Notes for ISyE 7401

Alexander Shapiro

April 8, 2020

Contents

1 Some matrix calculus

Consider $m \times 1$ random vector $\boldsymbol{X} = (X_1, ..., X_m)'$. Its expected value $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$ is defined as $\mathbb{E}[\boldsymbol{X}] = (\mathbb{E}[X_1], ..., \mathbb{E}[X_m])'$, i.e., the expectation is taken componentwise. The $m \times m$ covariance matrix of $\boldsymbol{X}$ is

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'] = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}'] - \boldsymbol{\mu}\boldsymbol{\mu}'.$$

Recall that if $\boldsymbol{A}$ and $\boldsymbol{B}$ are two matrices such that their product $\boldsymbol{AB}$ is well defined, then the transpose of $\boldsymbol{AB}$ is $\boldsymbol{B}'\boldsymbol{A}'$, i.e., $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$.

Covariance matrix $\boldsymbol{\Sigma}$ has the following properties. It is symmetric, i.e., $\boldsymbol{\Sigma}' = \boldsymbol{\Sigma}$. Consider a (deterministic) $k \times m$ matrix $\boldsymbol{A}$ and $k \times 1$ random vector $\boldsymbol{Y} = \boldsymbol{AX}$. Then

$$\boldsymbol{\mu}_Y = \mathbb{E}[\boldsymbol{Y}] = \mathbb{E}[\boldsymbol{AX}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{X}] = \boldsymbol{A}\boldsymbol{\mu}_X.$$

In particular, if $k = 1$ and $Y = \boldsymbol{a}'\boldsymbol{X} = a_1X_1 + ... + a_mX_m$, where $\boldsymbol{a} = (a_1, ..., a_m)'$, then $\mathbb{E}[Y] = \boldsymbol{a}'\boldsymbol{\mu}_X$. Now

$$\boldsymbol{\Sigma}_Y = \mathbb{E}[\boldsymbol{Y}\boldsymbol{Y}'] - \boldsymbol{\mu}_Y\boldsymbol{\mu}_Y' = \mathbb{E}[\boldsymbol{AXXA}'] - \boldsymbol{A}\boldsymbol{\mu}_X\boldsymbol{\mu}_X'\boldsymbol{A}'.$$

Since $\mathbb{E}[\boldsymbol{AXXA}'] = \boldsymbol{A}\mathbb{E}[\boldsymbol{XX}]\boldsymbol{A}'$ it follows that

$$\boldsymbol{\Sigma}_Y = \boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}'. \tag{1.1}$$

In particular, if $\boldsymbol{A} = \boldsymbol{a}'$ is a row vector, then

$$\mathrm{Var}(\boldsymbol{a}'\boldsymbol{X}) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \sum_{i,j=1}^{m} \sigma_{ij}a_ia_j, \tag{1.2}$$

where $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ is the ij-component of matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_X$. Since variance of a random variable is always nonnegative, it follows that $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} \geq 0$ for any $m \times 1$ vector $\boldsymbol{a}$. Such (symmetric) matrices are called *positive semidefinite*. If moreover matrix $\boldsymbol{\Sigma}$ is nonsingular, then it is said that $\boldsymbol{\Sigma}$ is *positive definite*.

Matrix $\boldsymbol{\Sigma}$ is positive definite iff $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} > 0$ for all $\boldsymbol{a} \neq \boldsymbol{0}$. If $\boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \boldsymbol{0}$ for some $\boldsymbol{a} \neq \boldsymbol{0}$, then this means that $\mathrm{Var}(\boldsymbol{a}'\boldsymbol{X}) = 0$ and hence $Y = \boldsymbol{a}'\boldsymbol{X}$ is constant. This means that random variables $X_1 - \mu_1, ..., X_m - \mu_m$ are linearly dependent. Therefore $\boldsymbol{\Sigma}$ is positive definite iff $X_1 - \mu_1, ..., X_m - \mu_m$ are linearly independent.

Trace of a square matrix is defined as the sum of its diagonal elements, i.e., $\mathrm{tr}(\boldsymbol{A}) = a_{11} + ... + a_{mm}$. It has the following important property. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two matrices such that their product $\boldsymbol{AB}$ is well defined. Then

$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA}). \tag{1.3}$$

As an example let us compute expectation of $\boldsymbol{X}'\boldsymbol{AX} = \sum_{i,j=1}^{m} a_{ij}X_iX_j$, where $\boldsymbol{A}$ is an $m \times m$ matrix. Note that using property (1.3) we can write $\boldsymbol{X}'\boldsymbol{AX} = \mathrm{tr}(\boldsymbol{X}'\boldsymbol{AX}) = \mathrm{tr}(\boldsymbol{AXX}')$. Hence

$$\mathbb{E}[\boldsymbol{X}'\boldsymbol{AX}] = \mathbb{E}[\mathrm{tr}(\boldsymbol{AXX}')] = \mathrm{tr}(\boldsymbol{A}\mathbb{E}[\boldsymbol{XX}']) = \mathrm{tr}(\boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}. \tag{1.4}$$

2 Multivariate normal distribution

Recall that a random variable X has normal distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, if its probability density function (pdf) is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Now let $X_1, ..., X_m$ be an iid sequence of standard normal variables, i.e., $X_i \sim N(0,1)$, $i = 1, ..., m$, and these random variables are independent of each other. Then the pdf of random vector $\boldsymbol{X} = (X_1, ..., X_m)'$ is

$$f_X(\boldsymbol{x}) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \frac{1}{(2\pi)^{m/2}} e^{-\frac{x_1^2+...+x_m^2}{2}} = \frac{1}{(2\pi)^{m/2}} \exp(-\boldsymbol{x}'\boldsymbol{x}/2).$$

Consider $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{A}$ is an $m \times m$ nonsingular matrix. Note that $\boldsymbol{X} = \boldsymbol{A}^{-1}\boldsymbol{Y}$. Then the pdf of $\boldsymbol{Y}$ is

$$f_Y(\boldsymbol{y}) = f_X\left(\boldsymbol{A}^{-1}\boldsymbol{y}\right)|\boldsymbol{A}^{-1}| = \frac{1}{(2\pi)^{m/2}|\boldsymbol{A}|} \exp(-\boldsymbol{y}'\boldsymbol{A}'^{-1}\boldsymbol{A}^{-1}\boldsymbol{y}/2) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}_Y|^{1/2}} \exp(-\boldsymbol{y}'\boldsymbol{\Sigma}_Y^{-1}\boldsymbol{y}/2).$$

Here $|\boldsymbol{A}|$ denotes determinant of (square) matrix $\boldsymbol{A}$, and we used the properties that $|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$, $\boldsymbol{A}'^{-1}\boldsymbol{A}^{-1} = (\boldsymbol{A}\boldsymbol{A}')^{-1}$, and $\boldsymbol{\Sigma}_Y = \boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{A}$ since $\boldsymbol{\Sigma}_X = \boldsymbol{I}_m$ is the identity matrix, $|\boldsymbol{\Sigma}_Y| = |\boldsymbol{A}\boldsymbol{A}'| = |\boldsymbol{A}|\,|\boldsymbol{A}'| = |\boldsymbol{A}|^2$. Note also that $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{A}\boldsymbol{\mu}_X = \boldsymbol{0}$.

Finally consider $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$. The pdf of this random vector is

$$f_Y(\boldsymbol{y}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\big\{-(\boldsymbol{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})/2\big\}. \tag{2.1}$$

If random vector $\boldsymbol{Y}$ has pdf of the form (2.1), where $\boldsymbol{\Sigma}$ is a symmetric positive definite matrix, then it is said that $\boldsymbol{Y}$ has multivariate normal distribution, denoted $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Note that then $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{Y}$.

Suppose that $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is partitioned $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$ with the corresponding partitioning of $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$. Suppose futher that $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$ and hence $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}' = \boldsymbol{0}$, i.e., matrix $\boldsymbol{\Sigma}$ is block diagonal. Then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}|\,|\boldsymbol{\Sigma}_{22}|$, $\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix}$ and hence

$$f_X(\boldsymbol{x}) = f_{X_1}(\boldsymbol{x}_1) f_{X_2}(\boldsymbol{x}_2).$$

It follows that random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent. That is, for multivariate normal distribution "independent" and "uncorrelated" are equivalent.

Moment generating function of a random variable X is defined as $M(t) = \mathbb{E}[e^{tX}]$. Two random variables X and Y have the same distribution if their moment generating functions $M_X(t)$ and $M_Y(t)$ are equal to each other for all t in some neighborhood of zero (provided these moment generating functions are finite valued in this neighborhood). Similarly moment generating function of a random vector $\boldsymbol{X} = (X_1, ..., X_m)'$ is defined as $M(\boldsymbol{t}) = \mathbb{E}[e^{t_1X_1+...+t_mX_m}] = \mathbb{E}[\exp(\boldsymbol{t}'\boldsymbol{X})]$. If $M_X(\boldsymbol{t})$ is finite valued in a neighborhood of $\boldsymbol{0}$, then

$$\partial M_X(\boldsymbol{t})/\partial \boldsymbol{t}\big|_{t=0} = \mathbb{E}[\partial \exp(\boldsymbol{t}'\boldsymbol{X})/\partial \boldsymbol{t}]\big|_{t=0} = \mathbb{E}[\boldsymbol{X}\exp(\boldsymbol{t}'\boldsymbol{X})]_{t=0} = \mathbb{E}[\boldsymbol{X}].$$

Similarly the Hessian matrix of second order partial derivatives

$$\partial^2 M_X(\boldsymbol{t})/\partial \boldsymbol{t}\partial \boldsymbol{t}'\big|_{t=0} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}'].$$

Let us compute the moment generating function of $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{X})$. For standard normal random variable $X \sim N(0,1)$ we have

$$M(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(x-t)^2/2} e^{t^2/2} dx = e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} dx = e^{t^2/2}.$$

Let $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_m)$. It follows that components X_i of $\boldsymbol{X}$ are independent. Thus

$$M(\boldsymbol{t}) = \mathbb{E}[e^{t_1X_1+\ldots+t_mX_m}] = \mathbb{E}[e^{t_1X_1} \times \cdots \times e^{t_mX_m}] = \prod_{i=1}^{m} \mathbb{E}[e^{t_iX_i}] = \prod_{i=1}^{m} e^{t_i^2/2} = \exp(\boldsymbol{t}'\boldsymbol{t}/2).$$

Consider now $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$. Since $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_m)$ we have that $\boldsymbol{e}[\boldsymbol{Y}] = \boldsymbol{\mu}$ and its covariance matrix is $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}'$. Then

$$\begin{aligned} M_Y(\boldsymbol{t}) &= \mathbb{E}[\exp(\boldsymbol{t}'(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}))] = \mathbb{E}[\exp(\boldsymbol{t}'\boldsymbol{\mu})\exp(\boldsymbol{t}'\boldsymbol{A}\boldsymbol{X})] = \exp(\boldsymbol{t}'\boldsymbol{\mu})\mathbb{E}[\exp((\boldsymbol{A}'\boldsymbol{t})'\boldsymbol{X})] \\ &= \exp(\boldsymbol{t}'\boldsymbol{\mu})M_X(\boldsymbol{A}'\boldsymbol{t}) = \exp(\boldsymbol{t}'\boldsymbol{\mu})\exp(\boldsymbol{t}'\boldsymbol{A}\boldsymbol{A}'\boldsymbol{t}/2) = \exp(\boldsymbol{t}'\boldsymbol{\mu} + \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2). \end{aligned}$$

That is, for $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ its moment generating function is $M_Y(\boldsymbol{t}) = \exp(\boldsymbol{t}'\boldsymbol{\mu} + \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2)$.

