_n with

$$\lambda = \sum_{i=1}^{n} \frac{\mu_i^2}{\sigma_i^2} = \frac{\|\mu\|^2}{\sigma^2}$$

□

3. Partitioning a Multivariate Normal Random Vector

Proposition 1.2. *Let $X \sim N_n(\mu, \Sigma)$. Let $X = (X_1^T \ X_2^T)^T$ where X_1 has n_1 components and X_2 has $n_2 = n - n_1$ components. Let $\mu = (\mu_1^T \ \mu_2^T)$ and*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

Then:

(a) $X_2 \sim N_{n_2}(\mu_2, \Sigma_{22})$

(b) X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$

(c) If Σ_{22} is positive definite then

$$X_1|X_2 = x_2 \sim N_{n_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)$$

Proof. Let X, X_1, X_2, μ and Σ be as above.

(a) We use the previous proposition with a matrix:

$$A = \begin{pmatrix} 0 & 0 \\ 0 & I_{n_2} \end{pmatrix}$$

(b) We use the characteristic equations. Let $t = (t_1 \ t_2)$

$$\begin{aligned} \psi_X(t) =& \exp\left[it^T\mu - \frac{t\Sigma t^T}{2}\right] \\ =& \exp\left[it_1^T\mu_1 - \frac{t_1\Sigma_{11}t_1^T}{2} + it_2^T\mu_2 - \frac{t_2\Sigma_{22}t_2^T}{2} - t_1\Sigma_{12}t_2^T\right] \\ =& \psi_{X_1}(t_1)\psi_{X_2}(t_2)\exp\left[-t_1\Sigma_{12}t_2^T\right] \end{aligned}$$

Therefore, when $\Sigma_{12} = 0$, X_1 and X_2 are independent.

(c) First we (cleverly) define a random variable Y in terms of X_1 and X_2, which is independent of X_2 and determine its distribution. Then, from this we can recover the distribution of $X_1|X_2 = x_2$.

i. Let $Y = X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2$. By Proposition 1.1, this a multivariate random vector. We compute:

$$\begin{aligned} \mathbb{E}[Y] &= \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2 \\ \mathbf{Cov}[Y] &= \mathbf{Cov}[X_1] - \Sigma_{12}\Sigma_{22}^{-1}\mathbf{Cov}[X_2]\Sigma_{22}^{-1}\Sigma_{12}^T \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T \end{aligned}$$

ii. Independence of Y and X_2:

$$\begin{aligned}\mathbf{Cov}\left[Y, X_2\right] &= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} \\ &= 0\end{aligned}$$

iii. $X_1 = Y + \Sigma_{12}\Sigma_{22}^{-1}X_2$. Therefore, $X_1|X_2 = x_2 = Y + \Sigma_{12}\Sigma_{22}^{-1}x_2$. Computing its distribution gives us the desired result.

□

4. Conditional independence of components of X, a Random Multivariate Normal Vector.

   Proposition 1.3. *Let $X \sim N_n(\mu, \Sigma)$. Let $\Sigma^{-1} = \Omega = (\omega_{ij})_{i,j=1}^n$ where Ω is the precision matrix. Let $X_{-ij} = \{X_k | k \neq i, j\}$. Then, X_i and X_j are conditionally independent given X_{-ij} if and only if $\omega_{ij} = 0$.*

1.2 Cochran's Theorem

Theorem 1.1. *Let $Y \sim N_n(0, \sigma^2 I)$ and suppose $A_1, \ldots, A_k \in \mathbb{R}^{n\times n}$ are symmetric with $rank(A_i) = r_i$ and $\sum_i A_i = I$. If $\sum_i r_i = n$ then $Y^T A_i Y \sim \sigma^2 \chi^2_{r_i}$ and $Y^T A_1 Y, \ldots, Y^T A_k Y$ are independent.*

Proof. First we prove that the distribution of $Y^T A_i Y$ is χ^2 and then we prove independence.

1. We will do this using a diagonilisation argument to decompose $A_i = QDQ^T$ where Q is an orthogonal matrix ($QQ^T = I$), so that a new random variable $Z = Q^T Y$ is such that $Y^T A_i Y = Z^T Q^T DQZ \sim \sigma^2 \chi^2_{r_i}$

   (a) Decompose $A_i = QDQ^T$ where Q is an orthogonal matrix. Then $D = Q^T A_i Q$. Since $rank(A_i) = r_i$, the first r_i diagonals of D are not zero while the remaining $n - r_i$ are zero.

   (b) Using this fact, $Q^T(I - A_i)Q$ is also diagonal with $n - r_i$ diagonals all equal to 1. So $rank(Q^T(A_i - I)Q) \geq n - r_i$. By assumption, $I - A_i = \sum_{j\neq i} A_j$ and $\sum_{j\neq i} r_j = n - r_i$. Therefore $rank(Q^T(\sum_{j\neq i} A_j)Q) \leq n - r_i$

   (c) Therefore, $rank(Q^T(I - A_i)Q) = n - r_i$. So the diagonals of D must all be 1.

   (d) Let $Q^T Y = Z \sim N(0, \sigma^2 QQ^T) = N(0, \sigma^2 I)$. Therefore,

$$\begin{aligned}Y^T A_i Y = Y^T(QDQ^T)Y = Y^T QDDQ^T Y = Z^T DDZ \\ = \sum_{l=1}^{r_i} Z_l^2 \sim \sigma^2 \chi^2_{r_i}\end{aligned}$$

2. For independence, we want to show that the moment generating functions of $\frac{1}{\sigma^2}(Y^T A_1 Y, \ldots Y^T A_k Y)$ can be split into the moment generating function for $\frac{1}{\sigma^2} Y^T A_i Y$ which is $M(t_i) = (1 - 2t_i)^{-r_1/2}$ for $t_i < 1/2$.

   (a) Note that: $A_i^2 = QDQ^T QDQ^T = QDQ^T = A_i$ so A_i is an orthogonal projection.

(b) Note that: For $j \neq i$, $A_i A_j = 0$.

(c) For sufficiently small t_i, the matrices $I - 2t_i A_i$ and $I - 2\sum_i t_i A_i$ are positive definite. Therefore:

$$\begin{aligned}
M(t_1, \ldots, t_k) &= \int_{\mathbb{R}^n} \frac{\exp\left[\frac{-1}{2\sigma^2} y^T (I - 2\sum_i t_i A_i) y\right]}{(2\pi\sigma^2)^{n/2}} \\
&= \det\left(I - 2\sum_i t_i A_i\right)^{-1/2} \\
&= \det\left(\prod_i (I - 2t_i A_i)\right)^{-1/2} \\
&= \prod_i \det\left(I - 2t_i A_i\right)^{-1/2} \\
&= \prod_i \det\left[Q^T \left(I - 2t_i A_i\right) Q\right]^{-1/2} \\
&= \prod_i \det\left(I - 2t_i D_i\right)^{-1/2} \\
&= \prod_i (1 - 2t_i)^{-r_i/2} \\
&= \prod_i M(t_i)
\end{aligned}$$

□

1.3 Convergence

1. Forms of Convergence

Definition 1.2. *Let (Y_n) be a sequence of R.V.*

(a) $Y_n \xrightarrow{a.s.} Y$ if $\forall \epsilon > 0$ as $n \to \infty$

$$\mathbb{P}\left[\sup_{m \geq n} \|Y_m - Y\| > \epsilon\right] \to 0$$

(b) $Y_n \xrightarrow{P} Y$ if $\forall \epsilon >> 0$, as $n \to \infty$:

$$\mathbb{P}\left[\|Y_n - Y\| > \epsilon\right] \to 0$$

(c) $Y_n \xrightarrow{d} Y$ if for all bounded, continuous, real-valued functions f, either

$$\mathbb{E}\left[f(Y_n)\right] \to \mathbb{E}\left[f(Y)\right]$$
$$\mathbb{P}\left[Y_n \leq y\right] \to \mathbb{P}\left[Y \leq y\right] \text{ at all continuity points}$$

2. Strong Law of Large Numbers

Proposition 1.4. *If Y_n are i.i.d R.V. with mean $\mu < \infty$ then*

$$n^{-1} \sum_i Y_n \xrightarrow{a.s.} \mu$$

3. Central Limit Theorem.

   Proposition 1.5. *Suppose Y_n are i.i.d. with mean μ and positive definite covariance Σ. Then*

$$\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N_m(0, \Sigma)$$

2 Theory of Matrices

2.1 Matrix Norms

1. Operator Norms

   (a) Definition of Operator Norm

   Definition 2.1. *The p^{th} operator norm of A is*

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \|Ax\|_p$$

   (b) Properties:

   i. $\|A\|_1$ is the maximum absolute column sum
   ii. $\|A\|_\infty$ is the maximum absolute row sum
   iii. When $m = n$, $\|A\|_2$ is the spectral norm, and its square root is the largest eigenvalue of $A^T A$, and for P, Q orthogonal,

$$\|PAQ\|_2 = \|A\|_2$$

2. Entrywise Norms:

$$\|A\|_p = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p \right)^{1/p}$$

3. Schatten Norms: The Schatten p-norm of A is the L^p norm of the vector of A's singular values

2.2 Singular Value Decomposition

1. Definitions of Singular Value, and Left and Right Singular Vectors

   Definition 2.2. *Let $A \in \mathbb{R}^{m \times n}$. $\sigma \geq 0$ is a singular value of A if $\exists u \in \mathbb{R}^m$ and $\exists v \in \mathbb{R}^n$ such that $\|u\|_2 = \|v\|_2 = 1$ and $Av = \sigma u$ and $A^T u = \sigma v$*

   (a) v is the right singular vector

   (b) u is the left singular vector

2. Singular Value Decomposition

Theorem 2.1. *Let $A \in \mathbb{R}^{m\times n}$. $\exists U \in \mathbb{R}^{m\times m}$ and $\exists V \in \mathbb{R}^{n\times n}$ orthogonal such that $U^T AV = diag(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{m\times n}$ where*

(a) $r = \min(m, n)$

(b) $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$

Proof. This is a proof by contradiction. We assume that a SVD for A is not possible and is not possible for matrices whose smallest dimension is larger than $\min(m, n) = r$. To achieve the contradiction, we want to show for an A_1 with an SVD, that $A = \begin{pmatrix} \sigma_1 & 0 \\ 0 & A_1 \end{pmatrix}$ has an SVD.

(a) Let $\sigma_1 = \|A\|_2$. Then $\exists v_1 \in \mathbb{R}^n$ such that $\|v_1\|_2 = 1$ and $\|Av_1\|_2 = \sigma_1$. Then for $u_1 = \frac{Av_1}{\sigma_1}$, $\|u_1\|_2 = 1$. Therefore, σ_1 is a singular value of A with singular vectors v_1 and u_1. Let V_1 be an $n \times n$ orthogonal matrix with its first column as v_1 and U_1 be an orthogonal matrix with its first column u_1. Then

$$U_1^T AV_1 = \begin{pmatrix} \sigma_1 & \omega^T \\ 0 & A_1 \end{pmatrix} = B$$

(b) Notice that

$$\sigma_1^2 = \|A\|_2^2 = \left\|U_1^T AV_1\right\|_2^2 = \|B\|_2^2 = \max_{x\neq 0} \frac{\|Bx\|_2^2}{\|x\|_2^2}$$

Cleverly choosing $X = \begin{pmatrix} \sigma_1 \\ \omega \end{pmatrix}$, we can show that $\omega = 0$:

$$\begin{aligned} \sigma_1^2 &\geq \frac{\|BX\|_2^2}{\|X\|_2^2} \\ &= \frac{\left(\sigma_1^2 + \|\omega\|_2^2\right)^2 + \|A_1\omega\|_2^2}{\sigma_1^2 + \|\omega\|_2^2} \\ &\geq \sigma_1^2 + \|\omega\|_2^2 \end{aligned}$$

Therefore, $\omega = 0$

(c) Now $A_1 \in \mathbb{R}^{m-1\times n-1}$ so $r > \min(m-1, n-1)$. We can find the SVD of A_1: $\tilde{U}_2^T A_1 \tilde{V}_2 = diag(\sigma_2, \ldots, \sigma_r)$. Let:

$$U_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{U}_2 \end{pmatrix} \qquad V_2 = \begin{pmatrix} 1 & 0 \\ 0 & \tilde{V}_2 \end{pmatrix}$$

And define:

$$V = V_1 V_2 \qquad U = U_1 U_2$$

Then V and U are orthogonal, and:

$$U_2^T U_1^T AV_1 V_2 = U_2^T BV_2 = diag(\sigma_1, \sigma_2, \ldots, \sigma_r)$$

(d) Finally, $\sigma_1 = \|A\|_2 = \left\|U^T A V\right\| = \|diag(\sigma_1, \ldots, \sigma_r)\|$. Then $\sigma_1 \geq \sigma_2$.

□

Corollary 2.1. *Let $A \in \mathbb{R}^{m\times n}$. $\exists r \leq \min(m,n)$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, and $\Sigma = diag(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{n\times n}$ with $\sigma_1 \geq \cdots \geq \sigma_r$ such that $A = U\Sigma V^T$*

3. Approximating low rank matrices:

Theorem 2.2. *Suppose $A \in \mathbb{R}^{m\times n}$ has SVD $U\Sigma V^T$ where $U = (u_1\ u_2 \cdots u_r)$ and $V = (v_1\ v_2 \cdots v_r)$. Let $k < r$. Define*

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

Then

$$\inf\{\|A - B\|_2 \,|B \in \mathbb{R}^{m\times n}, rank(B) = k\} = \|A - A_k\|_2 = \sigma_{k+1}$$

Proof. We state two important facts:

(a) $rank(A_k) = rank(U^T A_k V) = rank(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) = k$

(b) $\|A - A_k\|_2 = \left\|U^T(A - A_k)V\right\|_2 = \|diag(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_r)\| = \sigma_{k+1}$

Suppose $B \in \mathbb{R}^{m\times n}$ with $rank(B) = k$. The null space of B is of dimension $n-k$ so we can find its basis. The $\dim(span(v_1, \ldots, v_{k+1})) = k+1$ so the

$$(B) \cap span(v_1, \ldots, v_{k+1}) \neq \{0\}$$

Let z be in the intersection such that $\|z\|_2 = 1$. Then $Bz = 0$ and we can represent $z = \sum_{i=1}^{k+1} \alpha_i v_i$ with $\sum \alpha_i^2 = 1$. Then:

$$\|A - B\|_2^2 \geq \|(A - B)z\|_2^2 = \|Az\|_2^2 = \sum_{i=1}^{k+1} \alpha_i^2 \sigma_i^2 \geq \sigma_{k+1}^2$$

□

2.3 Moore-Penrose Pseudoinverses

1. Existence and Uniqueness

Theorem 2.3. *$\forall A \in \mathbb{R}^{m\times n}, \exists A^+ \in \mathbb{R}^{m\times n}$, a Moore-Penrose pseudoinverse, such that A^+A and AA^+ are symmetric, and*

(a) $A^+AA^+ = A^+$

(b) $AA^+A = A$

Proof. First we prove uniqueness and then existence.

(a) Uniqueness: Suppose $B, C \in \mathbb{R}^{m\times n}$ that satisfy Moore-Penrose requirements. By symmetry:

$$AB = B^TA^T = B^T(ACA)^T = B^TA^TC^TA^T = (AB)^T(AC)^T = ABAC = AC$$
$$BA = A^TB^T = (ACA)^TB^T = A^TC^TA^TB^T = (CA)^T(BA)^T = CABA = CA$$
$$C = CAC = (CA)C = (BA)C = B(AC) = BAB = B$$

(b) Existence: First we start with the special case of a singular value matrix, then use SVD of general matrices to prove the result.

i. Let $D = diag(\sigma_1, \ldots, \sigma_r)$. Let $D^+ = diag(\sigma_1', \ldots, \sigma_r'$ where $\sigma_i' = \sigma_i^{-1}\mathbf{1}\,[\sigma_i \neq 0]$. Then we notice that D^+ has the desired properties of a Moore-Penrose Pseudoinverse

ii. Let $A = U\Sigma V^T$. Since Σ is a diagonal matrix, we see that its pseudoinverse exists. Let $A^+ = V\Sigma^+U^T$. Then:

$$AA^+ = U\Sigma V^TV\Sigma^+U^T = U\Sigma\Sigma^+U^T$$
$$A^+A = V\Sigma^+U^TU\Sigma V^T = V\Sigma^+\Sigma V^T$$
$$AA^+A = U\Sigma V^TV\Sigma^+U^TU\Sigma V^T = U\Sigma V^T = A$$
$$A^+AA^+ = V\Sigma^+U^TU\Sigma V^TV\Sigma^+U^T = V\Sigma^+U^T = A^+$$

□

2. Closeness to a solution

Theorem 2.4. *Let $A \in \mathbb{R}^{m\times n}$. For $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$, for $z = A^+b$:*

$$\|Ax - b\|_2 \geq \|Az - b\|_2$$

Proof. Note that

$$A^T(Az - b) = A^TAA^+b - A^Tb = (AA^+A)^Tb - A^Tb = A^Tb - A^Tb = 0$$

Therefore:

$$\begin{aligned}\|Ax - b\|_2^2 &= \|A(x - z) + Az - b\|_2^2 \\ &= \|Az - b\|_2^2 + 2(x - z)^TA^T(Az - b) + \|A(x - z)\|_2^2 \\ &= \|Az - b\|_2^2 + \|A(x - z)\|_2^2 \geq + \|A(z - b)\|_2^2\end{aligned}$$

□

3. Closeness of a solution

Theorem 2.5. *Let $A \in \mathbb{R}^{m\times n}$. If $z = A^+b$ satisfies $Az = b$, then among all solutions $x \in \mathbb{R}^n$ of $Ax = b$, z has the smallest Euclidean norm.*

Proof. We expand as we did in the previous proof given that $Ax = b$ for some x:

Notice that:

$$z^T(x - z) = (A^+Az)^T(x - z) = z^T(A^+A)(x - z) = z^TA^+(Ax - AZ) = 0$$

Then:

$$\|x\|_2^2 = \|z\|_2^2 + 2z^T(x - z) + \|x - z\|_2^2 \geq \|z\|_2^2$$

□

3 Convex Analysis

1. Definitions

   Definition 3.1. *Convex functions and Subgradients:*

   (a) A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\forall x, y \in \mathbb{R}^n$ and $t \in (0, 1)$

   $$f((1-t)x + ty) \leq (1-t)f(x) + tf(y)$$

   (b) A vector $v \in \mathbb{R}^n$ is a subgradient of f at x if $\forall y \in \mathbb{R}^n$

   $$v^T(y - x) + f(x) \leq f(y)$$

   (c) The set of all subgradients of f at x is denoted $\partial f(x)$.

2. Basic Properties

   (a) If f is finite and convex, $\partial f(x)$ is a nonempty, compact and convex set.

   (b) Karush-Kuhn-Tucker Criterion:

   $$x^* \in \arg\min_{x \in \mathbb{R}^n} f(x) \Longleftrightarrow 0 \in \partial f(x)$$

3. Subgradients and Differentiability

   Proposition 3.1. *Let $f : f : \mathbb{R}^n \to \mathbb{R}$ be convex. Suppose f is differentiable at x. Then*

   $$\partial f(x) = \{\bigtriangledown f(x)\}$$

4. Sum of two functions and Subgradients

   Proposition 3.2. *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be convex. Suppose f is differentiable at x. Then*

   $$\partial(f + g) = \{\bigtriangledown f(x)\} + \partial g(x)$$

Part II
Classical Theory

4 Linear Models

1. Definition of a Linear Model

   Definition 4.1. *Let $Y_1, \ldots, Y_n$ be independent responses with $Y_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \sum_{j=1}^{p} x_{ij}\beta_j$ for x_{ij} explanatory variables and β_j unknown regression coefficients. This is a linear model and is written as*

   $$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$$

2. Rank of X

   (a) We normally assumpe that the $n \times p$ matrix X is of full rank p when $n \geq p$

   (b) In practice if $rank(X) < p$, we remove redundant variables until X has full rank

   (c) If X has full rank then $X^T X$ is non-signular, and is positive definite.

3. Maximum Likelihood Estimator

   (a) Definition of the MLE

   Definition 4.2. *Let*

   $$\theta = \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}$$

   Let $Y \sim N(X\beta, \sigma^2 I)$. Then likelihood function is $L(\theta) = N(X\beta, \sigma^2 I)$. The maximum likelihood estimator $\hat{\theta}$ maximises $L(\theta)$. So the MLE explains the data optimally.

   (b) Values of the MLE for β and θ for Normally distributed Y

   Lemma 4.1. *Let $L(\theta) = N(X\beta, \sigma^2 I)$. Then*

   i. $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$

   ii. $\hat{\sigma}^2 = \frac{1}{n}\left\|Y - X\hat{\beta}\right\|^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$

   iii. $\hat{\beta}$ *and* $\hat{\sigma}^2$ *are independent*

   Proof. The log-likelihood function is

   $$l(\theta) = \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(Y - X\beta)^T(Y - X\beta)$$

   i. Taking the derivative with respect to β, solving for $\hat{\beta}$ when the derivative is 0 gives:

   $$\frac{\partial l}{\partial \beta} = \frac{-1}{2\sigma^2}\left[-2X^T(Y - X\hat{\beta})\right] = 0$$
   $$\implies \hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

$$\begin{aligned}\hat{\beta} &= \left(X^T X\right)^{-1} X^T Y \\ &\sim N\left(\left(\left(X^T X\right)^{-1} X^T X\beta, \left(X^T X\right)^{-1} X^T\right)\sigma^2 I X \left(X^T X\right)^{-1}\right) \\ &\sim N\left(\beta, \sigma^2 \left(X^T X\right)^{-1}\right)\end{aligned}$$

ii. Taking the derivative with respect to σ^2, solving for $\hat{\sigma}^2$ when the derivative is 0 and using $\hat{\beta}$ for our estimate of β:

$$\begin{aligned}\frac{\partial l}{\partial \sigma^2} &= \frac{n}{-2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\left\|Y - X\hat{\beta}\right\|_2^2 = 0 \\ \implies \hat{\sigma}^2 &= \frac{\left\|Y - X\hat{\beta}\right\|_2^2}{n}\end{aligned}$$

iii. To get the distribution of $\hat{\sigma}^2$ we make use of Cochran's Theorem

A. Let $P = X(X^T X)^{-1} X^T$ so that $X\hat{\beta} = PY$. We note the following properties of P:

$$\begin{aligned}P^T &= P \\ P^2 &= P \\ PX &= X\end{aligned}$$

B. So we have that

$$\left\|Y - X\hat{\beta}\right\|^2 = Y^T(I-P)(I-P)Y = Y^T(I-P)Y$$

C. However, $Y \sim N(X\beta, \sigma^2 I)$, and to apply Cochran's Theorem, we need to centre Y:

$$\begin{aligned}Y^T(I-P)Y &= \left[(Y - X\beta)^T - (X\beta)^T\right](I-P)\left[X\beta + (Y - X\beta)\right] \\ &= (Y - X\beta)^T(I-P)(Y - X\beta) \\ &\quad + 2\left[\beta^T X^T (Y - X\beta) - \beta^T (PX)^T (Y - X\beta)\right] \\ &= (Y - X\beta)^T(I-P)(Y - X\beta)\end{aligned}$$

D. Therefore, by Cochran's Theorem, since $rank(I - P) = p$, $Y^T(I-P)Y\sigma^2\chi^2_{n-p}$. Therefore:

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}$$

□

(c) Rudimentary Variable Selection

i. First note that P is an $n \times n$ matrix that is an orthogonal projection onto a p-dimensional subspace:

$$U = \{Xb | b \in \mathbb{R}^p\} = \{Pb | b \in \mathbb{R}^n\}$$

ii. Suppose we want to know for $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, if β_1, which has $p - p_0$ components, is 0. Then we need to test the hypothesis that $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

A. Under H_0, we compute:

$$\hat{\hat{\beta}} = \left(X_0^T X_0\right)^{-1} X_0^T Y$$
$$\hat{\hat{\sigma}}^2 = n^{-1} \left\| Y - X_0 \hat{\hat{\beta}} \right\|^2$$
$$\hat{\hat{Y}} = X_0 \hat{\hat{\beta}} = P_0 Y$$

B. Under H_1, we compute:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$
$$\hat{\sigma}^2 = n^{-1} \left\| Y - X\hat{\beta} \right\|^2$$
$$\hat{Y} = X\hat{\beta} = PY$$

iii. Then the log-likelihood ratio, since $\|Y - P_0 Y\|^2 = \|Y - PY\|^2 + \|PY - P_0 Y\|^2$ is:

$$\begin{aligned} w_{LR}(H_0) =& 2\left[l(\hat{\theta}) - l(\hat{\hat{\theta}})\right] \\ =& n \log\left(\frac{\|Y - P_0 Y\|^2}{\|Y - PY\|^2}\right) \\ =& n \log\left(1 + \frac{\|PY - P_0 Y\|^2}{\|Y - PY\|^2}\right) \end{aligned}$$

iv. Under H_0, and noting that $P - P_0$ and $I - P_0$ are symmetric with ranks $p - p_0$ and $n - p$, Cochran's Theorem gives that:

$$\|(P - P_0)Y\|^2 \sim \sigma^2 \chi^2_{p-p_0}$$
$$\|(I - P)Y\|^2 \sim \sigma^2 \chi^2_{n-p}$$

v. Therefore:

$$w_{LR} \sim F_{p-p_0, n-p}$$

(d) Estimation Performance of the MLE

Lemma 4.2. *Given the MLE for multivariate normal data Y, and letting $\Omega = n\left[X^T X\right]^{-1}$*

i. $\mathbb{E}\left[\frac{1}{n}\left\|X\hat{\beta} - X\beta\right\|_2^2\right] = \frac{\sigma^2 p}{n}$

ii. $\mathbb{E}\left[\left\|\hat{\beta} - \beta\right\|_1\right] = \frac{\sigma}{\sqrt{n}} \sum_{i=1}^{p} \sqrt{\frac{2\omega_{ii}}{\pi}}$