Now let $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\eta}$. Then $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\eta}$, $\boldsymbol{\Sigma}_Y = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ and

$$M_Y(\boldsymbol{t}) = \exp(\boldsymbol{t}'(\boldsymbol{A}\boldsymbol{X} + \boldsymbol{\eta})) = \exp(\boldsymbol{t}'\boldsymbol{\eta})M_X(\boldsymbol{A}'\boldsymbol{t}) = \exp(\boldsymbol{t}'(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\eta}))\exp(\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'\boldsymbol{t}/2).$$

That is, $\boldsymbol{Y}$ has multivariate normal distribution with the respective mean and covariance matrix. A delicate point of this result is that the respective covariance matrix $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ should be nonsingular, i.e. *positive* definite. In particular, marginal distribution of every subvector of $\boldsymbol{X}$ is multivariate normal.

Exercise 1 *Show that $\boldsymbol{X}$ has multivariate normal distribution iff $Y = \boldsymbol{a}'\boldsymbol{X}$ is normally distributed for any vector $\boldsymbol{a} \neq \boldsymbol{0}$.*

Conditional normal distribution

Suppose that $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is partitioned $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$ with the corresponding partitioning of $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$. We want to compute the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$. Consider

$$\boldsymbol{Y} = \boldsymbol{X}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2 = \left[\boldsymbol{I}_{m_1},\ -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\right] \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}.$$

Note that vector $(\boldsymbol{Y}', \boldsymbol{X}_2')'$ has multivariate normal distribution. Moreover $\boldsymbol{X}_2 = [\boldsymbol{0}, \boldsymbol{I}_{m_2}] \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$ and

$$\mathrm{Cov}[\boldsymbol{Y}, \boldsymbol{X}_2] = \left[\boldsymbol{I}_{m_1},\ -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\right] \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{I}_{m_2} \end{bmatrix} = \left[\boldsymbol{I}_{m_1},\ -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\right] \begin{bmatrix} \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{22} \end{bmatrix} = \boldsymbol{0}.$$

It follows that $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are independent. Since $\boldsymbol{X}_1 = \boldsymbol{Y} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{X}_2$ it follows that the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is the same as the distribution of $\boldsymbol{Y} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{x}_2$. Now $\boldsymbol{Y}$ has multivariate normal distribution with mean

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_2$$

and covariance matrix

$$\boldsymbol{\Sigma}_Y = \left[\boldsymbol{I}_{m_1},\ -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\right] \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{m_1} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \end{bmatrix} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

That is the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is multivariate normal

$$N\left(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right). \tag{2.2}$$

2.1 Schur complement

Consider $(n+m)\times(n+m)$ matrix

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

where A, B, C, D are matrices of respective dimensions $n\times n$, $n\times m$, $m\times n$, $m\times m$, Suppose that D is invertible. Then

$$\begin{bmatrix} I_n & -BD^{-1} \\ 0 & I_m \end{bmatrix}\begin{bmatrix} A & B \\ C & D \end{bmatrix}\begin{bmatrix} I_n & 0 \\ -D^{-1}C & I_m \end{bmatrix} = \begin{bmatrix} A-BD^{-1}C & 0 \\ 0 & D \end{bmatrix}. \tag{2.3}$$

The matrix $A-BD^{-1}C$ is called the Schur complement of M with respect to D.

Note that

$$\begin{bmatrix} I_n & -BD^{-1} \\ 0 & I_m \end{bmatrix}^{-1} = \begin{bmatrix} I_n & BD^{-1} \\ 0 & I_m \end{bmatrix} \tag{2.4}$$

and

$$\begin{bmatrix} I_n & 0 \\ -D^{-1}C & I_m \end{bmatrix}^{-1} = \begin{bmatrix} I_n & 0 \\ D^{-1}C & I_m \end{bmatrix}. \tag{2.5}$$

Hence it follows from (2.3) that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I_n & BD^{-1} \\ 0 & I_m \end{bmatrix}\begin{bmatrix} A-BD^{-1}C & 0 \\ 0 & D \end{bmatrix}\begin{bmatrix} I_n & 0 \\ D^{-1}C & I_m \end{bmatrix}. \tag{2.6}$$

This implies that

$$\det(M) = \det(A-BD^{-1}C)\det(D). \tag{2.7}$$

Also it follows that matrix M is invertible iff the matrix $A-BD^{-1}C$ is invertible (recall that it is assumed that D is invertible), in which case

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} I_n & 0 \\ -D^{-1}C & I_m \end{bmatrix}\begin{bmatrix} (A-BD^{-1}C)^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix}\begin{bmatrix} I_n & -BD^{-1} \\ 0 & I_m \end{bmatrix}. \tag{2.8}$$

Using (2.8) it is possible to compute

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1}+D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}. \tag{2.9}$$

3 Quadratic forms

In this section we discuss distribution of quadratic forms $Q = \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{A}$ is an $m\times m$ symmetric matrix. Recall that the expected value of $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ was computed in (1.4). Let us first consider simple case where $\boldsymbol{A} = \boldsymbol{I}_m$ and $\boldsymbol{X} \sim N(\boldsymbol{0}, \boldsymbol{I}_m)$. Then $Q = X_1^2 + ... + X_m^2$ has chi-square distribution with m degrees of freedom, denoted $Q \sim \chi^2_m$. In order to proceed we need some matrix algebra.

Let $\boldsymbol{A}$ be an $m\times m$ symmetric matrix. Then it has m real valued eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$ and a corresponding set of eigenvectors $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$ such that

$$\boldsymbol{A}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i,\ i = 1, ..., m. \tag{3.1}$$

The eigenvectors can be chosen in such a way that $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$ for $i \neq j$ and $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$ for $i = 1, ..., m$, i.e., these eigenvectors are orthogonal to each other and of length one. The matrix $\boldsymbol{A}$ is positive semidefinite iff $\lambda_m \geq 0$ and is positive definite iff $\lambda_m > 0$.

Consider $m \times m$ matrix $\boldsymbol{T} = [\boldsymbol{e}_1, ..., \boldsymbol{e}_m]$ whose columns are formed from a set of orthonormal eigenvectors. Matrix $\boldsymbol{T}$ has the following property $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}_m$ and $\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{I}_m$. Such matrices are called *orthogonal*. Equations (3.1) can be written in the form $\boldsymbol{A}\boldsymbol{T} = \boldsymbol{T}\boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_m)$ is the diagonal matrix. By multiplying both sides of this matrix equation by $\boldsymbol{T}'$ we obtain

$$\boldsymbol{A} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}' = \sum_{i=1}^{m} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i'. \tag{3.2}$$

The representation (3.2) is called spectral decomposition of matrix $\boldsymbol{A}$.

It also follows that $\boldsymbol{T}'\boldsymbol{A}\boldsymbol{T} = \boldsymbol{\Lambda}$, and that $\text{tr}(\boldsymbol{A}) = \lambda_1 + \cdots + \lambda_m$ and $\boldsymbol{A}^{-1} = \boldsymbol{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{T}'$, provided that all $\lambda_i \neq 0$, $i = 1, ..., m$.

Now an $m \times m$ matrix $\boldsymbol{P}$ is said to be *idempotent* or *projection* matrix if $\boldsymbol{P}^2 = \boldsymbol{P}$. All eigenvalues of a projection matrix are either 1 or 0. If moreover $\boldsymbol{P}$ is symmetric, then $\boldsymbol{P} = \boldsymbol{T}_1\boldsymbol{T}_1'$, where $\boldsymbol{T}_1$ is an $m \times r$ matrix whose columns are orthonormal eigenvectors corresponding to eigenvalues 1, i.e., $\boldsymbol{T}_1'\boldsymbol{T}_1 = \boldsymbol{I}_r$. Then $\text{rank}(\boldsymbol{P}) = r = \text{tr}(\boldsymbol{P})$.

Theorem 3.1 *Let $\boldsymbol{X} \sim N_m(\boldsymbol{0}, \boldsymbol{\Sigma})$. Then $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2_m$.*

Proof. Consider spectral decomposition $\boldsymbol{\Sigma} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}'$ of the covariance matrix $\boldsymbol{\Sigma}$, and random vector $\boldsymbol{Y} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$, where $\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{T}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}'$. Note that $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{0}$, the covariance matrix of $\boldsymbol{Y}$ is $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{I}_m$ and $\boldsymbol{Y} \sim N(0, \boldsymbol{I}_m)$. Moreover $\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$. Hence $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = Y_1^2 + ... + Y_m^2 \sim \chi^2_m$. ∎

Theorem 3.2 *Let $\boldsymbol{X} \sim N_m(\boldsymbol{0}, \boldsymbol{I}_m)$ and $\boldsymbol{P}$ be symmetric projection matrix of rank r. Then $\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X} \sim \chi^2_r$.*

Proof. Consider spectral decomposition $\boldsymbol{P} = \boldsymbol{T}_1\boldsymbol{T}_1'$. Then $\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{T}_1\boldsymbol{T}_1'\boldsymbol{X} = \boldsymbol{Z}'\boldsymbol{Z}$, where $\boldsymbol{Z} = \boldsymbol{T}_1'\boldsymbol{X}$. We have that the $r \times 1$ vector $\boldsymbol{Z}$ has normal distribution with zero mean vector and covariance matrix $\boldsymbol{T}_1'\boldsymbol{T}_1 = \boldsymbol{I}_r$. It follows that $\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X} \sim \chi^2_r$. ∎

Noncentral chi square distribution

Let $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{I}_m)$ and consider $Q = \boldsymbol{X}'\boldsymbol{X} = X_1^2 + ... + X_m^2$. Note that if $\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{X}$, where $\boldsymbol{T}$ is an orthogonal matrix, then $\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}$ and $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{T}\boldsymbol{\mu}$ and the covariance matrix of $\boldsymbol{Y}$ is $\boldsymbol{I}_m$. It follows that the distribution of Q depends on $\delta = \mu_1^2 + ... + \mu_m^2$ rather than individual values of the components of the mean vector $\boldsymbol{\mu}$. Distribution of Q is called noncentral chi square with the noncentrality parameter $\delta = \mu_1^2 + ... + \mu_m^2$ and m degrees of freedom, denoted $Q \sim \chi^2_m(\delta)$. Similar to Theorem 3.1 it is possible to show the following.