Proof. Note that $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$. Therefore:

i. $X(\hat{\beta}-\beta) \sim N(0, \sigma^2 P)$. Then

$$P^{-1/2}X(\hat{\beta}-\beta) \sim N(0, \sigma^2 I)$$

Applying Cochran's theorem:

$$\left\|X(\hat{\beta}-\beta)\right\|_2^2 = (X\hat{\beta}-X\beta)^T P^{-1/2} P P^{-1/2}(X\hat{\beta}-X\beta) \sim \sigma^2 \chi_p^2$$

Therefore,

$$\mathbb{E}\left[\frac{1}{n}\left\|X(\hat{\beta}-\beta)\right\|_2^2\right] = \frac{1}{n}\sigma^2 p$$

ii. We work with $\left|\hat{\beta}_k - \beta_k\right|$ first:

$$\begin{aligned}\hat{\beta}_k - \beta_k =& e_k^T\left(\hat{\beta}-\beta\right) \sim N\left(0, \frac{\sigma^2}{n} e_k^T \Omega e_k\right)\\ \sim& N(0, \frac{\sigma^2}{n}\omega_{kk})\end{aligned}$$

Therefore, $\left|\hat{\beta}_k - \beta_k\right|$ is a half-normal random variable, and its mean is:

$$\mathbb{E}\left[\left|\hat{\beta}_k - \beta_k\right|\right] = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{2\omega_{kk}}{\pi}}$$

□

4. Linear Estimators

(a) Definition of a Linear Estimator

Definition 4.3. *Suppose we have a linear model $Y = X\beta + \epsilon$ with an unknown β. A linear estimator of β is of the form AY where A is a $p \times n$ matrix.*

(b) The Ordinary Least Squares Estimator

i. Definition of RSS and Ordinary LSE:

Definition 4.4. *The residual sum of squares for some β' is defined as $RSS(\beta') = \|Y - X\beta'\|^2$. The ordinary least squares estimator is:*

$$\tilde{\beta} = \arg\min_{\beta \in \mathbb{R}^p} RSS(\beta)$$

ii. In the linear model, given that $\mathbb{E}[\epsilon] = 0$ and $\mathbf{Cov}[\epsilon] = \sigma^2 I$, the Ordinary LSE $\tilde{\beta} = (X^T X)^{-1} X^T Y$.

iii. Adding the assumption that $\epsilon \sim N(0, \sigma^2 I)$, gives us that $\tilde{\beta} = \hat{\beta}$

(c) The Gauss-Markov Theorem

Theorem 4.1. *Let $Y = X\beta + \epsilon$ where X is of full rank $p < n$. Suppose $\mathbb{E}[\epsilon] = 0$ and $\boldsymbol{Cov}[\epsilon] = \sigma^2 I$. If $\tilde{\beta}$ is the Ordinary LSE and β' is any other linear unbiased estimator, then $\boldsymbol{Cov}\left[\tilde{\beta}\right] \leq \boldsymbol{Cov}[\beta']$ in the usual sense (i.e. $\forall t \in \mathbb{R}^p,\ t^T \boldsymbol{Cov}\left[\tilde{\beta}\right] < t^T \boldsymbol{Cov}[\beta']\, t$).*

Proof. We compute the covariance of any linear estimator in terms of the covariance of the Ordinary LSE:

i. Let β' be a linear unbiased estimator of β. Then $\beta' = AY$. Therefore:

$$\begin{aligned}\mathbb{E}[\beta'] &= AX\beta \\ \mathbf{Cov}[\beta'] &= \sigma^2 AA^T\end{aligned}$$

ii. However, β' is unbiased. Therefore, $AX\beta = \beta \implies AX = I$. Now, let $P = (X^TX)^{-1}X^T$, and let $B = A - P$. Then:

$$\begin{aligned}\mathbf{Cov}[\beta'] &= \sigma^2(B+P)(B+P)^T \\ &= \sigma^2(BB^T + (BP^T) + (BP^T)^T + PP^T) \\ &= \sigma^2(BB^T + PP^T) \\ &= \sigma^2(BB^T) + \mathbf{Cov}\left[\tilde{\beta}\right]\end{aligned}$$

Since:

$$\begin{aligned}BP^T &= (X^TX)^{-1}X^TA^T - (X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= (X^TX)^{-1}(AX)^T - (X^TX)^{-1} \\ &= 0\end{aligned}$$

Moreover, for any $t \in \mathbb{R}^p$, $t^TBB^Tt = \left\|B^Tt\right\|^2 \geq 0$. Therefore, $\mathbf{Cov}[\beta'] \geq \mathbf{Cov}\left[\tilde{\beta}\right]$

□

5 Parametric Theory

5.1 Introduction

1. : Framework: Suppose $Y_1, \ldots, Y_n$ are random variables, and let Y be the vector of these random variables. Let E be the state space of Y. Let $\Theta \subset \mathbb{R}^p$ be the parameter space of a model $f : E \times \Theta \to (0, \infty)$ such that $\forall\theta \in \Theta$, $f(\theta, \cdot)$ is a density with respect to a dominating measure μ.

2. Examples

Example 5.1. ***Nonlinear Least Squares Estimator.*** *Suppose $Z_i = g(\theta, x_i) + u_i$ where (x_i, Z_i) are i.i.d., g is unknown, and $\mathbb{E}[u_i|Z_i] = 0$. The Nonlinear LSE is:*

$$\hat{\theta}_n \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n}(Z_i - g(x_i, \theta))^2$$

Example 5.2. ***Maximum Likelihood Estimator.*** *Suppose $Y_1, \ldots, Y_n$ are i.i.d with density $f(\theta, y)$. Let $l(\theta) = \sum_{i=1}^{n} \log f(\theta, Y_i)$ be the log-likelihood function. The maximum likelihood estimator $\hat{\theta}_n$ maximises $l(\theta)$.*

5.2 Likelihood

1. Definitions of Score, Information Matrix and Fisher Information Matrix

   Definition 5.1. *For fixed $y \in E$, let $L(\theta) = f(\theta, y)$ be the likelihood function and $l(\theta) = \log L(\theta)$ be the log-likelihood function.*

   (a) The MLE is $\hat{\theta}_n \in \arg\max_{\theta\in\Theta}(l(\theta, y))$

   (b) If $f(\theta, y)$ is differentiable in θ, the score function $U(\theta) = \nabla_\theta l(\theta, y)$

   (c) If $f(\theta, y)$ is twice differentiable in θ, the information matrix is $j(\theta) = \nabla_\theta \nabla_\theta^T l(\theta, y)$

   (d) The Fisher Information Matrix is:

   $$i(\theta) = -\mathbb{E}_\theta \left[j(\theta)\right] = -\int_E j(\theta) f(\theta, y) dy$$

   d

2. Equivariance

   (a) Definition of Equivariance

   Definition 5.2. *A parameter is equivariant if it preserves the problem under an injective transformation of the space.*

   (b) Types

   i. Shift Invariance: if the data values undergo a shift, the parameter undergoes the same shift

   ii. Scale Invariant: if the data is scaled by a fixed amount,t he parameter is scaled accordingly

   iii. Parameterisation Invariant: inferences made about data using a model with parameter θ should be the same if the model uses φ where φ is an injective transformation of θ.

   (c) Properties

   i. The likelihood function, and (hence) the MLE, are parameterisation invariant. (i.e. suppose $\varphi_0 = h(\theta_0)$, h is one-to-one, then $L^{(\varphi)}(\varphi_0) = L^{(\theta)}(\theta_0)$ and $\hat{\varphi} = h(\hat{\theta})$)

   ii. The score function and fisher information matrix are not parameterisation invariant.

5.3 Consistency of M-Estimators

1. Definitions of Criterion Function and M-Estimators

   Definition 5.3. *Let $Y_1, \ldots, Y_n$ be data.*

   (a) A function $Q_n(\theta) = Q_n(\theta|Y_1, \ldots, Y_n)$ is a criterion function if it can be used to find an estimator for the data. It is almost always selected to be close to some function $Q(\theta)$ which is minimised at $\theta_0 \in \Theta$

   (b) An M-estimator is the minimum of $Q_n(\theta)$

2. Examples

Example 5.3. *The MLE minimises* $Q_n(\theta) = \frac{-1}{n} l(\theta, y)$ *for fixed* y*, and the LSE minimises* $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g(x_i,\theta)\right)^2$ *when* $\mathbb{E}\left[\epsilon_i\right] = 0$ *and* $\mathbb{E}\left[\epsilon_i^2\right] < \infty$.

Example 5.4. *Assuming that the* τ^{th} *quantile of* $\epsilon_i = 0$ *for* $i = 1, \ldots, n$*. The* τ^{th} *quantile estimator minimises:*

$$Q_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(y_i - g(x_i,\theta))$$

where

$$\rho_\tau(y) = \tau y \mathbf{1}\left[y \geq 0\right] + (1-\tau)|y|\mathbf{1}\left[y < 0\right]$$

3. Technical Points

   (a) We assume that $Q_n(\theta) \to Q(\theta)$. This is true in many cases.

   (b) Definition of Outer Measure

   Definition 5.4. *Let* $\mathbb{P}\left[\right]$ *be a probability measure for a space* $(\Omega, \mathcal{F})$*. The corresponding outer measure* $\mathbb{P}*$ *for any* $B \subset \Omega$ *is:*

$$\mathbb{P}^*(B) = \inf\{\mathbb{P}\left[A\right] | A \in \mathcal{F} \text{ s.t. } B \subseteq A\}$$

   Note 5.1. *We still write* $X_n \xrightarrow{\mathbb{P}} X$ *even if* $X_n \xrightarrow{\mathbb{P}^*} X$.

4. Consistency of M-Estimators

Theorem 5.1. *Suppose* $\Theta \subset \mathbb{R}^p$ *is compact. Suppose* $Q : \Theta \to \mathbb{R}$ *is continuous with a unique minimiser* θ_0*. If* $\sup_\theta |Q_n(\theta) - Q(\theta)| \to 0$ *then any* $\hat{\theta}_n \in \arg\min Q_n(\theta)$ *is consistent with* θ_0*.i.e.:*

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$$

Proof. Given that θ_0 minimises Q:

(a) By continuity, and since θ_0 is a unique minimum,

$$\forall \epsilon > 0, \exists \delta > 0 \text{ such that } \|\theta - \theta_0\| > \epsilon \implies Q(\theta) - Q(\theta_0) > \delta$$

(b) Now we can use the convergence of Q_n to Q to upper bound the probability that the M-estimator and the true value are different as $n \to \infty$. Then, as $n \to \infty$:

$$\begin{aligned}
\mathbb{P}^*_{\theta_0}\left(\left\|\hat{\theta}_n - \theta_0\right\| > \epsilon\right) &\leq \mathbb{P}^*_{\theta_0}\left(Q(\hat{\theta}_n) - Q(\theta_0) > \delta\right) \\
&= \mathbb{P}^*_{\theta_0}\left(Q(\hat{\theta}_n) - Q_n(\hat{\theta}_n) + Q_n(\hat{\theta}_n) - Q(\theta_0) > \delta\right) \\
&\leq \mathbb{P}^*_{\theta_0}\left(2\sup_{\theta\in\Theta}|Q_n(\theta) - Q(\theta)| > \delta\right) \\
&\to 0
\end{aligned}$$

□

5.4 Asymptotic Normality of MLE

1. Motivation: Consistency discusses global behaviour of an estimator. To discuss the accuracy of $\hat{\theta}_n$ (e.g. how much it deviates from θ_0), we need to look at its asymptotic distribution.

2. Notation:
$$\frac{\partial}{\partial\theta} l(\theta_0) = \left[\frac{\partial}{\partial\theta} l(\theta)\right]_{\theta=\theta_0}$$

3. Properties of the Score Function

   (a) Behavioural condition on $\nabla_\theta f$ for determining the mean of the score function:

   Lemma 5.1. *Suppose $\exists\delta > 0$ such that $B(\delta, \theta_0) \subset \Theta$ on which $l(y, \theta)$ is differentiable in θ, $\forall y \in E$. If:*
$$\int_E \sup_{\theta\in B(\delta,\theta_0)} \|\nabla_\theta f\| \, d\mu(y) < \infty$$
   Then, $\mathbb{E}_{\theta_0}[U(\theta_0)] = 0$.

   Proof. We will prove this component wise:
$$\begin{aligned}
\mathbb{E}_{\theta_0}[U_j(\theta_0)] &= \int_E U_j(\theta_0) f(y, \theta_0) dy \\
&= \int_E \frac{\partial \log(f(y, \theta_0))}{\partial\theta_j} f(y, \theta_0) dy \\
&= \int_E \frac{\partial f(y, \theta_0)}{\theta_j} dy \\
\text{By the hypothesis: } &= \frac{\partial}{\partial\theta_j}\left[\int_E f(y, \theta) dy\right]_{\theta=\theta_0} \\
&= \frac{\partial}{\partial\theta_j} 1 = 0
\end{aligned}$$
   □

   (b) Behaviour conditions on first and second derivatives for computing the covariance of the score function:

   Lemma 5.2. *Suppose $\exists\delta > 0$ such that $B(\delta, \theta_0) \subset \Theta$ on which $l(y, \theta)$ is differentiable in θ, $\forall y \in E$. If:*
$$\int_E \sup_{\theta\in B(\delta,\theta_0)} \|\nabla_\theta f\| \, d\mu(y) < \infty$$
$$\int_E \sup_{\theta\in B(\delta,\theta_0)} \left\|\nabla_\theta \nabla_\theta^T f(y, \theta)\right\| dy < \infty$$
$$\int_E \|U(\theta_0)\|^2 \, d\mu(y) = \mathbb{E}_{\theta_0}\left[\|U(\theta_0)\|^2\right] < \infty$$
   Then $\boldsymbol{Cov}_{\theta_0}[U(\theta_0)] = i(\theta_0)$.

Proof. Using the previous lemma, satisfied by the first hypothesis:

$$\mathbf{Cov}_{\theta_0}\left[U(\theta_0)\right] = \mathbb{E}_{\theta_0}\left[U(\theta_0)U^T(\theta_0)\right]$$

Starting componentwise, we start with $i(\theta_0)$ and work backwards. By the second hypothesis:

$$\begin{aligned} i_{rs}(\theta_0) &= -\int_E \partial_{\theta_r}\left(\partial_{\theta_s}\log f(y,\theta_0)\right) f(y,\theta_0)dy \\ &= -\int_E \frac{\partial^2}{\partial\theta_r\partial\theta_s} f(y,\theta_0)dy + \int_E \partial_{\theta_r}\log f(y,\theta_0)\partial_{\theta_s}\log f(y,\theta_0) f(y,\theta)dy \\ &= -\frac{\partial^2}{\partial\theta_r\partial\theta_s}\left[\int_E f(y,\theta)dy\right]_{\theta=\theta_0} + \mathbb{E}_{\theta_0}\left[U_r(\theta_0)U_s(\theta_0)\right] \end{aligned}$$

The first integral is 0 and the expectation exists by the third hypothesis. □

4. The fisher information matrix says that when $i(\theta_0)$ is small, θ near θ_0 will have very similar distributions, while when $i(\theta_0)$ is very large, θ and θ_0 will have very distinguishable distributions.

5. Technical Point:

   Note 5.2. *Let A_n be a sequence of events such that $\mathbb{P}\left[A_n^C\right] \to 0$. It is sufficient to show that $\mathbb{P}^*\left[\{X_n \le x\} \cap A_n\right] \to \mathbb{P}\left[X \le x\right]$ on all continuity points $x \in C$ to have that $\mathbb{P}^*\left[X_n \le x\right] \to \mathbb{P}\left[X \le x\right]$*

6. Asymptotic Normality

   Theorem 5.2. *Let Let Θ be a correctly specified parameter space. $f : E \times \Theta \to (0,\infty)$ be a family of density functions satisfying the following regularity conditions, given that $Y_1,\ldots,Y_n$ are i.i.d. with density $f(y,\theta_0)$:*

   (a) Consistency: $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0$

   (b) θ_0 is an interior point of Θ and $f(y,\theta)$ is twice continuously differentiable in a neighbourhood $B(\delta,\theta_0)$ for all $y \in E$.

   (c) Let $U(\theta) = \nabla_\theta \log(f(y,\theta))$. Then:

$$\begin{aligned} &\int_E \sup_{\theta\in B(\delta,\theta_0)} \|\nabla_\theta f\|\, d\mu(y) < \infty \\ &\int_E \sup_{\theta\in B(\delta,\theta_0)} \left\|\nabla_\theta \nabla_\theta^T f(y,\theta)\right\| dy < \infty \\ &\int_E \|U(\theta_0)\|^2\, d\mu(y) = \mathbb{E}_{\theta_0}\left[\|U(\theta_0)\|^2\right] < \infty \end{aligned}$$

   (d) $\mathbb{E}_{\theta_0}\left[\sup_{\theta\in B(\delta,\theta_0)} \|j(\theta)\|\right] < \infty$

   (e) $i(\theta_0)$ is positive definite.

   Then:

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, i^{-1}(\theta_0))$$

Proof. The idea is that we use the mean value theorem (which can only be applied to components) to get a function with known mean and variance, to which we can apply the CLT. We will be proving convergence in distribution on events $A_n = \{\hat{\theta}_n in B(\delta/2, \theta_0)\}$ for which, by Hypothesis (a), $\mathbb{P}\left[A_n^C\right] \to 0$.

(a) Notation:

i. Score Function:
$$U(y, \theta) = \nabla_\theta \log f(y, \theta)$$

ii. Information Matrix:
$$j(y, \theta) = \nabla_\theta \nabla_\theta^T \log f(y, \theta)$$

iii. Mean Score Function:
$$U_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} U(Y_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log f(Y_i, \theta)$$

iv. Mean Information Matrix:
$$j_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} j(Y_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \nabla_\theta^T \log f(Y_i, \theta)$$

v. k^{th} component of $U_n(\theta)$ is denoted $U_{n,k}(\theta)$

vi. k^{th} column of $j_n(\theta)$ is denoted $j_{n,k}(\theta)$

(b) The proof relies on applying the Mean Value Theorem to $U_{n,k}$ at values θ_0 and $\hat{\theta}_n$ on events A_n. For $\bar{\theta}_n^{(k)} \in \{\hat{\theta}_n(1-t) + \theta_0 t | t \in (0,1)\}$:
$$U_{n,k}(\hat{\theta}_n) - U_{n,k}(\theta_0) = j_{n,k}(\bar{\theta}_n^{(k)}) \left[\hat{\theta}_n - \theta_0\right]$$

i. Since $\hat{\theta}_n$ is the MLE, the derivative of the log-likelihood function at the MLE must be zero, and since we are on A_n, this implies that $U_{n,k}(\hat{\theta}_n) = 0$.

ii. Now letting:
$$M_n = \left[j_{n,1}(\bar{\theta}_n^{(1)}) \;\; j_{n,2}(\bar{\theta}_n^{(2)}) \;\; \cdots \;\; j_{n,p}(\bar{\theta}_n^{(p)})\right]$$
we have that:
$$-U_n(\theta_0) = M_n^T \left[\hat{\theta}_n - \theta_0\right]$$

(c) We now look at the behaviour of $U_n(\theta_0)$. Since we satisfy Lemmas 5.1 and 5.2 by hypotheses (b) and (c):

i. By Lemma 5.1, $\mathbb{E}_{\theta_0}\left[U(\theta_0)\right] = 0$ and by Lemma 5.2,
$$\mathbf{Cov}_{\theta_0} U(\theta_0) = i(\theta)$$

ii. By the Central Limit Theorem, as $n \to \infty$:
$$\sqrt{n} U_n(\theta_0) \xrightarrow{d} N(0, i(\theta_0))$$

(d) We now look at the convergence of M_n column wise, which we suspect by the Uniform Law of Large Numbers will yield $\mathbb{E}_{\theta_0}[j(\theta_0)]$.

i. Let $h_k(\theta) = \mathbb{E}_\theta[j_k(\theta)]$. Then:

$$\begin{aligned}\left\|j_{n,k}(\bar{\theta}_n^{(k)}) - h_k(\theta_0)\right\| &\leq \left\|j_{n,k}(\bar{\theta}_n^{(k)}) - h_k(\bar{\theta}_n^{(k)})\right\| \\ &\quad + \left\|h_k(\bar{\theta}_n^{(k)}) - h_k(\theta_0)\right\| \\ &\leq \sup_{\theta \in B(\delta/2, \theta_0)} \|j_{n,k}(\theta) - h_k(\theta)\| \\ &\quad + \left\|h_k(\bar{\theta}_n^{(k)}) - h_k(\theta_0)\right\| \\ &\xrightarrow{\mathbb{P}} 0\end{aligned}$$

ii. The first term: By hypothesis (d), we can apply the Uniform Law of Large Numbers so that the term converges to 0 in probability.

iii. The second term: Hypothesis (a) implies that $\bar{\theta}_n^{(k)} \xrightarrow{\mathbb{P}} \theta_0$ and since i is continuous by Hypothesis (b), we have that:

$$\left\|i(\bar{\theta}_n^{(k)}) - i(\theta_0)\right\| \xrightarrow{\mathbb{P}} 0$$

iv. Therefore,

$$M_n \xrightarrow{\mathbb{P}} h(\theta_0) = -i(\theta_0)$$

v. Let $B_n = \{-M_n \text{ is positive defnite}\}$. Since $-M_n \xrightarrow{i} (\theta_0)$ and by Hypothesis (e), $\mathbb{P}\left[B_n^C\right] \to 0$ as $n \to \infty$.

(e) Therefore, on events $A_n \cap B_n$ and by Slutsky's Lemma:

$$\begin{aligned}\sqrt{n}\left[\hat{\theta}_n - \theta_0\right] &= \sqrt{n}\left(-M_n^T\right) U_n(\theta_0) \\ &\xrightarrow{d} N_p(0, i^{-1}(\theta_0) i(\theta_0) i^{-1}(\theta_0))\end{aligned}$$

□

Part III
High-dimensional Parametric Theory

6 Traditional Model Selection

By reducing the number of parameters we increase the bias but smaller models have advantages such as:

1. Prediction error may be reduced by not over-fitting data
2. Interpretation is much easier
3. In many applications, it is believed that the true underlying model is sparse

6.1 Variable Selection

1. Single Coefficient Selection
   (a) Hypotheses: $H_0 : \beta_j = 0$ and $H_1 : \beta_j \neq 0$
   (b) Test Statistic: $z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{\omega_{jj}}} \sim t_{n-p-1}$ where $\omega_{jj} = \left[X^T X\right]^{-1}_{jj}$
2. Group Coefficient Selection
   (a) Consider two models:
      i. One with $p_0 + 1$ parameters and RSS_0
      ii. One with $p_1 + 1 > p_0 + 1$ parameters and RSS_1
   (b) Hypothesis: H_0 : the model with $p_0 + 1$ parameters is the correct model
   (c) Test Statistic:

$$F = \frac{(RSS_0 - RSS_1)(p_1 - p_0)^{-1}}{RSS_1 (n - p_1 - 1)^{-1}} \sim F_{p_1 - p_o, n - p_1 - 1}$$

      i. The numerator is the reduction in the residual sum of squares when we change to a more complex model, and the denominator is an estimate of σ^2
      ii. The statistic tests for if the reduction is significant enough to warrant the increased complexity in the model

6.2 Subset Selections

1. Best Subsect Selection: Find $K \subset \{1, \ldots, p\}$ such that the parameter with only $|K|$ non-zero components has the lowest RSS.
   (a) Typically, RSS decreases as $|K|$ increases, so we need an auxiliary criterion:

i. Akaike Information Criterion (AIC): find $\hat{\beta}_K$ which minimises

$$n \log \left[\frac{1}{n} RSS \left(\hat{\beta}_K \right) \right] + 2|K|$$

ii. Bayesian Information Criterion (BIC): find $\hat{\beta}_K$ which minimises

$$n \log \left[\frac{1}{n} RSS \left(\hat{\beta}_K \right) \right] + |K| \log n$$

(b) Best Subset Selection is possible by the Leaps and Bounds Algorithm when the number of parameters is less than 40

2. Forward Selection: Start with the intercept term $\hat{\beta}_0 = \bar{y}$ and add parameters step-wise that results in the greatest decrease in RSS

(a) This will produce a sequence of estimators $\hat{\beta}_{[1]}, \hat{\beta}_{[2]}, \ldots, \hat{\beta}_{[p]}$. We stop the process when the following F statistic achieves a predetermined significant value:

$$\frac{RSS_{[k]} - RSS_{[k+1]}}{RSS_{[k+1]} \left(n - (k+1) - 1 \right)^{-1}} \sim F_{1, n-k-2}$$

(b) Again this statistic is a comparison of the reduction in RSS with the increase in complexity in comparison to the variance.