Exercise 2 *Show that if $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2_m(\delta)$ with noncentrality parameter $\delta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. If $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{I}_m)$ and $\boldsymbol{P}$ is symmetric projection matrix of rank r, then $\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X} \sim \chi^2_r(\delta)$, where $\delta = \boldsymbol{\mu}'\boldsymbol{P}\boldsymbol{\mu}$.*

4 Statistical inference of linear models

Consider linear regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + ...\beta_k x_{ik} + \varepsilon_i, \ i = 1, ..., n. \tag{4.1}$$

In matrix form we can write this model as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.2}$$

where $\boldsymbol{Y} = (Y_1, ..., Y_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)'$ and $\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$ is $n \times p$, $p = k + 1$, design matrix. Note that the first column of $\boldsymbol{X}$ is formed by ones. We assume that $\boldsymbol{X}$ has full column rank p.

The least squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, provided columns of $\boldsymbol{X}$ are linearly independent. Suppose that $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$. Then (recall that the design matrix $\boldsymbol{X}$ is assumed to be deterministic)

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{Y}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}.$$

That is, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Consider the $n \times n$ matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Note that vector $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ of fitted values is given by $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$, and vector of residuals $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is given by $\boldsymbol{e} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$.

Exercise 3 *Show that matrix $\boldsymbol{H}$ has the following properties:*

(i) $\boldsymbol{H}$ is symmetric.

(ii) $\boldsymbol{H}$ and $\boldsymbol{I}_n - \boldsymbol{H}$ are idempotent (projection) matrices, i.e. $\boldsymbol{H}^2 = \boldsymbol{H}$ and $(\boldsymbol{I}_n - \boldsymbol{H})^2 = \boldsymbol{I}_n - \boldsymbol{H}$.

(iii) $\mathrm{tr}(\boldsymbol{H}) = p$ and $\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = n - p$.

(iv) $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$.

Suppose that the errors ε_i, are uncorrelated, $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, ..., n$. That is, $\mathrm{Cov}(\boldsymbol{Y}) = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_n$. Then the covariance matrix of $\hat{\boldsymbol{\beta}}$ can be computed as

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\mathrm{Cov}(\boldsymbol{Y})]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

It also follows that the covariance matrix of $\boldsymbol{e}$ is $\sigma^2(\boldsymbol{I}_n - \boldsymbol{H})$, and $\mathbb{E}[\boldsymbol{e}] = \boldsymbol{0}$ and

$$\mathbb{E}\left[\sum_{i=1}^{n}(e_1^2 + ... + e_n^2)\right] = \sum_{i=1}^{n}\mathrm{Var}(e_i) = \sigma^2\mathrm{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = \sigma^2(n - p).$$

That is, $S^2 = (n-p)^{-1}\sum_{i=1}^{n}(e_1^2 + ... + e_n^2)$ is an unbiased estimator of σ^2.

Also $\boldsymbol{e}'\boldsymbol{X} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$. Since the first column of $\boldsymbol{X}$ is vector $\boldsymbol{1}_n$, it follows that $\boldsymbol{e}'\boldsymbol{1}_n = 0$, that is $\sum_{i=1}^{n} e_i = 0$. In similar way we have $\boldsymbol{e}'\hat{\boldsymbol{Y}} = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H}\boldsymbol{Y} = 0$, and hence $\boldsymbol{e}$ and $\hat{\boldsymbol{Y}}$ are uncorrelated.

Theorem 4.1 (Gauss - Markov) *Suppose that $\mathbb{E}[\boldsymbol{\varepsilon}] = \boldsymbol{0}$ and $\mathrm{Cov}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{I}_n$. Then the LSE estimator $\hat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\beta}$. That is, if $\tilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{Y}$ is a linear unbiased estimator of $\boldsymbol{\beta}$ (i.e., $\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ for all $\boldsymbol{\beta}$), then*

$$\mathrm{Var}(\boldsymbol{a}'\tilde{\boldsymbol{\beta}}) \geq \mathrm{Var}(\boldsymbol{a}'\hat{\boldsymbol{\beta}}) \tag{4.3}$$

for any $p \times 1$ vector $\boldsymbol{a}$.

Proof. Since $\mathbb{E}[\tilde{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ for *all* $\boldsymbol{\beta}$, it follows that $\boldsymbol{\beta} = \boldsymbol{A}'\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{A}'\boldsymbol{X}\boldsymbol{\beta}$. Hence $(\boldsymbol{I}_p - \boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{0}$ for all $\boldsymbol{\beta}$, and thus $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_p$. Consider matrix $\boldsymbol{B} = \boldsymbol{A} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Note that since $\boldsymbol{X}'\boldsymbol{A} = \boldsymbol{I}_p$, it follows that

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{A} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \boldsymbol{0},$$

and hence $\boldsymbol{X}'\boldsymbol{B} = \boldsymbol{0}$.

Now since covariance matrix of $\boldsymbol{Y}$ is $\sigma^2\boldsymbol{I}_n$ it follows that

$$\mathrm{Var}(\boldsymbol{a}'\tilde{\boldsymbol{\beta}}) = \mathrm{Var}(\boldsymbol{a}'\boldsymbol{A}'\boldsymbol{Y}) = \sigma^2\boldsymbol{a}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{a}.$$

Also since $\boldsymbol{X}'\boldsymbol{B} = \boldsymbol{0}$ we have that

$$\boldsymbol{A}'\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1})\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{B}'\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{B}'\boldsymbol{B}.$$

Hence

$$\mathrm{Var}(\boldsymbol{a}'\tilde{\boldsymbol{\beta}}) = \sigma^2\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a} + \sigma^2\boldsymbol{a}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{a} = \mathrm{Var}(\boldsymbol{a}'\hat{\boldsymbol{\beta}}) + \sigma^2\boldsymbol{a}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{a}.$$

It remains to note that $\boldsymbol{a}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{a} = (\boldsymbol{B}\boldsymbol{a})'\boldsymbol{B}\boldsymbol{a} \geq 0$. ∎

Condition number Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix. Then

$$\max_{\|\boldsymbol{x}\|=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \lambda_{\max}(\boldsymbol{A}), \tag{4.4}$$

where $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}'\boldsymbol{x}}$ and $\lambda_{\max}(\boldsymbol{A})$ is the largest eigenvalue of $\boldsymbol{A}$. This can be shown by using the spectral decomposition of $\boldsymbol{A}$.

Now consider the system of linear equations $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, where $\boldsymbol{A}$ is a nonsingular $n \times n$ matrix (not necessarily symmetric) and $\boldsymbol{b}$ is $n \times 1$ nonzero vector. It has solution $\boldsymbol{x}_0 = \boldsymbol{A}^{-1}\boldsymbol{b}$. Consider perturbed system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a "small" vector of errors. This system has solution $\boldsymbol{x}_\varepsilon = \boldsymbol{x}_0 + \boldsymbol{A}^{-1}\boldsymbol{\varepsilon}$. Consider the following ratio of the relative error in the solution to the relative error in $\boldsymbol{b}$

$$\frac{\|\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\|/\|\boldsymbol{x}_0\|}{\|\boldsymbol{\varepsilon}\|/\|\boldsymbol{b}\|} = \frac{\|\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\|}{\|\boldsymbol{\varepsilon}\|} \cdot \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}^{-1}\boldsymbol{b}\|}.$$

The following maximum is called the conditional number of $\boldsymbol{A}$:

$$\mathrm{cond}(\boldsymbol{A}) = \max_{\boldsymbol{b}\neq\boldsymbol{0},\boldsymbol{\varepsilon}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\|}{\|\boldsymbol{\varepsilon}\|} \cdot \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}^{-1}\boldsymbol{b}\|} = \left(\max_{\boldsymbol{\varepsilon}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\|}{\|\boldsymbol{\varepsilon}\|}\right)\left(\max_{\boldsymbol{b}\neq\boldsymbol{0}} \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}^{-1}\boldsymbol{b}\|}\right).$$

Now

$$\max_{\boldsymbol{\varepsilon}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}^{-1}\boldsymbol{\varepsilon}\|}{\|\boldsymbol{\varepsilon}\|} = \max_{\|\boldsymbol{\varepsilon}\|=1} \sqrt{\boldsymbol{\varepsilon}'(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{\varepsilon}} = \frac{1}{\sigma_{\min}(\boldsymbol{A})},$$

where $\sigma_{\min}(\boldsymbol{A}) = \sqrt{\lambda_{\min}(\boldsymbol{A}'\boldsymbol{A})}$ is the minimal singular value of $\boldsymbol{A}$. Similarly

$$\max_{\boldsymbol{b}\neq\boldsymbol{0}} \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}^{-1}\boldsymbol{b}\|} = \max_{\boldsymbol{z}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{z}\|}{\|\boldsymbol{z}\|} = \sigma_{\max}(\boldsymbol{A}).$$

Therefore $\mathrm{cond}(\boldsymbol{A}) = \sigma_{\max}(\boldsymbol{A})/\sigma_{\min}(\boldsymbol{A})$. If matrix $\boldsymbol{A}$ is symmetric positive definite, then $\mathrm{cond}(\boldsymbol{A}) = \lambda_{\max}(\boldsymbol{A})/\lambda_{\min}(\boldsymbol{A})$.

4.1 Distribution theory

Suppose now that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ and hence $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$. It follows that the LSE estimator $\hat{\boldsymbol{\beta}}$ has normal distribution $N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$. Hence it follows by Theorem 3.1 that

$$\sigma^{-2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi^2_p, \tag{4.5}$$

where $p = k + 1$.

Also since $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X} = \mathbf{0}$,

$$\frac{(n-p)S^2}{\sigma^2} = \frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} = \frac{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})^2\boldsymbol{Y}}{\sigma^2} = \frac{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{\varepsilon}}{\sigma^2}.$$

Recall that $\boldsymbol{I}_n - \boldsymbol{H}$ is a projection matrix. Its rank

$$\operatorname{rank}(\boldsymbol{I}_n - \boldsymbol{H}) = \operatorname{tr}(\boldsymbol{I}_n - \boldsymbol{H}) = \operatorname{tr}(\boldsymbol{I}_n) - \operatorname{tr}(\boldsymbol{H}) = n - \operatorname{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = n - \operatorname{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = n - p.$$

By Theorem 3.2 it follows that

$$\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}. \tag{4.6}$$

Moreover

$$\operatorname{Cov}[\boldsymbol{e}, \hat{\boldsymbol{\beta}}] = (\boldsymbol{I}_n - \boldsymbol{H})\operatorname{Cov}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \mathbf{0}.$$

Hence $\boldsymbol{e}$ and $\hat{\boldsymbol{\beta}}$ are independent. It follows that S^2 and $\hat{\boldsymbol{\beta}}$ are independent, and hence S^2 and $\sigma^{-2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ are independent. It follows that

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/p}{S^2} \sim F_{p,n-p}. \tag{4.7}$$

This can be used to construct the following $(1-\alpha)$-confidence region for $\boldsymbol{\beta}$:

$$\left\{\boldsymbol{\beta} : (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \le pS^2 F_{\alpha;p,n-p}\right\}.$$