(c) Forward selection can be computed for very large p, and typically has lower variance than best subset selection

(d) Forward selection is a local method, not a global method.

3. Backward Selection: Analogous to forward selection, but we start with the complete model and remove the parameter component that results in the largest increase in RSS.

7 Shrinkage Estimators

7.1 Introduction and the Usual Estimator

1. Motivation: Selection methods typically have parameter estimations resulting in high variance. Shrinkage estimators exchange more bias for reduced variance to improve prediction performance.

2. Loss and Risk Functions

Definition 7.1. *The loss function $L(\hat{\theta}, \theta)$ is a measure of the proximity between the estimate and the parameter. The risk function is the expectation of the loss function.*

Example 7.1. *A simple example of loss would be square error loss: $L = (\hat{\theta} - \theta)^2$, which has a risk function: $R(\hat{\theta}, \theta) = MSE(\hat{\theta}) = n\mathbb{E}\left[L(\hat{\theta}, \theta) \right]$*

3. The Usual Estimator: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

(a) Let $X_1, \ldots, X_n$ be i.i.d. with a $N(\theta, 1)$ distribution.

(b) The Mean Square Error is: $R(\bar{X}_n, \theta) = 1$

Derivation. By independence:

$$\begin{aligned} n\mathbb{E}\left[\left(\left(\frac{1}{n}X_i\right) - \theta\right)^2\right] &= n\mathbb{E}\left[\frac{1}{n^2}\left(\sum_{i=1}^{n}(X_i - \theta)\right)^2\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)^2\right] \\ &= 1 \end{aligned}$$

□

7.2 Hodges Estimator

1. Let $X_1, \ldots, X_n$ be i.i.d. with $X \sim N(\theta, 1)$ for $\theta \in \mathbb{R}$

2. Definition:

   Definition 7.2. *Let $a \in [0, 1)$. The Hodges Estimator is:*

   $$\hat{\theta}_n^H = a\bar{X}_n \mathbf{1}\left[|\bar{X}_n| \leq n^{-1/4}\right] + \bar{X}_n \mathbf{1}\left[|\bar{X}_n| > n^{-1/4}\right]$$

3. Risk of the Hodges Estimator:

   $$R(\hat{\theta}_n^H, \theta) = a^2 \mathbf{1}\left[\theta = 0\right] + \mathbf{1}\left[\theta \neq 0\right]$$

   Derivation. We use of several simplifications to compute the risk of $\hat{\theta}_n^H$

   (a) Rewrite:

   $$\begin{aligned} a\bar{X}_n \mathbf{1}\left[|\bar{X}_n| \leq n^{-1/4}\right] + \bar{X}_n \mathbf{1}\left[|\bar{X}_n| > n^{-1/4}\right] \\ = \bar{X}_n\left[1 - (1-a)\mathbf{1}\left[|\bar{X}_n| \leq n^{-1/4}\right]\right] \end{aligned}$$

   (b) Using the Central Limit Theorem, for $z \sim N(0, 1)$

   $$\bar{X}_n \xrightarrow{d} \theta + n^{-1/2}z \sim N(\theta, \frac{1}{n})$$

   (c) Let $\gamma_n^+ = n^{1/4} - \theta n^{1/2}$ and $\gamma_n^- = -n^{1/4} - \theta n^{1/2}$, so that, for sufficiently large n:

   $$\mathbf{1}\left[|\bar{X}_n| \leq n^{-1/4}\right] = \mathbf{1}\left[\gamma_n^- \leq z \leq \gamma_n^+\right]$$

Then, we have:

$$\begin{aligned}
R =& n\mathbb{E}_\theta\left[\left(\bar{X}_n - \theta - \bar{X}_n(1-a)\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right)^2\right] \\
=& n\mathbb{E}\left[\left(\bar{X}_n - \theta\right)^2\right] + 2n(1-a)\mathbb{E}\left[\left(\bar{X}_n - \theta\right)\bar{X}_n\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
&+ n(1-a)^2\mathbb{E}\left[\bar{X}_n^2\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
=& 1 + 2n\theta(1-a)\mathbb{E}\left[\bar{X}_n\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
&+ n\left[(1-a)^2 - 2(1-a)\right]\mathbb{E}\left[\bar{X}_n^2\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
=& 1 + 2n\theta(1-a)\mathbb{E}\left[\left(\theta + zn^{-1/2}\right)\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
&+ n\left[a^2 - 1\right]\mathbb{E}\left[\left(\theta^2 + 2\theta z n^{-1/2} + z^2 n^{-1}\right)\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
=& 1 + n\theta^2\left[2(1-a) + a^2 - 1\right]\mathbb{E}\left[\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
&+ 2\sqrt{n}\theta\left[(1-a) + a^2 - 1\right]\mathbb{E}\left[z\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right] \\
&+ (a^2-1)\mathbb{E}\left[z^2\mathbf{1}\left[\gamma_n^- \le z \le \gamma_n^+\right]\right]
\end{aligned}$$

There are two cases:

$$\begin{aligned}
&\text{If } \theta = 0,\ \gamma_n^- \to -\infty \text{ and } \gamma_n^+ \to \infty \\
&\qquad \implies R = 1 - a^2 - 1 = a^2 \\
&\text{If } \theta \neq 0,\ \gamma_n^-, \gamma_n^+ \to -\infty \\
&\qquad \implies R = 1
\end{aligned}$$

□

7.3 James-Stein Estimator

1. Suppose $X \sim N_p(\theta, I)$ and $L(\hat{\theta}, \theta) = \left\|\hat{\theta} - \theta\right\|^2$
2. If we only have one data point, we can estimate θ using only this point $X = \hat{\theta}_0$. However, James and Stein have shown for $p \ge 3$, $\hat{\theta}_0$ is inadmissable (i.e. there is an estimator which dominates it, that is, has lower risk).

   Note 7.1. $R(\hat{\theta}_0, \theta) = p$

3. James-Stein Estimator

   Definition 7.3. *For $p \ge 3$, the James-Stein Estimator is:*

   $$\hat{\theta}^{JS}(X) = \left(1 - \frac{p-2}{\|X\|^2}\right)X$$

4. The Risk of the James-Stein Estimator
5. Properties of the James Stein Estimator
   (a) The James-Stein Estimator is dominated by its positive part, and Brown proved that there are still estimators which dominate $\hat{\theta}_+^{JS}$
   (b) Stein Phenomena: Estimating θ_i using X_i will result in greater risk in comparison of estimating θ using all components of X simultaneously

7.4 Ridge Regression

1. Formulations

   (a) Formal Definition

   Definition 7.4. *Let $\beta_0 \in \mathbb{R}$ such that we have a linear model*

   $$Y = \beta_0 \mathbf{1} + X\beta + \epsilon$$

   where $\epsilon \sim N(0, \sigma^2 I)$. Consider the criterion function, for $\lambda > 0$:

   $$Q_2(\beta_0, \beta) = \|Y - \beta_0 \mathbf{1} - X\beta\|^2 + \lambda \|\beta\|_2^2$$

   The Ridge Regression Estimator is: $\hat{\beta}_\lambda^R \in \arg\min Q_2$

   (b) Equivalent Formulation: $\hat{\beta}^R \in \arg\min \|Y - \beta_0 \mathbf{1} - X\beta\|^2$ subject to $\|\beta\|_2^2 \leq s$ where s and λ have a one-to-one correspondence

2. Motivation: Suppose X is nearly colinear, which occurs as $p \to n$. Then $\det X^T X$ will be small and at least one eigenvalue will be large. Since $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$, if one eigenvalue is large, $(X^T X)^{-1}$ will have a large trace, giving $\hat{\beta}$ a large variance. Ridge regression reduces this variance, but increases the bias.

3. Assumptions

   (a) Centred Columns: We centre the columns of X by replacing x_{ij} with $x_{ij} - \bar{x}_{\cdot,j}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

   (b) It is "easy" to infer that $\beta_0 = \bar{y}$. So we can replace y_i by $y_i - \bar{y}$

4. Given the assumptions, the solution is

   $$\hat{\beta}_\lambda^R = \left(X^T X + \lambda I\right)^{-1} X^T Y$$

   Derivation. Note that, under the assumptions, Q_2 can be written as

   $$Q_2(\beta) = [Y - X\beta]^T [Y - X\beta] + \lambda \beta^T \beta$$

   Then:

   $$\begin{aligned}
   \frac{\partial Q_2}{\partial \beta} &= -2X^T[Y - X\beta] + 2\lambda\beta \\
   0 &= -2X^T[Y - X\hat{\beta}_\lambda^R] + 2\lambda\hat{\beta}_\lambda^R \\
   X^T Y - X^T X \hat{\beta}_\lambda^R &= \lambda I \hat{\beta}_\lambda^R \\
   \hat{\beta}_\lambda^R &= (X^T X + \lambda I)^{-1} X^T Y
   \end{aligned}$$

   □

5. Improved performance of $\hat{\beta}_\lambda^R$ in comparison to MLE $\hat{\beta}$

   Theorem 7.1. *For sufficiently small $\lambda > 0$, in the usual sense:*

   $$\mathbb{E}\left[\left(\hat{\beta}_\lambda^R - \beta\right)\left(\hat{\beta}_\lambda^R - \beta\right)^T\right] \leq \mathbb{E}\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^T\right]$$

Proof. This is proved by direct computation, using several important facts:

(a) Important Facts

i. Note that $Y \sim N(\beta_0 \mathbf{1} + X\beta, \sigma^2 I)$ and $\bar{Y} = \beta_0 + \bar{\epsilon} \sim N(\beta_0, \frac{\sigma^2}{n})$, which can follow from the central limit theorem. Therefore, using the characteristic functions, we can have that:

$$Y - \bar{Y}\mathbf{1} \sim N\left(X\beta, \sigma^2\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right)$$

ii. Recall that $\hat{\beta}_\lambda^R = \left(X^T X + \lambda I\right)^{-1} X^T Y$ is the solution when $Y = Y - \bar{Y}\mathbf{1}$

iii. X has centred columns so that $\mathbf{1}^T X = X^T \mathbf{1} = 0$

iv. $MSE(\hat{\beta}_\lambda^R, \beta) = \mathbf{Cov}\left[\hat{\beta}_\lambda^R\right] - \left(\mathbb{E}\left[\hat{\beta}_\lambda^R - \beta\right]\right)\left(\mathbb{E}\left[\hat{\beta}_\lambda^R - \beta\right]\right)^T$

(b) Direct Computation

i. First:

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_\lambda^R - \beta\right] =& \left(X^T X - \lambda I\right)^{-1} X^T X\beta - \beta \\ =& \left(X^T X - \lambda I\right)^{-1}\left(X^T X - \lambda I\right)\beta \\ &+ \left(X^T X - \lambda I\right)^{-1}\lambda I\beta - \beta \\ =& \beta - \beta + \left(X^T X - \lambda I\right)^{-1}\lambda\beta\end{aligned}$$

ii. Second:

$$\begin{aligned}\mathbf{Cov}\left[\hat{\beta}_\lambda^R\right] =& \sigma^2\left(X^T X - \lambda I\right)^{-1} X^T\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X\left(X^T X - \lambda I\right)^{-1} \\ =& \sigma^2\left(X^T X - \lambda I\right)^{-1} X^T X\left(X^T X - \lambda I\right)^{-1}\end{aligned}$$

iii. Therefore:

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\beta} - \beta\right)\left(\hat{\beta} - \beta\right)^T\right] - \mathbb{E}\left[\left(\hat{\beta}_\lambda^R - \beta\right)\left(\hat{\beta}_\lambda^R - \beta\right)^T\right] \\ &= \sigma^2(X^T X)^{-1} - \sigma^2\left(X^T X - \lambda I\right)^{-1} X^T X\left(X^T X - \lambda I\right)^{-1} \\ &\quad - \lambda^2\left(X^T X - \lambda I\right)^{-1}\beta\beta^T\left(X^T X - \lambda I\right)^{-1} \\ &= \lambda\left(X^T X - \lambda I\right)^{-1}\left[2I\sigma^2 + \lambda(X^T X)^{-1}\sigma^2 - \lambda\beta\beta^T\right]\left(X^T X - \lambda I\right)^{-1}\end{aligned}$$

(c) As $\lambda \to 0$ the last two terms in the centre matrix will tend to the first, which is positive definite. So at sufficiently small λ the difference will be positive definite.