Now consider $S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2$, $SS_R = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ and $SS_E = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2$. Let us show that

$$S_{YY} = SS_R + SS_E. \tag{4.8}$$

Indeed, we have that

$$S_{YY} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$$

and

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum_{i=1}^n \hat{Y}_i e_i + \bar{Y}\sum_{i=1}^n e_i = 0.$$

Also

$$(n-p)S^2 = SS_E = \boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} \tag{4.9}$$

and

$$SS_R = (\hat{\boldsymbol{Y}} - \mathbf{1}_n\bar{Y})'(\hat{\boldsymbol{Y}} - \mathbf{1}_n\bar{Y}) = (\boldsymbol{H}\boldsymbol{Y} - n^{-1}\mathbf{1}_n\mathbf{1}_n'\boldsymbol{Y})'(\boldsymbol{H}\boldsymbol{Y} - n^{-1}\mathbf{1}_n\mathbf{1}_n'\boldsymbol{Y}) = \boldsymbol{Y}'(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')^2\boldsymbol{Y}.$$

Moreover, since $\boldsymbol{H}\mathbf{1}_n = \mathbf{1}_n$ (this holds since $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$ and the first column of $\boldsymbol{X}$ is $\mathbf{1}_n$) we obtain $(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')^2 = \boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n'$. and hence

$$SS_R = \boldsymbol{Y}'(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{Y}. \tag{4.10}$$

Since $(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H} = \mathbf{0}$ and $(\boldsymbol{I}_n - \boldsymbol{H})\mathbf{1}_n = \mathbf{0}$, we have that

$$(\boldsymbol{I}_n - \boldsymbol{H})(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n') = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{H} - n^{-1}(\boldsymbol{I}_n - \boldsymbol{H})\mathbf{1}_n\mathbf{1}_n' = \mathbf{0},$$

and hence SS_E and SS_R are independent.

Consider the following so-called F-statistic, for testing $H_0 : \beta_1 = ... = \beta_k = 0$,

$$F = \frac{SS_R/k}{SS_E/(n-p)}. \tag{4.11}$$

Recall that $SS_E/\sigma^2 \sim \chi^2_{n-p}$. Also under H_0 we have that $\boldsymbol{Y} = \beta_0\mathbf{1}_n$ and hence

$$(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{Y} = \beta_0(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{1}_n = \beta_0(\boldsymbol{H}\mathbf{1}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{1}_n) = \mathbf{1}_n - \mathbf{1}_n = \mathbf{0}.$$

Consequently

$$SS_R = \boldsymbol{\varepsilon}'(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{\varepsilon},$$

and hence $SS_R/\sigma^2 \sim \chi^2_k$. Note that

$$\text{rank}(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n') = \text{tr}(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n') = \text{tr}(\boldsymbol{H}) - 1 = k.$$

It follows that under H_0 the statistic F has $F_{k,n-p}$ distribution.

Under alternative H_1, SS_R/σ^2 has noncentral chi square distribution $SS_R/\sigma^2 \sim \chi^2_k(\delta)$ with noncentrality parameter

$$\delta = \sigma^{-2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')\boldsymbol{X}\boldsymbol{\beta} = \sigma^{-2}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} - n^{-1}(\mathbf{1}_n'\boldsymbol{X})'(\mathbf{1}_n'\boldsymbol{X}))\boldsymbol{\beta}.$$

4.2 Estimation with linear restrictions

Suppose that we want to test linear constraints $\boldsymbol{a}_i'\boldsymbol{\beta} = c_i$, $i = 1, ..., q$. We can write this as $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$, where $\boldsymbol{A}$ is the corresponding $q \times p$ matrix whose rows are formed from vectors $\boldsymbol{a}_i'$, $i = 1, ..., q$, and $\boldsymbol{c} = (c_1, ..., c_q)'$. We assume that vectors $\boldsymbol{a}_i$, $i = 1, ..., q$, are linearly independent, i.e., matrix $\boldsymbol{A}$ has rank q.

The respective constrained least squares estimator is obtained as a solution of the following optimization problem

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \text{ subject to } \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}. \tag{4.12}$$

Consider the Lagrangian of the above problem (4.12):

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\lambda}) &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\sum_{i=1}^{q} \lambda_i(\boldsymbol{a}_i'\boldsymbol{\beta} - c_i) \\ &= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}'(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{c}). \end{aligned}$$

Optimality conditions for problem (4.12) can be written as $\partial L(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial\boldsymbol{\beta} = \mathbf{0}$ and $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$. Note that

$$\partial L(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial\boldsymbol{\beta} = -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + 2\boldsymbol{A}'\boldsymbol{\lambda}.$$

Hence the optimality conditions can be written as the following system of linear equations

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{A'} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'Y} \\ \boldsymbol{c} \end{bmatrix}. \tag{4.13}$$

The corresponding estimators are given by

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_H \\ \hat{\boldsymbol{\lambda}}_H \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{A'} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X'Y} \\ \boldsymbol{c} \end{bmatrix}. \tag{4.14}$$

By using formula (2.9) for the inverse $\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{A'} \\ \boldsymbol{A} & \boldsymbol{0} \end{bmatrix}^{-1}$, after some algebraic calculations it is possible to write the estimate $\hat{\boldsymbol{\beta}}_H$ in the following form

$$\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}} + (\boldsymbol{X'X})^{-1}\boldsymbol{A'}[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\beta}}), \tag{4.15}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$. Geometrical interpretation, consider $\hat{\boldsymbol{Y}}_H = \boldsymbol{X}\hat{\boldsymbol{\beta}}_H$. Then $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is orthogonal to $\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H$,

$$\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\beta}}),$$

and

$$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_H\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H\|^2.$$

Statistic for testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ is

$$F = \frac{(SS_E(H) - SS_E(F))/q}{SS_E(F)/(n-p)}, \tag{4.16}$$

where $SS_E(F)$ is the sum of squares of residuals of the full (unconstrained) model and $SS_E(H)$ is the sum of squares of residuals of the reduced (constrained) model. We have here that

$$SS_E(H) - SS_E(F) = (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c}). \tag{4.17}$$

Recall that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1})$, and hence $\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim N(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{c}, \sigma^2\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'})$. It follows that under the H_0 hypothesis, $[SS_E(H) - SS_E(F)]/\sigma^2 \sim \chi^2_q$.

Under the H_0 hypothesis, the above F statistic has distribution $F_{q,n-p}$.

4.3 Polynomial Regression

Consider the polynomial regression model (one predictor)

$$Y_i = \beta_0 + \beta_1 x_i + ...\beta_k x_i^k + \varepsilon_i,\ i = 1, ..., n. \tag{4.18}$$

We can formulate this as the linear multivariate model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with the corresponding design matrix $\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^k \\ & & & \\ & \cdots & \cdots & \cdots \\ 1 & x_n & \cdots & x_n^k \end{bmatrix}$. We have here

$$[\boldsymbol{X'X}]_{st} = n\sum_{i=1}^{n} x_i^{s+t} \approx n\int_0^1 x^{s+t}dx = \frac{n}{s+t+1},\ s, t = 0, ..., k.$$

That is

$$\boldsymbol{X}'\boldsymbol{X} \approx n \begin{bmatrix} 1 & 1/2 & 1/3 & \cdots & 1/(k+1) \\ 1/2 & 1/3 & 1/4 & \cdots & 1/(k+2) \\ 1/3 & 1/4 & 1/5 & \cdots & 1/(k+3) \\ & \cdots & \cdots & \cdots & \\ 1/(k+1) & 1/(k+2) & 1/(k+3) & \cdots & 1/(2k+1) \end{bmatrix}.$$

This matrix is ill conditioned.

Chebishev polynomials

$$T_n(x) = \cos[n(\arccos x)], \quad -1 \le x \le 1.$$

Let $\theta = \arccos x$. Then

$$\begin{aligned} T_0(x) &= \cos 0 = 1, \\ T_1(x) &= \cos\theta = x, \\ T_2(x) &= \cos(2\theta) = 2\cos^2\theta - 1 = 2x^2 - 1. \end{aligned}$$

Recall that

$$\cos(m+1)\theta + \cos(m-1)\theta = 2\cos\theta\cos m\theta.$$

It follows that

$$T_{m+1}(x) + T_{m-1}(x) = 2xT_m(x),$$

and hence

$$T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x).$$

For example

$$T_3(x) = 2xT_2(x) - T_1(x) = 2x(2x^2-1) - x.$$

Note that, by $d\arccos x = \frac{1}{\sqrt{1-x^2}}dx$,

$$\int_{-1}^{1} \frac{T_k(x)T_\ell(x)}{\sqrt{1-x^2}}dx = \int_{-1}^{1} \cos(k\theta)\cos(\ell\theta)d\theta = 0, \text{ for } k \ne \ell.$$

Hence for $x_i = \cos(\pi/n)i$ and $\theta_i = (\pi/n)i$ we have

$$\sum_{i=0}^{n-1} T_k(x_i)T_\ell(x_i) = \sum_{i=0}^{n-1} \cos k\theta_i \cos \ell\theta_i = 0.$$

For

$$Y_i = \beta_0 T_0(x_i) + \beta_1 T_1(x_i) + ... + \beta_k T_k(x_i) + \varepsilon_i, \quad i = 1, ..., n,$$

we have that

$$\boldsymbol{X} = \begin{bmatrix} T_0(x_1) & \cdots & T_k(x_1) \\ & \cdots & \\ T_0(x_n) & \cdots & T_k(x_n) \end{bmatrix}.$$

5 Shrinkage Methods

For a vector $\boldsymbol{x} \in \mathbb{R}^m$ the ℓ_q norm is $\|\boldsymbol{x}\|_q = (|x_1|^q + ... + |x_m|^q)^{1/q}$, $q \ge 1$. In particular, the ℓ_2 norm is the Euclidean norm, and ℓ_1 norm is $\|\boldsymbol{x}\|_1 = |x_1| + ... + |x_m|$. Note that function $\|\cdot\|_q$ is positively homogeneous (i.e., $\|\lambda\boldsymbol{x}\|_q = \lambda\|\boldsymbol{x}\|_q$ for $\lambda \ge 0$) for any $q > 0$. However for $q \in (0,1)$, $\|\cdot\|_q$ is not convex.

5.1 Ridge Regression

Consider the following approach, called Ridge Regression, to estimation parameters of the linear model (4.2)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\varepsilon\|\boldsymbol{\beta}\|_2^2, \tag{5.1}$$

where $\varepsilon>0$. Solution $\tilde{\boldsymbol{\beta}}_\varepsilon$ of this problem satisfies optimality conditions

$$-\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})+\varepsilon\boldsymbol{\beta}=\boldsymbol{0}.$$

That is $\tilde{\boldsymbol{\beta}}_\varepsilon=(\boldsymbol{X}'\boldsymbol{X}+\varepsilon\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ (recall that $p=k+1$ is the number of estimated parameters). Of course for $\varepsilon=0$ the estimator $\tilde{\boldsymbol{\beta}}_\varepsilon$ coincides with Least Squares (LS) estimator $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. It is also possible to formulate problem (5.1) in the following form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \text{ s.t. } \|\boldsymbol{\beta}\|_2\le c, \tag{5.2}$$

for a certain value of $c>0$ (take $c=\|\tilde{\boldsymbol{\beta}}_\varepsilon\|_2$). Conversely solution of problem (5.2), for some $c>0$, is also the solution of problem (5.1) when ε is the corresponding Lagrange multiplier. (If $\|\hat{\boldsymbol{\beta}}\|_2\le c$, then the corresponding $\varepsilon=0$.) Therefore in a sense problems (5.1) and (5.2) are equivalent to each other for a proper choice of the respective positive constants ε and c.