□

6. V-fold Cross Validation

(a) From the proof it is clear that we must carefully select λ so that the Ridge Regression Estimator performs better than the MLE

(b) In v-fold cross validation, we compute $\hat{\beta}^R_{\lambda,k}$, where $\hat{\beta}^R_{\lambda,k}$ is the ridge estimator f the data when the k^{th} fold is removed:

i. We do this on a grid of values for λ

ii. The data is split into v-folds of roughly equal size

(c) We select the λ that minimises:

$$c_v(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}^R_{\lambda,k(i),j})^2$$

where $k(i)$ is the fold to which Y_i belongs

8 Least Absolute Selection and Shrinkage Operator Estimator

8.1 Introduction

1. Definition

Definition 8.1. *The Least Absolute Selection and Shrinkage Operator (LASSO) estimator is the pair* $(\hat{\beta}_0, \hat{\beta}^L_\lambda)$ *that minimises, for* $\lambda > 0$

$$Q_1(\beta_0, \beta) = \frac{1}{2n}\left\|Y - \beta_0\mathbf{1} - X\beta\right\|_2^2 + \lambda\left\|\beta\right\|_1$$

2. Equivalent Formulations

(a) $\hat{\beta}^L_\lambda \in \arg\min \frac{1}{2n}\left\|Y - \beta_0\mathbf{1} - X\beta\right\|_2^2$ subject to $\left\|\beta\right\|_1 \leq s$ where $s = s(\lambda)$

(b) $\hat{\beta}^L_\lambda \in \arg\min \left\|\beta\right\|_1$ subject to $\frac{1}{2n}\left\|Y - \beta_0\mathbf{1} - X\beta\right\|_2^2 \leq t$ where $t = t(\lambda)$

3. Assumptions: The design matrix X is centred, $\hat{\beta}_0 = \bar{Y}$, and we replace Y with $Y - \bar{Y}$

4. Interpretations of how LASSO performs variable selection

(a) Orthonormal Interpretation: suppose X is orthonormal (i.e. $X^TX = I$), and let the Ordinary LSE be $\hat{\beta} = (X^TX)^{-1}X^TY = X^TY$

i. The equivalent criterion function for the orthonormal situation is:

$$Q(\beta) = \frac{1}{2n}\left\|\beta - \hat{\beta}\right\|_2^2 + \lambda\left\|\beta\right\|_1$$

Derivation. We simply rescale λ in the penultimate line to get:

$$\begin{aligned} Q_1(\beta) =& \frac{1}{2n}\left\|Y - X\hat{\beta} + X\hat{\beta} - X\beta\right\|_2^2 + \lambda\|\beta\|_1 \\ =& \frac{1}{2n}\left\|Y - XX^TY + X(\hat{\beta} - \beta)\right\|_2^2 + \lambda\|\beta\|_1 \\ =& \frac{1}{2n}\left\|Y - Y + X(\hat{\beta} - \beta)\right\|_2^2 + \lambda\|\beta\|_1 \\ Q(\beta) =& \frac{1}{2n}\left\|\hat{\beta} - \beta\right\|_2^2 + \lambda\|\beta\|_1 \end{aligned}$$

□

ii. Minimising the criterion function, we have:

$$\hat{\beta}^L_{\lambda,j} = \text{sgn}(\hat{\beta}_j)\left[|\beta_j| - \lambda\right]_+$$

iii. Notice that $\hat{\beta}^L_\lambda$ is a soft thresh-holding estimator in comparison to a hard thresholding estimator:

(b) Geometric Interpretation:

i. $\frac{1}{2n}\|Y - X\beta\|_2^2$ have contours which are ellipses centred about the minimum $\hat{\beta}$

ii. $\|\beta\|_2^2$ (Ridge Regression) have contours which are circles centred about 0, and the intersection of the ellipses with the circles is the solution for the Ridge Regression Estimator

iii. $\|\beta\|_1$ forms L^1 circles, or diamond contours centred about 0, and the intersection of the diamonds and the ellipses is the solution of the LASSO estimator. This typically occurs at a corner of the diamond, resulting in some parameters being set to 0.

8.2 Existence and Uniqueness of LASSO

1. Index Extraction Notation: Let $s \subset \{1, \dots, p\}$
   (a) Let $A \in \mathbb{R}^{n \times p}$. A_s is the $n \times |s|$ matrix with columns from A indexed by s
   (b) Let $b \in \mathbb{R}^p$. b_s is the vector whose components are the components of b with indices in s
   (c) $A_{-s} = A_{\{1,\dots,p\} \setminus s}$ and $b_{-s} = b_{\{1,\dots,p\} \setminus s}$
2. Existence
   (a) Existence of Solution

   Lemma 8.1. *The LASSO solution always exists*

   Proof. $Q_1(\beta)$ is continuous and convex. Moreover, $\lim_{\|\beta\|_2 \to \infty} Q_1(\beta) = \infty$. Therefore, $\arg\min Q_1(\beta) \neq \emptyset$. So the LASSO exists. □

   (b) Useful convex optimality formulation

   Lemma 8.2. *$\hat{\beta}^L$ is the LASSO solution if and only if for γ the subgradient of $f(X) = \|X\|_1$, evaluated at $\hat{\beta}^L$ is:*

$$\frac{1}{n} X^T (Y - X\hat{\beta}^L) = \lambda\gamma$$

   where

$$\gamma_j = \begin{cases} \operatorname{sgn}(\hat{\beta}^L) & \text{if } \hat{\beta}^L \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}^L = 0 \end{cases}$$

   Proof. $\hat{\beta}^L$ is a LASSO solution if and only if it minimises $Q(\beta) = f(\beta) + g(\beta)$ where $f(\beta) = \frac{1}{2n} \|Y - X\beta\|_2^2$ and $g(\beta) = \lambda \|\beta\|_1$
      i. By KKT, $\hat{\beta}^L \in \arg\min Q(\beta)$ if and only if $0 \in \partial Q(\hat{\beta}^L)$. Since f is differentiable, this is equivalent to:

$$0 \in \{\nabla f\} + \partial g$$

      ii. We now study $h(x) = |x|$. We find that $\gamma \in \partial h(x)$ if

$$\gamma = \begin{cases} \operatorname{sgn}(x) & \text{if } x \neq 0 \\ [-1, 1] & \text{if } x = 0 \end{cases}$$

         A. $x > 0$. Then $\forall y \in \mathbb{R}$ if $\gamma \in \partial h(x)$ then $\gamma(y - x) + |x| = \gamma y + (1 - \gamma)x \leq |y|$. This is satisfied if $\gamma = 1$.
         B. $x < 0$. Then $\forall y \in \mathbb{R}$ if $\gamma \in partial h(x)$ then $\gamma(y - x) + |x| = \gamma y + (\gamma + 1)|x| \leq |y|$. This is satisfied if $\gamma = -1$.
         C. $x = 0$. Then $\gamma y \leq |y|$ which is satisfied for $\gamma \in [-1, 1]$

iii. So if $\gamma \in \partial \frac{1}{\lambda} g(\hat{\beta}^L)$ then γ_j has the form specified.

iv. Therefore, $0 \in \{\nabla f\} + \partial g$ if and only if $0 = \frac{-1}{n} X^T(Y - X\beta) + \lambda\gamma$

□

3. Uniqueness

   (a) Definitions of Equicorrelation and Vector of Equicorrelation Signs

   Definition 8.2. *The Equicorrelation set is*

   $$\epsilon = \{j = 1, \ldots, p | \frac{1}{n}(X^T(Y - X\hat{\beta}^L))\} = \{j = 1, \ldots, p | |\gamma_j| = 1\}$$

   where the last formulation follows from the convex optimality formulation of the LASSO solution

   Definition 8.3. *The vector of Equicorrelation signs is* $s = \text{sgn} X_\epsilon^T (Y - X\hat{\beta}^L)$

   (b) Uniqueness

   Proposition 8.1. *If* $rank(X_\epsilon) = |\epsilon|$ *then* $\hat{\beta}^L$ *is unique and:*

   i. $\hat{\beta}^L_\epsilon = (X_\epsilon^T X_\epsilon)^{-1}(X_\epsilon^T Y - n\lambda s)$

   ii. $\hat{\beta}^L_{-\epsilon} = 0$

   Proof. Uniqueness follows from the convexity of the problem. To show that $\hat{\beta}^L_\epsilon = (X_\epsilon^T X_\epsilon)(X_\epsilon^T Y - n\lambda s)$:

   i. First we note that since $\hat{\beta}^L$ is unique, then γ and s must also be unique. Moreover, by definition of s, we have that $s = \gamma_\epsilon$.

   ii. It follows from the convex formulation of the LASSO that:

   $$\lambda n s = \lambda n \gamma_\epsilon = X_\epsilon^T (Y - X_\epsilon \hat{\beta}^L_\epsilon)$$

   iii. Solving for $\hat{\beta}^L_\epsilon = (X_\epsilon^T X_\epsilon)^{-1}(X_\epsilon^T - n\lambda s)$ given the rank assumption on X_ϵ so that $X_\epsilon^T X_\epsilon$ is positive definite.

   iv. Finally, by definition, $\gamma_{-\epsilon,j} \in (-1, 1)$. SO $\hat{\beta}^L_{-\epsilon,j} = 0$.

   □

8.3 Estimation and Prediction Properties of LASSO

1. Notation: Let β^0 be the true parameter with active set $S_0 = \{j | \beta_j^0 \neq 0\}$, inactive set N_0, and index $s_0 = |S_0|$

2. Basic Inequality between Criterion function and the Empirical Process

   Lemma 8.3. *Given the linear model* $Y = X\beta + \epsilon$*:*

   $$\frac{1}{n} \left\| X(\hat{\beta} - \beta^0) \right\|_2^2 + 2\lambda \left\| \hat{\beta} \right\|_1 \leq 2\frac{\epsilon^T X}{n}(\hat{\beta} - \beta^0) + 2\lambda \left\| \beta^0 \right\|_1$$

Proof. By definition $Q(\hat{\beta}) \leq Q(\beta^0)$. So we are simply expand and rearranging this inequality:

$$\begin{aligned}
&\left\|Y - X\hat{\beta}\right\|_2^2 + 2n\lambda \left\|\hat{\beta}\right\|_1 \leq \left\|Y - X\beta^0\right\|_2^2 + 2n\lambda \left\|\beta^0\right\|_1 \\
&\|Y\|_2^2 - 2Y^T(X\hat{\beta}) + \left\|X\hat{\beta}\right\|_2^2 + 2n\lambda \left\|\hat{\beta}\right\|_1 \\
&\quad \leq \|Y\|_2^2 - 2Y^T(X\beta^0) + \left\|X\beta^0\right\|_2^2 + 2n\lambda \left\|\beta^0\right\|_1 \\
&\left\|X\hat{\beta}\right\|_2^2 - 2(\beta^0)^T X^T X\hat{\beta} + \left\|X\beta^0\right\|_2^2 + 2n\lambda \left\|\hat{\beta}\right\|_1 \\
&\quad \leq 2Y^T X\left(\hat{\beta} - \beta^0\right) - 2(X\beta^0)^T X\hat{\beta} + 2(X\beta^0)^T X\beta^0 + 2n\lambda \left\|\beta^0\right\|_1 \\
&\left\|X\left(\hat{\beta} - \beta^0\right)\right\|_2^2 + 2n\lambda \left\|\hat{\beta}\right\|_1 \\
&\quad \leq 2(Y - X\beta^0)^T X\left(\hat{\beta} - \beta^0\right) + 2n\lambda \left\|\beta^0\right\|_1 \\
&\quad \leq 2\epsilon^T X\left(\hat{\beta} - \beta^0\right) + 2n\lambda \left\|\beta^0\right\|_1
\end{aligned}$$

□

3. The Empirical Process $\frac{\epsilon^T X}{n}\left(\hat{\beta} - \beta^0\right)$ is what we want to control, which we can do using Cauchy-Schwarz and studying the process on the event:

$$\Omega_0 = \left\{\frac{2}{n}\left\|\epsilon^T X\right\|_\infty < \lambda\right\}$$

which has probability

$$\mathbb{P}\left[\Omega_0\right] \geq 1 - p^{-A^2/8-1}$$

Lemma 8.4. *On Ω_0,*

$$\frac{1}{n}\left\|X\left(\hat{\beta} - \beta^0\right)\right\|_2^2 + \lambda\left\|\hat{\beta}_{N_0}\right\|_1 \leq 3\lambda\left\|\left(\hat{\beta}_{S_0} - \beta^0_{S_0}\right)\right\|_1$$

Proof. From the previous lemma, Cauchy-Schwarz, and on Ω_0, we have:

(a)

$$\begin{aligned}
\frac{1}{n}\left\|X\left(\hat{\beta} - \beta^0\right)\right\|_2^2 + 2\lambda\left\|\hat{\beta}_{N_0}\right\|_1 + 2\lambda\left\|\hat{\beta}_{S_0}\right\|_1 & \\
&\leq \frac{2}{n}\left\|\epsilon^T X\right\|_\infty \left\|\left(\hat{\beta} - \beta^0\right)\right\|_1 + 2\lambda\left\|\beta^0_{S_0}\right\|_1 \\
&\leq \lambda\left\|\left(\hat{\beta} - \beta^0\right)\right\|_1 + 2\lambda\left\|\beta^0_{S_0}\right\|_1 \\
&= \lambda\left\|\left(\hat{\beta}_{S_0} - \beta^0_{S_0}\right)\right\|_1 + \lambda\left\|\hat{\beta}_{N_0}\right\| + 2\lambda\left\|\beta^0_{S_0}\right\|_1
\end{aligned}$$

(b) By Reverse triangle inequality:

$$\frac{1}{n}\left\|X\left(\hat{\beta} - \beta^0\right)\right\|_2^2 + \lambda\left\|\hat{\beta}_{N_0}\right\|_1 \leq 3\lambda\left\|\left(\hat{\beta}_{S_0} - \beta^0_{S_0}\right)\right\|_1$$

□

4. Compatibility Condition

   (a) Motivation: we want to bound the right hand side of the previous lemma by $\frac{1}{2n}\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2$ and some non-random term

   (b) Derivation:

      i. By Cauchy-Schwarz:

$$\begin{aligned}\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\| &= \left\langle \hat{\beta}_{S_0}-\beta^0_{S_0}, \operatorname{sgn}\left(\hat{\beta}_{S_0}-\beta^0_{S_0}\right)\right\rangle \\ &\leq \sqrt{s_0}\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_2\end{aligned}$$

      ii. On Ω_0, we suppose $\exists\phi_0>0$ such that

$$\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_2^2 \leq \frac{\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2^2}{n\phi_0^2}$$

      Since $\hat{\beta}$ is random, we need to require this for all $\beta\in\mathbb{R}^p$. But this is too restrictive, so we require, by the second lemma, that it only be true for $b\in\mathbb{R}^p$ such that $\|b_{N_0}\|_1 \leq 3\|b_{S_0}\|_1$. Note that the inequality holds for each term separately in the second lemma.

   (c) Definition

      Definition 8.4. *The compatibility condition holds for set S_0 if $\exists\phi_0>0$, $\forall b\in\mathbb{R}^p$ such that whenever $\|b_{N_0}\|_1\leq 3\|b_{S_0}\|$, we have that*

$$\|b_{S_0}\|_1^2 \leq \frac{s_0}{n\phi_0^2}\|Xb\|_2^2$$

   (d) Interpretation: Consider the strong condition:

$$\|b_{S_0}\|_1^2 \leq s_0\|b_{S_0}\|_2^2 \leq \frac{s_0}{n\phi_0^2}\|Xb\|_2^2$$

      Recalling the definition of the spectral norm of matrix $\frac{X^TX}{n}$, this requires that this matrix's smallest eigenvalue is at least ϕ_0.

5. The approximate error in the LASSO estimator

   Theorem 8.1. *Suppose the Compatibility Condition holds for set S_0. Then on Ω_0:*

$$\frac{1}{2n}\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2^2+\lambda\left\|\hat{\beta}-\beta^0\right\|_1 \leq \frac{8\lambda^2 s_0}{\phi_0^2}$$

   Proof. We make use of the previous theorems and the Compatibility condition:

(a) By the second lemma, on Ω_0

$$\frac{1}{n}\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2^2+\lambda\left\|\hat{\beta}_{N_0}\right\| \leq 3\lambda\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_1$$
$$\frac{1}{n}\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2^2+\lambda\left\|\hat{\beta}_{N_0}\right\|+\lambda\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_1 \leq 4\lambda\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_1$$
$$\frac{1}{n}\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2^2+\lambda\left\|\hat{\beta}-\beta^0\right\|_1 \leq 4\lambda\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_1$$

(b) By the Compatibility Condition:

$$4\lambda\left\|\hat{\beta}_{S_0}-\beta^0_{S_0}\right\|_1 \leq 4\left(\frac{\lambda\sqrt{s_0}}{\phi_0}\right)\left(\frac{\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2}{\sqrt{n}}\right)$$

Noting that:

$$0 \leq \left(\frac{u}{\sqrt{2}}-2\sqrt{2}v\right)^2 \implies 4uv \leq \frac{u^2}{2}+8v^2$$

We have that:

$$4\left(\frac{\lambda\sqrt{s_0}}{\phi_0}\right)\left(\frac{\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2}{\sqrt{n}}\right) \leq \frac{\left\|X\left(\hat{\beta}-\beta^0\right)\right\|_2^2}{2n}+\frac{8\lambda^2 s_0}{\phi_0^2}$$

(c) Therefore, we have the desired result.

□

8.4 Noiseless Variable Selection

1. Noiseless case: $\epsilon = 0$ so $Y = X\beta^0$
2. Irrepresentable Conditions:
   (a) The irrepresentable condition is met for the set S if $\forall \tau_S \in \mathbb{R}^S$ satisfying $\|\tau_S\|_\infty \leq 1$ we have that $\left\|X_{N_0}^T X_{S_0}(X_{S_0}^T X_{S_0})^{-1}\tau_S\right\|_\infty < 1$.
   (b) The weak irrepresentable condition holds for some fixed $\tau_S \in \mathbb{R}^S$ with $\|\tau_S\|_\infty \leq 1$ if $\left\|X_{N_0}^T X_{S_0}(X_{S_0}^T X_{S_0})^{-1}\tau_S\right\|_\infty < 1$
3. Irrepresentability and Recovering β^0:

   Theorem 8.2. *Assume $X_S^T X_S$ is positive definite. Suppose for S_0 and γ the subgradient of $f(X) = \|X\|_1$ at $\hat{\beta}^L$, the weak irrepresentable condition holds.*

   (a) Then, the LASSO solution $\hat{\beta}^L_{N_0} = 0$
   (b) If, in addition, $|\beta^0_j| > \lambda\left\|n(X_S^T X_S)^{-1}\gamma_{S_0}\right\|_\infty$ then $\operatorname{sgn}(\hat{\beta}^L) = \operatorname{sgn}(\beta^0)$

   Proof. The proof of (a) relies on the KKT and using the fact that

$$A_0 = \frac{1}{n}(X_{N_0}^T X_{N_0} - X_{N_0}^T X_{S_0}(X_{S_0}^T X_{S_0})^{-1} X_{S_0}^T X_{N_0})$$

.

(a) By the KKT condition:

$$\frac{X^T X}{n}(\hat{\beta} - \beta^0) = -\frac{X^T}{n}(Y - X\hat{\beta}) = -\lambda\gamma$$

Then

$$\begin{aligned} -\lambda\gamma_{S_0} &= \frac{X_{S_0}^T X_{S_0}}{n}\left(\hat{\beta}_{S_0} - \beta_{S_0}^0\right) + \frac{X_{S_0}^T X_{N_0}}{n}\left(\hat{\beta}_{N_0}\right) \\ -\lambda n\left(X_{S_0}^T X_{S_0}\right)^{-1}\gamma_{S_0} &= \left(X_{S_0}^T X_{S_0}\right)^{-1} X_{S_0}^T X_{N_0}\left(\hat{\beta}_{N_0}\right) + \left(\hat{\beta}_{S_0} - \beta_{S_0}^0\right) \end{aligned}$$

And

$$-\lambda\gamma_{N_0} = \frac{X_{N_0}^T X_{N_0}}{n}\left(\hat{\beta}_{N_0}\right) + \frac{X_{N_0}^T X_{S_0}}{n}\left(\hat{\beta}_{S_0} - \beta_{S_0}^0\right)$$

(b) Since we are only interested in the behaviour of the $\hat{\beta}_{N_0}$ term, we can combine the two equations to remove the other term:

$$-\lambda\gamma_{N_0} + \lambda X_{N_0}^T X_{S_0}\left(X_{S_0}^T X_{S_0}\right)^{-1}\gamma_{S_0} = A_0\hat{\beta}_{N_0}$$

To use the fact that A_0 is positive semi-definite, we multiply by $\hat{\beta}_{N_0}^T$. Note that by definition of γ, $\hat{\beta}_{j,N_0}\gamma_{j,N_0} = |\hat{\beta}_{j,N_0}|$. Therefore:

$$-\lambda\left\|\hat{\beta}_{N_0}\right\|_1 + \lambda\hat{\beta}_{N_0}^T X_{N_0}^T X_{S_0}\left(X_{S_0}^T X_{S_0}\right)^{-1}\gamma_{S_0} = \hat{\beta}_{N_0}^T A_0\hat{\beta}_{N_0} \geq 0$$

(c) Using the weak irrepresentability condition and Cauchy-Schwarz:

$$\begin{aligned} &|\hat{\beta}_{N_0}^T X_{N_0}^T X_{S_0}\left(X_{S_0}^T X_{S_0}\right)^{-1}\gamma_{S_0}| \\ &\qquad \leq \left\|\hat{\beta}_{N_0}\right\|_1 \left\|X_{N_0}^T X_{S_0}\left(X_{S_0}^T X_{S_0}\right)^{-1}\gamma_{S_0}\right\|_\infty \\ &\qquad \leq \left\|\hat{\beta}_{N_0}\right\|_1 \end{aligned}$$

Therefore,

$$0 \leq \hat{\beta}_{N_0}^T A_0\hat{\beta}_{N_0} \leq -\lambda\left\|\hat{\beta}_{N_0}\right\|_1 + \lambda\left\|\hat{\beta}_{N_0}\right\|_1 = 0$$

Implying that $\hat{\beta}_{N_0} = 0$.

The second part requires the KKT and the result of the first part:

(a) Let $B_0 = \left(\frac{X_{S_0}^T X_{S_0}}{n}\right)^{-1}\gamma_{S_0}$.

(b) By KKT and the first result: $\hat{\beta}_{S_0} - \beta_{S_0}^0 = -\lambda B_0$. Therefore:

$$\left|\hat{\beta}_{S_0} - \beta_{S_0}^0\right| \leq \lambda\left\|B_0\right\|_\infty$$

Implying:

$$-\lambda\left\|B_0\right\|_\infty \leq \hat{\beta}_{S_0} - \beta_{S_0}^0 \leq \lambda\left\|B_0\right\|_\infty$$

(c) By assumption, $\left|\beta_j^0\right| > \lambda \left\|B_0\right\|_\infty$ for $j \in S \subseteq S_0$ (by the first result):

i. When $\beta_j^0 < 0$, then

$$\hat{\beta}_{j,S_0} < \lambda \left\|B_0\right\|_\infty + \beta_{j,S_0}^0 < \lambda \left\|B_0\right\|_\infty - \lambda \left\|B_0\right\|_\infty = 0$$

ii. When $\beta_j^0 > 0$, then

$$\hat{\beta}_{j,S_0} > -\lambda \left\|B_0\right\|_\infty + \beta_{j,S_0}^0 > -\lambda \left\|B_0\right\|_\infty + \lambda \left\|B_0\right\|_\infty = 0$$

□

4. An almost converse:

Theorem 8.3. *Suppose* $\operatorname{sgn}(\hat{\beta}^L) = \operatorname{sgn}(\beta^0)$ *then*

$$\left\|X_{N_0}^T X_{S_0} (X_{S_0}^T X_{S_0})^{-1} \gamma_{S_0}\right\|_\infty \leq 1$$

Proof. By assumption, $\hat{\beta}_{N_0} = \beta_{N_0}^0 = 0$.

(a) By KKT:

$$-\lambda X_{N_0}^T X_{S_0} \left(X_{S_0}^T X_{S_0}\right)^{-1} \gamma_{S_0} = \frac{X_{N_0}^T X_{S_0}}{n} \left(\hat{\beta}_{S_0} - \beta_{S_0}^0\right)$$
$$-\lambda \gamma_{N_0} = \frac{X_{N_0}^T X_{S_0}}{n} \left(\hat{\beta}_{S_0} - \beta_{S_0}^0\right)$$

(b) By definition $\left\|\gamma_{N_0}\right\|_\infty \leq 1$, and combining the two equations above:

$$\left\|X_{N_0}^T X_{S_0} \left(X_{S_0}^T X_{S_0}\right)^{-1} \gamma_{S_0}\right\|_\infty = \left\|\gamma_{N_0}\right\|_\infty \leq 1$$

□

8.5 Computing LASSO Paths

1. Coordinate Descent Algorithms

(a) Motivation: A very fast algorithm that finds $\hat{\beta}^L$ in a variety of LASSO extensions

(b) We can compute $\hat{\beta}_\lambda^L$ on a value of

$$\lambda \in \{0 \leq \lambda_{grid,1} < \lambda_{grid,2}, \ldots, \lambda_{grid,g}\}$$

(c) Algorithm

i. Let $\beta^{[0]} \in \mathbb{R}^p$. Let $m = 0$. Let $\mathcal{J}^{[0]} = 0$

ii. Repeat:

A. $m = m + 1$

B. $j = \mathcal{J}^{[m]} = \mathcal{J}^{[m-1]} + 1 \mod p$.

C. If $\left|G_j\left(\beta_{-j}^{[m-1]}\right)\right| \leq \lambda$ then $\beta_j^{[m]} = 0$

D. else $\beta_j^{[m]} = \arg\min_{\beta_j \in \mathbb{R}} Q_\lambda\left(\beta_{+j}^{[m]}\right)$

iii. Until numerical convergence

(d) Remarks

i. $\beta^{[0]}$ is arbitrary. $\mathcal{J}^{[m]} \in \{1, \ldots, p\}$ cycles through the coordinates.

ii. $Q_\lambda = \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$, the Criterion function.

iii. $G_j(\beta) = \frac{-1}{n} X^T (Y - X\beta)$, the derivative of the first component of the criterion function, which in the algorithm is compared to λ according to the KKT formulation of the LASSO problem.

iv. Note that $\beta_{-j}^{[m]}$ is $\beta^{[m]}$ with the j^{th} component set to 0.

v. Note that $\beta_{+j}^{[m]}$ is $\beta^{[m]}$ with the j^{th} component a variable β_j ranging over $\mathbb{R}$

vi. By KKT, if $\left|G_j\left(\beta_{-j}^{[m-1]}\right)\right| \leq \lambda$ then the j^{th} component is 0.

2. Theory shows that the algorithm does converge.

8.6 Extensions

1. Motivation: traditional LASSO penalises large coefficients, hence has a large absolute downward bias for large coefficients in magnitude.

2. Extended LASSO (simple)

   (a) Given a centred Y, $Q(\beta) = \frac{1}{2n} \|Y - X\beta\|_2^2 + \sum_{i=j}^{p} p_\lambda(|\beta_j|)$

   (b) p_λ is a non-negative non-decreasing, and continuously differentiable function on $(0, \infty)$. Moreover¡ $p_\lambda(0) = \lim_{t \to 0} p_\lambda(t) = 0$ and $p'_\lambda(0) = \lim_{t \to 0} p'_\lambda(t)$

3. Desirable properties of $\hat{\beta}$ given p_λ

   (a) Approximates unbiasedness: the estimator should be nearly unbiased for large coefficients in modulus

   (b) Sparsity: many components should be set to 0

   (c) Continuity: the estimator should be continuous with respect to X to avoid instability in prediction.

4. Convexity and Computation

   (a) Any $p_\lambda(t)$ that results in a parameter estimator with these properties cannot be convex, and can only guarantee local solutions

   (b) Local Linear Approximation:

      i. Note: $p_\lambda(|\beta|) \approx p_\lambda(|\beta^0|) + p'_\lambda(|\beta^0|)\left(|\beta| - |\beta^0|\right)$

      ii. Letting β^0 we can find local solutions numerically by finding:

$$\hat{\beta}^{k+1} \in \arg\min \frac{1}{2n} \|Y - X\beta\|_2^2 + \sum_{j=1}^{p} p'_\lambda(|\hat{\beta}_j^k|)|\beta_j|$$

5. Example of SCAD

Example 8.1. *The Smooth Absolutely Continuous Deviation (SCAD) Penalty is ideal for* $3 \leq a \leq 7$*:*

$$p_\lambda(0) = 0$$

$$p'_\lambda(t) = \lambda \mathbf{1}[\lambda \geq t] + \frac{a\lambda - t}{a-1}\mathbf{1}[\lambda < t]$$

6. Complicated LASSO Extensions
   (a) Pseudo-likelihood models: in generalised linear models or quantised regression models, we minimise $\sum_{i=1}^{n} L(Y_i, \sum_{j=1}^{p} x_{ij}\beta_j)$. To control variance we include a penalty function as above.
   (b) Group LASSO
      i. Consider a group model of the form $Y = \beta_0 \mathbf{1} + \sum_{j=1}^{J} X_j \beta_j + \epsilon$ where $X_j \in \mathbb{R}^{n \times p_j}$, $Y \in \mathbb{R}^{n \times 1}$, and $\beta_j \in \mathbb{R}^{p_j}$
      ii. The criterion function is:

$$Q(\beta^{GL}) = \frac{1}{2n}\left\| Y - \beta_0 \mathbf{1} - \sum_{j=1}^{J} X_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} p_j \left(\|\beta_j\|_1 \right)$$

      iii. $\hat{\beta}^{GL} = \left(\hat{\beta}_1^{GL}, \hat{\beta}_2^{GL}, \ldots, \hat{\beta}_J^{GL}\right)$ is sparse at the group level.
   (c) Additive Models
      i. Consider the nonparametric model $Y_i = m(X_i) + \epsilon$. This model becomes infeasible for large p and has very slow convergence with respect to n
      ii. Additive models assume $Y_i = m(X_i) + \epsilon = \sum_{j=1}^{p} m_j(x_{ij}) + \epsilon_i$ so the problem is reduced to a group LASSO

7. Stability Selection
   (a) Conduct LASSO regression on subsamples of the data
   (b) Select only variables that occur most frequently over all subsamples
   (c) This results in stability selection, which is closely related to bagging and sub-bagging (bagging = bootstrap aggregating).

Part IV
Multiple Hypothesis Testing

9 Framework

1. Motivation: In many applications, we are interested in testing many hypotheses simultaneously (e.g. variable selection).

2. General Problem

	Claimed Non-significant	Claimed Significant	Total
True Nulls	N_{00}	N_{01}	m_0
False Nulls	N_{10}	N_{11}	$m - m_0$
Total	$m - R$	R	m

(a) Suppose that we want to test the null hypotheses $H_1, \ldots, H_m$. We suppose that m_0 of these are actually true.

(b) Contingency Table

i. N_{00} are the true nulls that are correctly accepted

ii. N_{10} are the false nulls that are incorrectly accepted

iii. N_{01} are the true nulls that are incorrectly rejected

iv. R is the total number of nulls claimed to be false

10 Classical Theory: FWER and Bonferri

1. Definitions of Family Wise Error Rate and Bonferri Correction

   Definition 10.1. *The Family Wise Error Rate (FWER) is* $\mathbb{P}[N_{01} \geq 1]$. *The Bonferri Correction says that if the* i^{th} *p-value corresponding to hypothesis* H_i *is more extreme than* $\frac{\alpha}{m}$ *then we reject it.*

2. The FWER under Bonferri Correction

   Lemma 10.1. *The Bonferri Correction controls the FWER at level* α.

   Proof. Suppose that $H_1, \ldots, H_{m_0}$ are the true nulls with test statistics that have continuous distributions. Then the p-values, $p_1, \ldots, p_{m_0} \sim U(0,1)$. Therefore:

$$\mathbb{P}[N_{01} \geq 1] = \mathbb{P}\left[\bigcup_{i=1}^{m_0} \{p_i \leq \frac{\alpha}{m}\}\right] \leq \sum_{i=1}^{m_0} \mathbb{P}\left[p_i \leq \frac{\alpha}{m}\right] = \sum_{i=1}^{m_0} \frac{\alpha}{m} = \alpha$$

   □

11 False Discovery Proportion

1. Motivation: the Bonferri Correction is very conservative and has low power. And often, having a few true nulls claimed false does not necessarily invalidate the conclusions.

2. Definitions

   Definition 11.1. *Given* N_{01} *and* R *as above:*

   (a) False Discovery Proportion: $FDP = \frac{N_{01}}{R}$

   (b) False Discovery Rate: $FDR = \mathbb{E}[FDP]$

   (c) Benjamini-Hockberg Procedure: given an $\alpha > 0$, *order the p-values* $\{p_{(1)} \leq \cdots \leq p_{(m)}$. *Define* $k = \max\left(j | p_{(j)} \leq \frac{\alpha j}{m}\right)$. *The procedure rejects all* $H_{(1)}, \ldots, H_{(k)}$.

3. The Benjamini-Hockberg Procedure and Controlling the False Discovery Rate

Theorem 11.1. *Suppose $p_{(1)}, \ldots, p_{(m_0)}$ are identically $U(0,1)$ distributed and independent of $p_{(m_0+1)}, \ldots, p_{(m)}$. Then the Benjamini-Hockberg procedure controls the FDR at level α.*

Proof. Let $p_1, \ldots, p_m$ correspond to $H_1, \ldots, H_m$. And $p_{(1)} \leq \cdots \leq p_{(m)}$ be the ordered p-values.

(a) consider the following modified BH procedure applied to the set $P^{(1)} = \{p_{(2)} \leq \cdots \leq p_{(m)}\}$. Let $R^{(1)}$ be the number of rejections in $P^{(1)}$ where $k = \max\left(i | p^{(1)}_{(i)} \leq \frac{\alpha(i+1)}{m}\right)$. Hence, $R^{(1)} = k$.

(b) For $r = 1, \ldots, m$ consider the event:

$$\begin{aligned}
\{p_{(1)} \leq \frac{\alpha r}{m}\} \cap \{R = r\} &= \{p_{(1)} \leq \frac{\alpha r}{m}, R = r\} \\
&= \{p_1 \leq \frac{\alpha r}{m}, R = r, p_{(r)} \leq \frac{\alpha r}{m}, \forall s > r : \ p_{(s)} > \frac{\alpha s}{m}\} \\
&= \{p_1 \leq \frac{\alpha r}{m}, R = r, p^{(1)}_{(r-1)} \leq \frac{\alpha r}{m}, \forall s > r-1 : \ p^{(1)}_{(s)} > \frac{\alpha(s+1)}{m}\} \\
&= \{p_1 \leq \frac{\alpha r}{m}, R^{(1)} = r - 1\}
\end{aligned}$$

(c) Now we compute the FDR:

$$\begin{aligned}
FDR &= \sum_{r=1}^{m} \mathbb{E}\left[\frac{N_{01}}{r} \mathbf{1}\left[R = r\right]\right] \\
&= \sum_{r=1}^{m} \mathbb{E}\left[\frac{\mathbf{1}\left[R = r\right]}{r} \sum_{s=1}^{m_0} \mathbf{1}\left[p_s \leq \frac{\alpha r}{m}\right]\right] \\
&= \sum_{r=1}^{m} \frac{1}{r} \sum_{s=1}^{m_0} \mathbb{E}\left[\mathbf{1}\left[R = r\right] \mathbf{1}\left[p_s \leq \frac{\alpha r}{m}\right]\right] \\
&= \sum_{r=1}^{m} \frac{m_0}{r} \mathbb{E}\left[\mathbf{1}\left[R = r\right] \mathbf{1}\left[p_1 \leq \frac{\alpha r}{m}\right]\right] \\
&= \sum_{r=1}^{m} \frac{m_0}{r} \mathbb{E}\left[\mathbf{1}\left[R^{(1)} = r - 1\right] \mathbf{1}\left[p_1 \leq \frac{\alpha r}{m}\right]\right] \\
&= \sum_{r=1}^{m} \frac{m_0}{r} \mathbb{P}\left[R^{(1)} = r - 1\right] \mathbb{P}\left[p_1 \leq \frac{\alpha r}{m}\right] \\
&= \sum_{r=1}^{m} \frac{m_0}{r} \frac{\alpha r}{m} \mathbb{P}\left[R^{(1)} = r - 1\right] \\
&= \frac{m_0}{m} \alpha \leq \alpha
\end{aligned}$$

□

4. Extensions

   (a) let $\pi_0 = \frac{m_0}{m}$. Let $\hat{\pi}_0$ be an estimate to π_0. A more accurate, modified BH procedure has $k = \max\left(j | p_{(j)} \leq \frac{\alpha j}{\hat{\pi}_0 m}\right)$

   (b) One such estimator is $\hat{\pi}_0(\lambda) = \frac{|\{i|p_i > \lambda\}|}{(1-\lambda)m}$

      i. Note since the p_i are $U(0,1)$:

$$\mathbb{E}\left[\hat{\pi}_0(\lambda)\right] \geq \frac{1}{(1-\lambda)m}\mathbb{E}\left[|\{i|p_i > \lambda\}|\right] = \frac{m_0(1-\lambda)}{m(1-\lambda)}$$

      ii. As λ decreases, bias will decrease but variance will increase.