The estimator $\tilde{\boldsymbol{\beta}}_\varepsilon$ shrinks the LS estimator to the origin. In particular if columns of the design matrix $\boldsymbol{X}$ are orthogonal, i.e., matrix $\boldsymbol{X}'\boldsymbol{X}=\mathrm{diag}(\lambda_1,...,\lambda_p)$ is diagonal. Then $\boldsymbol{X}'\boldsymbol{X}+\varepsilon\boldsymbol{I}_p=\mathrm{diag}(\lambda_1+\varepsilon,...,\lambda_p+\varepsilon)$ and $\tilde{\beta}_{\varepsilon,i}=(1+\varepsilon/\lambda_i)^{-1}\hat{\beta}_i$. Let $\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}'$ be the spectral decomposition of matrix $\boldsymbol{X}'\boldsymbol{X}$, with $\lambda_1\ge\cdots\ge\lambda_p>0$ being the eigenvalues and $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,...,\lambda_p)$. Then $\boldsymbol{X}'\boldsymbol{X}+\varepsilon\boldsymbol{I}_p=\boldsymbol{T}(\boldsymbol{\Lambda}+\varepsilon\boldsymbol{I}_p)\boldsymbol{T}'$. Number λ_1/λ_p is called the condition number of matrix $\boldsymbol{X}'\boldsymbol{X}$. The condition number of matrix $\boldsymbol{X}'\boldsymbol{X}+\varepsilon\boldsymbol{I}_p$ is $(\lambda_1+\varepsilon)/(\lambda_p+\varepsilon)$, and can be much smaller than λ_1/λ_p even for small values of $\varepsilon>0$. Moreover $\tilde{\boldsymbol{\beta}}_\varepsilon=\boldsymbol{T}(\boldsymbol{\Lambda}+\varepsilon\boldsymbol{I}_p)^{-1}\boldsymbol{T}'\boldsymbol{X}'\boldsymbol{Y}$, and hence

$$\tilde{\boldsymbol{\gamma}}_\varepsilon=(\boldsymbol{\Lambda}+\varepsilon\boldsymbol{I}_p)^{-1}\tilde{\boldsymbol{X}}'\boldsymbol{Y},$$

where $\tilde{\boldsymbol{\gamma}}_\varepsilon=\boldsymbol{T}'\tilde{\boldsymbol{\beta}}_\varepsilon$ and $\tilde{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{T}$. Note that $\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}=\boldsymbol{\Lambda}$ and hence $\tilde{\gamma}_{\varepsilon,i}=(1+\varepsilon/\lambda_i)^{-1}\hat{\gamma}_i$, where $\hat{\boldsymbol{\gamma}}$ is the LS estimator of the corresponding linear model with $\boldsymbol{X}$ replaces by $\tilde{\boldsymbol{X}}$.

The estimator $\tilde{\boldsymbol{\beta}}_\varepsilon$ is biased, that is $\mathbb{E}[\tilde{\boldsymbol{\beta}}_\varepsilon]=(\boldsymbol{X}'\boldsymbol{X}+\varepsilon\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$. It is possible to show that there exists $\varepsilon>0$ such that the components of $\tilde{\boldsymbol{\beta}}_\varepsilon$ have smaller mean square error than the respective components of the LS estimator $\hat{\boldsymbol{\beta}}$. That is, let $\theta=\boldsymbol{a}'\boldsymbol{\beta}$ for some given vector $\boldsymbol{a}\ne\boldsymbol{0}$, and let $\tilde{\theta}_\varepsilon=\boldsymbol{a}'\tilde{\boldsymbol{\beta}}_\varepsilon$ and $\hat{\theta}=\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ be estimators of θ. We show that there exists $\varepsilon>0$ such that

$$MSE(\tilde{\theta}_\varepsilon)<MSE(\hat{\theta}),$$

where $MSE(\tilde{\theta})=\mathbb{E}[(\tilde{\theta}-\theta)^2]$ is the mean square error of an estimator $\tilde{\theta}$. Recall that

$$\tilde{\boldsymbol{\beta}}_\varepsilon=(\boldsymbol{X}'\boldsymbol{X}+\varepsilon\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y}=[\boldsymbol{I}_p+\varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\hat{\boldsymbol{\beta}},$$

and hence

$$\begin{aligned}\mathbb{E}[\tilde{\theta}_\varepsilon] &= \boldsymbol{a}'[\boldsymbol{I}_p+\varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\boldsymbol{\beta}\\ &= \boldsymbol{a}'[\boldsymbol{I}_p-\varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{\beta}+o(\varepsilon)\\ &= \theta-\varepsilon\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\beta}+o(\varepsilon),\end{aligned}$$

where $o(\varepsilon)/\varepsilon\to 0$ as $\varepsilon\to 0$. This uses the geometric series expansion $(\boldsymbol{I}+\boldsymbol{A})^{-1}=\boldsymbol{I}-\boldsymbol{A}+\boldsymbol{A}^2-...=\boldsymbol{I}-\boldsymbol{A}+o(\|\boldsymbol{A}\|)$, which holds for matrix $\boldsymbol{A}$ sufficiently small. It follows that

$$\mathrm{Bias}[\tilde{\theta}_\varepsilon]=\mathbb{E}[\tilde{\theta}_\varepsilon]-\theta=-\varepsilon\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\beta}+o(\varepsilon),$$

and hence $\mathrm{Bias}^2[\tilde{\theta}_\varepsilon] = o(\varepsilon)$.

We also have that

$$\begin{aligned}\mathrm{Var}[\tilde{\theta}_\varepsilon] &= \sigma^2 \boldsymbol{a}'[\boldsymbol{I}_p + \varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}(\boldsymbol{X}\boldsymbol{X}')^{-1}[\boldsymbol{I}_p + \varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\boldsymbol{a} \\ &= \sigma^2 \boldsymbol{a}'[\boldsymbol{I}_p - \varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}](\boldsymbol{X}\boldsymbol{X}')^{-1}[\boldsymbol{I}_p - \varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}]\boldsymbol{a} + o(\varepsilon) \\ &= \sigma^2 \boldsymbol{a}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{a} - 2\varepsilon\sigma^2 \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-2}\boldsymbol{a} + o(\varepsilon) \\ &= \mathrm{Var}[\hat{\theta}] - 2\varepsilon\sigma^2 \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-2}\boldsymbol{a} + o(\varepsilon).\end{aligned}$$

Therefore

$$MSE(\hat{\theta}) - MSE(\tilde{\theta}_\varepsilon) = \mathrm{Var}[\hat{\theta}] - \mathrm{Var}[\tilde{\theta}_\varepsilon] - \mathrm{Bias}^2[\tilde{\theta}_\varepsilon] = 2\varepsilon\sigma^2 \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-2}\boldsymbol{a} + o(\varepsilon).$$

Since matrix $\boldsymbol{X}'\boldsymbol{X}$ is positive definite, and hence $(\boldsymbol{X}'\boldsymbol{X})^{-2}$ is positive definite, and $\boldsymbol{a} \neq \boldsymbol{0}$, we have that $\sigma^2 \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-2}\boldsymbol{a} > 0$. It follows that for $\varepsilon > 0$ small enough the term $2\varepsilon\sigma^2 \boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-2}\boldsymbol{a} + o(\varepsilon)$ is positive, and hence $MSE(\hat{\theta}) > MSE(\tilde{\theta}_\varepsilon)$.

5.2 Lasso method

The Least Absolute Shrinkage and Selection Operator (Lasso) method is based on using regularization term of the form $\varepsilon\|\boldsymbol{\beta}\|_1$. That is the Lasso estimator $\tilde{\boldsymbol{\beta}}_\varepsilon$ is obtained as a solution of the following optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \varepsilon\|\boldsymbol{\beta}\|_1. \tag{5.3}$$

Equivalently this can be formulated as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \text{ s.t. } \|\boldsymbol{\beta}\|_1 \leq c, \tag{5.4}$$

for an appropriate choice of the constant $c > 0$. If $c < \|\hat{\boldsymbol{\beta}}\|_1$, then the Lasso estimator performs shrinkage of the LS estimator $\hat{\boldsymbol{\beta}}$.

Note that

$$\frac{\partial\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2}{\partial\boldsymbol{\beta}} = 2(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{X}'\boldsymbol{Y}) = 2(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is the usual least squares estimator. When $c < \|\hat{\boldsymbol{\beta}}\|_1$, an optimal solution of problem (5.4) is on the boundary of the feasible set $C = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq c\}$ and the corresponding optimality conditions are

$$-2(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \in N_C(\boldsymbol{\beta}),$$

where $N_C(\boldsymbol{\beta})$ is the normal cone to C at $\boldsymbol{\beta} \in C$.

Optimality conditions for problem (5.3) are

$$\boldsymbol{0} \in 2(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \varepsilon\partial\|\boldsymbol{\beta}\|_1,$$

where $\partial\|\boldsymbol{\beta}\|_1$ is the subdifferential of the function $f(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$. The subdifferential $\partial\|\boldsymbol{\beta}\|_1$ consists of vectors $\boldsymbol{g}$ (the so-called subgradients) such that $g_i = 1$ if $\beta_i > 0$, $g_i = -1$ if $\beta_i < 0$, and g_i can be any number of the interval $[-1, 1]$ if $\beta_i = 0$. It follows that if ε is bigger than the absolute value of every component $[\boldsymbol{X}'\boldsymbol{Y}]_i$ of vector $\boldsymbol{X}'\boldsymbol{Y}$, then $\tilde{\boldsymbol{\beta}}_\varepsilon = \boldsymbol{0}$. If $\boldsymbol{X}'\boldsymbol{X}$ is diagonal, then $\tilde{\beta}_{\varepsilon,i} = 0$, when $\varepsilon > [\boldsymbol{X}'\boldsymbol{Y}]_i$.

It is possible to look at Lasso estimation form the following point of view. By definition $\|\boldsymbol{\beta}\|_0$ is equal to the number of nonzero components of vector $\boldsymbol{\beta}$. Note that $\|\boldsymbol{\beta}\|_0 = \lim_{q \downarrow 0} \sum |\beta_i|^q$. Consider the problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \text{ s.t. } \|\boldsymbol{\beta}\|_0 \leq c, \tag{5.5}$$

This is a difficult combinatorial problem. Problem (5.4) can be viewed as a convex approximation of problem (5.5). Problem (5.4) can be written as the following quadratic programming problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}, \boldsymbol{\xi}} & \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} & \beta_i \leq \xi_i, \ -\beta_i \leq \xi_i, \ i = 0, ..., k, \\ & \sum_{i=0}^k \xi_i \leq c. \end{array} \tag{5.6}$$

Problem (5.3) can be formulated as the following problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta} \in \mathbb{R}^p, \, z \in \mathbb{R}} & \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + z \\ \text{s.t.} & \varepsilon \|\boldsymbol{\beta}\|_1 \leq z, \end{array} \tag{5.7}$$

which in turn can be written as a quadratic programming problem similar to (5.6).

6 Elements of large samples theory

Let Y_n, $n = 1, ...$, be a sequence of random variables. It is said that Y_n converges in probability to a number a, denoted $Y_n \xrightarrow{p} a$, if for any $\varepsilon > 0$ it follows that

$$\lim_{n \to \infty} \mathbb{P}\{|Y_n - a| \geq \varepsilon\} = 0.$$

Weak Law of Large Numbers (WLLN) can be proved by using Chebishev inequality. That is, let X be a nonnegative valued random variable. Then for any $\varepsilon > 0$ we have

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{E}\left[\mathbf{1}_{[\varepsilon, \infty)}(X)\right] \leq \mathbb{E}[\varepsilon^{-1} X] = \varepsilon^{-1} \mathbb{E}[X].$$

The above inequality sometimes is called Markov inequality.

Now let X be a random variable with finite second order moment. By taking $Y = (X - \mu)^2$, where $\mu = \mathbb{E}[X]$, we obtain from Markov inequality the following Chebishev inequality:

$$\mathbb{P}\{|X - \mu| \geq \varepsilon\} = \mathbb{P}\{(X - \mu)^2 \geq \varepsilon^2\} \leq \varepsilon^{-2} \mathbb{E}[(X - \mu)^2] = \varepsilon^{-2} \mathrm{Var}(X).$$

It follows that if Y_n is a sequence of random variables such that $\mathbb{E}[Y_n] = \mu$, for all n, and $\mathrm{Var}(Y_n)$ tends to zero as $n \to \infty$, then $Y_n \xrightarrow{p} \mu$. In particular if $X_1, ..., X_n$ is iid with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$, then $\mathrm{Var}(\bar{X}) = \sigma^2 / n \to 0$, and hence $\bar{X} \xrightarrow{p} \mu$ as $n \to \infty$.

Now let X_n and Y_n be two sequences of random variables. The notation $Y_n = o_p(X_n)$ means that $Y_n / X_n \xrightarrow{p} 0$ as $n \to \infty$. Usually it is used when X_n is deterministic. In particular $Y_n = o_p(1)$ means that $Y_n \xrightarrow{p} 0$. It is said that Y_n is *bounded in probability* if for any $\varepsilon > 0$ there exists $c > 0$ such that $\mathbb{P}\{|Y_n| > c\} \leq \varepsilon$ for all n. The notation $Y_n = O_p(X_n)$ means that Y_n / X_n is bounded in probability.

Recall that it is said that X_n converges in distribution to a random variable X, denoted $X_n \xrightarrow{\mathcal{D}} X$, if for any number x such that $\mathbb{P}\{X = x\} = 0$ it follows that

$$\lim_{n \to \infty} \mathbb{P}\{X_n \leq x\} = \mathbb{P}\{X \leq x\}.$$

Exercise 4 *(i) Show that if $X_n = o_p(1)$ and $Y_n = O_p(1)$, then $X_nY_n = o_p(1)$. (ii) Show that if $X_n \xrightarrow{\mathcal{D}} X$, then $X_n = O_p(1)$, i.e., X_n is bounded in probability.*

Theorem 6.1 *If $X_n - X \xrightarrow{p} 0$, then $X_n \xrightarrow{\mathcal{D}} X$.*

Proof. For $\varepsilon > 0$ we have

$$\begin{aligned}\mathbb{P}(X_n \le x) &= \mathbb{P}(X_n \le x, |X_n - X| \le \varepsilon) + \mathbb{P}(X_n \le x, |X_n - X| > \varepsilon) \\ &\le \mathbb{P}(X \le x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).\end{aligned}$$

Similarly

$$\mathbb{P}(X \le x - \varepsilon) \le \mathbb{P}(X_n \le x) + \mathbb{P}(|X_n - X| > \varepsilon),$$

and hence

$$\mathbb{P}(X \le x - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon) \le P(X_n \le x).$$

Since $X_n - X \xrightarrow{p} 0$, letting n tend to ∞ and then letting ε tend to 0, we obtain

$$\mathbb{P}(X < x) \le \liminf_{n\to\infty} \mathbb{P}(X_n \le x) \le \limsup_{n\to\infty} \mathbb{P}(X_n \le x) \le \mathbb{P}(X \le x).$$

It follows that $\mathbb{P}(X_n \le x) \to \mathbb{P}(X \le x)$ if $\mathbb{P}(X = x) = 0$. This shows that $X_n \xrightarrow{\mathcal{D}} X$. ∎

Exercise 5 *In a similar way show that if $X_n \xrightarrow{\mathcal{D}} X$ and $Y_n \xrightarrow{p} a$, then $X_n + Y_n \xrightarrow{\mathcal{D}} X + a$. This result is known as* **Slutsky's theorem**.

Theorem 6.2 (Delta theorem) *Let $\boldsymbol{X}_n$ be a sequence of $m \times 1$ random vectors and $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^k$ be a function. Suppose that $\lambda_n(\boldsymbol{X}_n - \boldsymbol{\mu}) \xrightarrow{\mathcal{D}} \boldsymbol{Z}$, where $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\lambda_n \to \infty$, and that $\boldsymbol{g}(\cdot)$ is differentiable at $\boldsymbol{\mu}$ with $\nabla \boldsymbol{g}(\boldsymbol{\mu})$ being the $m \times k$ matrix of partial derivatives. Then*

$$\lambda_n\big(g(\boldsymbol{X}_n) - g(\boldsymbol{\mu})\big) \xrightarrow{\mathcal{D}} [\nabla g(\boldsymbol{\mu})]'\boldsymbol{Z}. \tag{6.1}$$

Proof. Since $\boldsymbol{g}(\cdot)$ is differentiable at $\boldsymbol{\mu}$ we have that

$$\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{\mu}) = [\nabla \boldsymbol{g}(\boldsymbol{\mu})]'(\boldsymbol{x} - \boldsymbol{\mu}) + \boldsymbol{r}(\boldsymbol{x}),$$

where $\boldsymbol{\varepsilon}(\boldsymbol{x}) = \boldsymbol{r}(x)/\|\boldsymbol{x} - \boldsymbol{\mu}\|$ tends to $\mathbf{0}$ as $\boldsymbol{x} \to \boldsymbol{\mu}$. Hence

$$\lambda_n\big(\boldsymbol{g}(\boldsymbol{X}_n) - \boldsymbol{g}(\boldsymbol{\mu})\big) = [\nabla \boldsymbol{g}(\boldsymbol{\mu})]'[\lambda_n(\boldsymbol{X}_n - \boldsymbol{\mu})] + \boldsymbol{\varepsilon}(\boldsymbol{X}_n)[\lambda_n\|\boldsymbol{X}_n - \boldsymbol{\mu}\|]. \tag{6.2}$$

Now since $\lambda_n(\boldsymbol{X}_n - \boldsymbol{\mu})$ converges in distribution, it follows that $\lambda_n(\boldsymbol{X}_n - \boldsymbol{\mu})$ is bounded in probability. Moreover since $\lambda_n \to \infty$ it follows that $\boldsymbol{X}_n \xrightarrow{p} \boldsymbol{\mu}$. Hence $\boldsymbol{\varepsilon}(\boldsymbol{X}_n) \xrightarrow{p} \mathbf{0}$, and thus $\boldsymbol{\varepsilon}(\boldsymbol{X}_n)[\lambda_n\|\boldsymbol{X}_n - \boldsymbol{\mu}\|] \xrightarrow{p} \mathbf{0}$. By Slutsky's theorem the convergence (6.1) follows from (6.2). ∎

In particular it follows that if in addition to the assumptions of Theorem 6.2, $\sqrt{n}(\boldsymbol{X}_n - \boldsymbol{\mu})$ converges in distribution to normal $N(\mathbf{0}, \boldsymbol{\Sigma})$, then $\sqrt{n}\big(\boldsymbol{g}(\boldsymbol{X}_n) - \boldsymbol{g}(\boldsymbol{\mu})\big)$ converges in distribution to normal with zero mean and covariance matrix $[\nabla \boldsymbol{g}(\boldsymbol{\mu})]'\boldsymbol{\Sigma}[\nabla \boldsymbol{g}(\boldsymbol{\mu})]$. For example let X_n and Y_n be two independent sequences of random variables such that $\sqrt{n}\big(X_n - \mu_x\big) \xrightarrow{\mathcal{D}} N(0, \sigma_x^2)$ and $\sqrt{n}\big(Y_n - \mu_y\big) \xrightarrow{\mathcal{D}} N(0, \sigma_y^2)$, $\mu_y \neq 0$. Let us find the asymptotic distribution of (V_n, W_n), where $V_n = X_nY_n$ and $W_n = X_n/Y_n$. Consider $\boldsymbol{g}(x, y) = (xy, x/y)$. Note that $\boldsymbol{g}(X_n, Y_n) = (V_n, W_n)$. By Delta Theorem we have that $\sqrt{n}\begin{bmatrix} V_n - \mu_x\mu_y \\ W_n - \mu_x/\mu_y \end{bmatrix}$ converges in distribution to normal $N(\mathbf{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mu_y & \mu_x \\ 1/\mu_y & -\mu_x/\mu_y^2 \end{bmatrix} \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} \begin{bmatrix} \mu_y & 1/\mu_y \\ \mu_x & -\mu_x/\mu_y^2 \end{bmatrix}.$$

6.1 Maximum likelihood method

Consider a parametric family of distributions defined by probability density functions (pdf) $f(\boldsymbol{x},\boldsymbol{\theta})$, $\boldsymbol{x}\in\mathbb{R}^m$, with parameter vector $\boldsymbol{\theta}\in\Theta\subset\mathbb{R}^k$. Given an iid sample $\boldsymbol{X}_1,...,\boldsymbol{X}_N$, the Maximum Likelihood (ML) estimator of $\boldsymbol{\theta}$ is the maximizer $\hat{\boldsymbol{\theta}}_N$ of the likelihood function $L_N(\boldsymbol{\theta})=\prod_{i=1}^N f(\boldsymbol{X}_i,\boldsymbol{\theta})$ over $\boldsymbol{\theta}\in\Theta$. Note that both $L_N(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_N$ are functions of the sample, this is suppressed in the notation. We can write that as

$$\hat{\boldsymbol{\theta}}_N\in\arg\max_{\boldsymbol{\theta}\in\Theta}\log L_N(\boldsymbol{\theta}). \tag{6.3}$$

Note that such maximizer may not exist or could be not unique. We assume that the random sample is generated according a distribution with pdf $g(\boldsymbol{x})$, written $\boldsymbol{X}_1,...,\boldsymbol{X}_N\sim g(\cdot)$. In particular if $g(\cdot)=f(\cdot,\boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^*\in\Theta$, we say that the model is *correctly specified*. It is said that the model is *identified* at $\boldsymbol{\theta}^*$ if $f(\cdot,\boldsymbol{\theta})=f(\cdot,\boldsymbol{\theta}^*)$, $\boldsymbol{\theta}\in\Theta$, implies that $\boldsymbol{\theta}=\boldsymbol{\theta}^*$. That is, $\boldsymbol{\theta}^*$ is the unique value of the parameter vector which defines the model.

Since $\log L_N(\boldsymbol{\theta})=\sum_{i=1}^N\log f(\boldsymbol{X}_i,\boldsymbol{\theta})$, it follows by the LLN that for a given $\boldsymbol{\theta}$ the average $N^{-1}\log L_N(\boldsymbol{\theta})$ converges w.p.1 as $N\to\infty$ to

$$\mathbb{E}_g[\log f(\boldsymbol{X},\boldsymbol{\theta})]=\int[\log f(\boldsymbol{x},\boldsymbol{\theta})]g(\boldsymbol{x})d\boldsymbol{x},$$

provided this expectation is well defined. The notation $\mathbb{E}_g$ emphasizes that the expectation is taken with respect to the distribution of the sample. It is natural then to expect that the ML estimator $\hat{\boldsymbol{\theta}}_N$ will converge w.p.1 to a maximizer of $\mathbb{E}_g[\log f(\boldsymbol{X},\boldsymbol{\theta})]$ over $\boldsymbol{\theta}\in\Theta$. And indeed it is possible to prove that such converges holds under certain regularity conditions. In order to understand what such maximizer is, we need the following inequality.

Theorem 6.3 (Jensen inequality) *Let $\phi:\mathbb{R}^m\to\mathbb{R}$ be a convex function and $\boldsymbol{X}$ be an $m\times 1$ random vector with mean $\boldsymbol{\mu}=\mathbb{E}[\boldsymbol{X}]$. Then*

$$\mathbb{E}[\phi(\boldsymbol{X})]\geq\phi(\boldsymbol{\mu}). \tag{6.4}$$

Proof. Since $\phi(\cdot)$ is convex we have that there exists $\boldsymbol{\gamma}\in\mathbb{R}^m$ such that

$$\phi(\boldsymbol{x})\geq\phi(\boldsymbol{\mu})+\boldsymbol{\gamma}'(\boldsymbol{x}-\boldsymbol{\mu})$$

for any $\boldsymbol{x}\in\mathbb{R}^m$ (vector $\boldsymbol{\gamma}$ is called subgradient of ϕ at $\boldsymbol{\mu}$). It follows that

$$\mathbb{E}[\phi(\boldsymbol{X})]\geq\phi(\boldsymbol{\mu})+\mathbb{E}[\boldsymbol{\gamma}'(\boldsymbol{X}-\boldsymbol{\mu})].$$

Since $\mathbb{E}[\boldsymbol{\gamma}'(\boldsymbol{x}-\boldsymbol{\mu})]=\boldsymbol{\gamma}'(\mathbb{E}[\boldsymbol{X}]-\boldsymbol{\mu})=0$, the inequality (6.4) follows. ∎

Kullback-Leibler divergence of pdf $f(\cdot)$ from pdf $g(\cdot)$:

$$D(g\|f)=\int\log\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}g(\boldsymbol{x})d\boldsymbol{x}=\mathbb{E}_g\left[\log\frac{g(\boldsymbol{X})}{f(\boldsymbol{X})}\right].$$

Since $-\log$ is a convex function we have by Jensen inequality

$$D(g\|f)=-\mathbb{E}_g\left[\log\frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right]\geq-\log\mathbb{E}_g\left[\frac{f(\boldsymbol{X})}{g(\boldsymbol{X})}\right]=-\log\int\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}g(\boldsymbol{x})d\boldsymbol{x}=-\log\int f(\boldsymbol{x})d\boldsymbol{x}=-\log 1=0.$$

That is, $D(g\|f)\geq 0$ and $D(g\|f)=0$ iff $f=g$.

Since

$$D(g(\cdot)\|f(\cdot,\boldsymbol{\theta})) = \mathbb{E}_g\left[\log g(\boldsymbol{X})\right] - \mathbb{E}_g\left[\log f(\boldsymbol{X},\theta)\right],$$

we have that maximizing $\mathbb{E}_g[\log f(\boldsymbol{X},\boldsymbol{\theta})]$, over $\boldsymbol{\theta}\in\Theta$, is equivalent to minimizing of the KL divergence of $f(\cdot,\boldsymbol{\theta})$ from $g(\cdot)$. In particular, if the model is correctly specified, i.e., $g(\cdot)=f(\cdot,\boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^*\in\Theta$, then $\boldsymbol{\theta}^*$ is a maximizer of $\mathbb{E}[f(\boldsymbol{X},\boldsymbol{\theta})]$, over $\boldsymbol{\theta}\in\Theta$, where the expectation is taken with respect to the true distribution $g(\cdot)=f(\cdot,\boldsymbol{\theta}^*)$. It follows that if the model is identified at $\boldsymbol{\theta}^*$ and some regularity conditions are satisfied, then the ML estimator $\hat{\boldsymbol{\theta}}_N$ converges w.p.1 to $\boldsymbol{\theta}^*$. In that case it is said that $\hat{\boldsymbol{\theta}}_N$ is a *consistent* estimator of $\boldsymbol{\theta}^*$.

6.1.1 Asymptotic distribution of the ML estimators

The following $k\times k$ matrix is called (Fisher) information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathbb{E}_\theta\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{X},\boldsymbol{\theta})\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{X},\boldsymbol{\theta})\right]'\right\}.$$

The notation $\mathbb{E}_\theta$ emphasises that the expectation is taken with respect to the distribution $f(\cdot,\boldsymbol{\theta})$. Note that $\boldsymbol{I}(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$. Let us show that

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbb{E}_\theta\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f(\boldsymbol{X},\boldsymbol{\theta})\right\}.$$

We need to show that

$$\mathbb{E}_\theta\left\{\frac{\partial\log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial\log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_j}\right\} = -\mathbb{E}_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{X},\boldsymbol{\theta})\right\}, \tag{6.5}$$

$i,j=1,...,k$. We have (assuming that the partial derivatives can be taken inside the integral) that

$$\mathbb{E}_\theta\left\{\frac{\partial\log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\right\} = \mathbb{E}_\theta\left\{\frac{\partial f(\boldsymbol{X},\boldsymbol{\theta})/\partial\theta_i}{f(\boldsymbol{X},\boldsymbol{\theta})}\right\} = \frac{\partial}{\partial\theta_i}\int f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x} = 0.$$

Consequently

$$\frac{\partial}{\partial\theta_j}\int\frac{\partial\log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x} = 0. \tag{6.6}$$

By taking the derivative, in the left hand side of (6.6), inside the integral one obtains

$$\begin{aligned}
0 &= \int\frac{\partial}{\partial\theta_j}\left[\frac{\partial\log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}f(\boldsymbol{x},\boldsymbol{\theta})\right]d\boldsymbol{x}\\
&= \int\frac{\partial^2\log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x} + \int\frac{\partial\log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial\log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_j}f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}\\
&= \mathbb{E}_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{X},\boldsymbol{\theta})\right\} + \mathbb{E}_\theta\left\{\frac{\partial\log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial\log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_j}\right\},
\end{aligned}$$

and hence (6.5) follows.

Let us show that $\boldsymbol{I}(\boldsymbol{\theta})$ is positive semidefinite. We have that, for $\boldsymbol{a}\in\mathbb{R}^k$,

$$\boldsymbol{a}'\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{a} = \sum_{i,j=1}^{k} I_{ij}(\boldsymbol{\theta})a_ia_j,$$

where

$$I_{ij}(\theta) = \mathbb{E}_\theta\left\{\frac{\partial\log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial\log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_j}\right\},$$

and hence

$$a_i a_j I_{ij}(\boldsymbol{\theta}) = \mathbb{E}_\theta \left\{ a_i \frac{\partial \log f(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_i} \, a_j \frac{\partial \log f(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_j} \right\}.$$

It follows that

$$\boldsymbol{a}' \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{a} = \mathbb{E}_\theta \left\{ \left[\sum_i^k a_i \frac{\partial \log f(\boldsymbol{X}, \boldsymbol{\theta})}{\partial \theta_i} \right]^2 \right\},$$

and hence $\boldsymbol{a}' \boldsymbol{I}(\theta) \boldsymbol{a} \geq 0$.

Consider now the ML estimation procedure. Suppose that the model is correctly specified and let $\hat{\boldsymbol{\theta}}_N$ be the ML estimator of the true parameter value $\boldsymbol{\theta}^*$. Assume that $\hat{\boldsymbol{\theta}}_N$ is a consistent estimator of $\boldsymbol{\theta}^*$. Since $\hat{\boldsymbol{\theta}}_N$ is a maximizer of $\log L_N(\boldsymbol{\theta})$ the following optimality condition should hold

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{i=1}^N \log f(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_N) \right] = \mathbf{0}. \tag{6.7}$$

By the Mean Value Theorem we can write

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{i=1}^N \log f(\boldsymbol{X}_i, \hat{\boldsymbol{\theta}}_N) \right] = \frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{i=1}^N \log f(\boldsymbol{X}_i, \boldsymbol{\theta}^*) \right] + \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \boldsymbol{\theta}'} \sum_{i=1}^N \log f(\boldsymbol{X}_i, \tilde{\boldsymbol{\theta}}_N) \right] (\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*),$$

for some $\tilde{\boldsymbol{\theta}}_N$ between $\hat{\boldsymbol{\theta}}_N$ and $\boldsymbol{\theta}^*$. It follows that

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*) = - \left[\frac{1}{N} \frac{\partial^2}{\partial \boldsymbol{\theta} \boldsymbol{\theta}'} \sum_{i=1}^N \log f(\boldsymbol{X}_i, \tilde{\boldsymbol{\theta}}_N) \right]^{-1} \left[\frac{1}{\sqrt{N}} \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^N \log f(\boldsymbol{X}_i, \boldsymbol{\theta}^*) \right]. \tag{6.8}$$

Since $\hat{\boldsymbol{\theta}}_N$, and hence $\tilde{\boldsymbol{\theta}}_N$, converge to $\boldsymbol{\theta}^*$ w.p.1, we have by the LLN that

$$\frac{1}{N} \frac{\partial^2}{\partial \, \boldsymbol{\theta} \boldsymbol{\theta}'} \sum_{i=1}^N \log f(\boldsymbol{X}_i, \tilde{\boldsymbol{\theta}}_N) = \frac{1}{N} \sum_{i=1}^N \frac{\partial^2}{\partial \, \boldsymbol{\theta} \boldsymbol{\theta}'} \log f(\boldsymbol{X}_i, \tilde{\boldsymbol{\theta}}_N)$$

converges to $-\boldsymbol{I}(\boldsymbol{\theta}^*)$. Now note that

$$\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log f(\boldsymbol{X}, \boldsymbol{\theta}) \right] = \mathbb{E}_\theta \left[\frac{\frac{\partial}{\partial \theta_i} f(\boldsymbol{X}, \boldsymbol{\theta})}{f(\boldsymbol{X}, \boldsymbol{\theta})} \right] = \int \frac{\partial}{\partial \theta_i} f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = \frac{\partial}{\partial \theta_i} \int f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = 0.$$

Therefore by the CLT we have that $\frac{1}{\sqrt{N}} \frac{\partial}{\partial \boldsymbol{\theta}} \sum_{i=1}^N \log f(\boldsymbol{X}_i, \boldsymbol{\theta}^*)$ converges in distribution to normal with zero mean vector and covariance matrix $\boldsymbol{I}(\boldsymbol{\theta}^*)$. Together with (6.8) this implies that

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}^*)^{-1}\right).$$

- Under what conditions expectation and differentiation can be interchanged, i.e.,

$$\frac{\partial}{\partial \theta} \mathbb{E}[g(X, \theta)] = \mathbb{E}\left[\frac{\partial}{\partial \theta} g(X, \theta) \right].$$

Lebesgue Dominated Convergence Theorem:
if $f_n, g : \Omega \to \mathbb{R}$ are such that $|f_n| \leq g$, $\int_\Omega g dP < \infty$ and $f_n(\omega) \to f(\omega)$ for a.e. $\omega \in \Omega$, then $\int_\Omega f_n dP \to \int_\Omega f dP$.

We have

$$\frac{\partial}{\partial\theta}\mathbb{E}[g(X,\theta)] = \lim_{h\to 0}\mathbb{E}\left[\frac{g(X,\theta+h)-g(X,\theta)}{h}\right] = \mathbb{E}\left[\lim_{h\to 0}\frac{g(X,\theta+h)-g(X,\theta)}{h}\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}g(X,\theta)\right],$$

provided

$$\big|g(X,\theta+h)-g(X,\theta)\big| \le K(X)|h|,$$

where $\mathbb{E}[K(X)] < \infty$.

6.1.2 Cramer - Rao lower bound

Let $X_1,...,X_n$ be an iid sample from $f(x,\theta)$, $\theta\in\mathbb{R}$, and $T(\boldsymbol{X})$ be a statistic, i.e., $T(\boldsymbol{X})$ is a function of $\boldsymbol{X}=(X_1,...,X_n)$. Then under some regularity conditions

$$\mathrm{Var}_\theta[T(\boldsymbol{X})] \ge i_X(\theta)^{-1}[\partial g(\theta)/\partial\theta]^2, \tag{6.9}$$

where $g(\theta)=\mathbb{E}_\theta[T(\boldsymbol{X})]$ and $i_X(\theta)=\mathbb{E}_\theta\left[(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta))^2\right]$ is Fisher's information of

$$f(\boldsymbol{x},\theta)=\prod_{i=1}^n f(x_i,\theta).$$

In particular, if $\mathbb{E}_\theta[T(\boldsymbol{X})]=\theta$, i.e. $T(\boldsymbol{X})$ is an unbiased estimator of θ, then

$$\mathrm{Var}_\theta[T(\boldsymbol{X})] \ge i_X(\theta)^{-1}.$$

Indeed, we have that

$$\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right] = \int\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{x},\theta)}{f(\boldsymbol{x},\theta)}f(\boldsymbol{x},\theta)d\boldsymbol{x} = \int\frac{\partial}{\partial\theta}f(\boldsymbol{x},\theta)d\boldsymbol{x} = \frac{\partial}{\partial\theta}\int f(\boldsymbol{x},\theta)d\boldsymbol{x} = 0,$$

provided the derivative can be taken inside the integral. Then

$$\begin{array}{rcl}\mathrm{Cov}_\theta\left(T(\boldsymbol{X}),\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right) &=& \mathbb{E}_\theta\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right] = \mathbb{E}_\theta\left[T(\boldsymbol{X})\frac{\partial}{\partial\theta}f(\boldsymbol{X},\theta)/f(\boldsymbol{X},\theta)\right]\\ &=& \int T(\boldsymbol{x})\partial f(\boldsymbol{x},\theta)/\partial\theta d\boldsymbol{x} = \frac{\partial}{\partial\theta}\int T(\boldsymbol{x})f(\boldsymbol{x},\theta)d\boldsymbol{x}.\end{array}$$

That is,

$$\mathrm{Cov}_\theta\left(T(\boldsymbol{X}),\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right) = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[T(\boldsymbol{X})] = \partial g(\theta)/\partial\theta.$$

Now by Cauchy inequality we have

$$\left[\mathrm{Cov}_\theta\left(T(\boldsymbol{X}),\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right)\right]^2 \le \mathrm{Var}_\theta[T(\boldsymbol{X})]\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right].$$

Moreover

$$\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right] = \mathbb{E}_\theta\left[(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta))^2\right] = i(\theta),$$

and hence the inequality (6.9) follows. Note that, by the independence of $X_1,...,X_n$,

$$i_X(\theta) = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X},\theta)\right] = \sum_{i=1}^n\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_i,\theta)\right] = ni(\theta),$$

where $i(\theta)=\mathbb{E}_\theta\left[(\frac{\partial}{\partial\theta}\log f(X_i,\theta))^2\right]$.

This can be extended to a multivariate setting. Suppose that $\boldsymbol{\theta}=(\theta_1,...,\theta_m)'$ and let $\boldsymbol{T}(\boldsymbol{X})=(T_1(\boldsymbol{X}),...,T_m(\boldsymbol{X}))'$ be an unbiased estimator of $\boldsymbol{\theta}$, i.e., $\mathbb{E}_{\boldsymbol{\theta}}[T_i]=\theta_i$, $i=1,...,m$. Then for any vector $\boldsymbol{a}\in\mathbb{R}^m$ it follows that

$$\boldsymbol{a}'\boldsymbol{\Sigma}_T\boldsymbol{a} \ge \boldsymbol{a}'\boldsymbol{I}(\boldsymbol{\theta})^{-1}\boldsymbol{a}, \tag{6.10}$$

where $\boldsymbol{\Sigma}_T$ is the covariance matrix of $\boldsymbol{T}(\boldsymbol{X})$ and $\boldsymbol{I}(\boldsymbol{\theta})$ is Fisher information matrix.

7 Hypotheses testing

Consider testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta_1$, where $\Theta_0 \subset \Theta_1 \subset \mathbb{R}^m$. Given data (sample) $\boldsymbol{X} = (X_1, ..., X_N)$ reject H_0 if $\boldsymbol{X} \in R$, where $R \subset \mathbb{R}^N$ is the rejection region and $R^c = \mathbb{R}^N \setminus R$ is the acceptance region. Two types of error, type I error - reject H_0 when H_0 is true, type II error - accept H_0 when H_0 is false. The corresponding probabilities $\alpha = P(\text{type I error})$ and $\beta = P(\text{type II error})$.

Theorem 7.1 (Neyman - Pearson Lemma) *Consider simple alternatives $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ with pdfs $f(x, \boldsymbol{\theta}_i)$, $i = 0, 1$. Then the minimal error rejection region is $R = \{\boldsymbol{x} \in \mathbb{R}^N : f(\boldsymbol{x}, \boldsymbol{\theta}_1) \geq \kappa f(\boldsymbol{x}, \boldsymbol{\theta}_0)\}$, where $\kappa > 0$ is such that $\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \alpha$ with $f(\boldsymbol{x}, \boldsymbol{\theta}) = \prod_{i=1}^N f(x_i, \boldsymbol{\theta})$.*

Proof Consider minimization of

$$\int_{R^c} f(\boldsymbol{x}, \boldsymbol{\theta}_1) d\boldsymbol{x} + \kappa \int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x}.$$

Note that $\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) = \alpha$ and $\int_{R^c} f(\boldsymbol{x}, \boldsymbol{\theta}_1) d\boldsymbol{x} = \beta$. Therefore we need to choose R^c to minimize the above. Since $R = \mathbb{R}^N \setminus R^c$ we have that

$$\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \int_{\mathbb{R}^N} f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} - \int_{R^c} f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x}$$

and $\int_{\mathbb{R}^N} f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = 1$, and hence

$$\int_{R^c} f(\boldsymbol{x}, \boldsymbol{\theta}_1) d\boldsymbol{x} + \kappa \int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \kappa + \int_{R^c} [f(\boldsymbol{x}, \boldsymbol{\theta}_1) - \kappa f(\boldsymbol{x}, \boldsymbol{\theta}_0)] d\boldsymbol{x}.$$

It follows that the minimum is attained for

$$R^c = \{\boldsymbol{x} : f(\boldsymbol{x}, \boldsymbol{\theta}_1) - \kappa f(\boldsymbol{x}, \boldsymbol{\theta}_0) < 0\},$$

and

$$R = \{\boldsymbol{x} : f(\boldsymbol{x}, \boldsymbol{\theta}_1) - \kappa f(\boldsymbol{x}, \boldsymbol{\theta}_0) \geq 0\}.$$

7.1 Likelihood Ratio Test

Consider

$$\lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})},$$

where $\Theta = \Theta_0 \cup \Theta_1$ and $L(\boldsymbol{\theta}) = \prod_{i=1}^N f(x_i, \boldsymbol{\theta})$ is the corresponding likelihood function. Note that $0 \leq \lambda(\boldsymbol{x}) \leq 1$. The rejection region of the Likelihood Ratio Test (LRT) is

$$R = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\},$$

for some $c \in (0, 1)$.

For simple alternatives when $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ we have that

$$\lambda(\boldsymbol{x}) = \frac{L(\boldsymbol{\theta}_0)}{\max\{L(\boldsymbol{\theta}_0), L(\boldsymbol{\theta}_1)\}},$$

and hence this is equivalent to the rejection region of the Neyman - Pearson Lemma.

Let us discuss asymptotics of the LRT. We will discuss this for the simple $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the unrestricted alternative $H_1 : \boldsymbol{\theta} \in \mathbb{R}^m$. We have that

$$-2\log\lambda(\boldsymbol{X}) = -2\log L(\boldsymbol{\theta}_0) - 2\inf_{\boldsymbol{\theta}\in\mathbb{R}^m}[-\log L(\boldsymbol{\theta})].$$

Note that

$$\inf_{\boldsymbol{\theta}\in\mathbb{R}^m}[-\log L(\boldsymbol{\theta})] = -\log L(\hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimator under the unrestricted alternative H_1. Consider

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial\theta}\log L(\boldsymbol{\theta}) = \sum_{i=1}^{N}\frac{\partial}{\partial\theta}\log f(X_i,\boldsymbol{\theta}) \tag{7.1}$$

is called the score function. Note that $S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ (necessary optimality condition) and $\mathbb{E}_{\boldsymbol{\theta}}[S(\boldsymbol{\theta})] = \mathbf{0}$. Now using second order Taylor approximation,

$$\log L(\hat{\boldsymbol{\theta}}) \approx \log L(\boldsymbol{\theta}_0) + \left[\tfrac{\partial}{\partial\theta}\log L(\boldsymbol{\theta}_0)\right]'(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\left[\tfrac{\partial^2}{\partial\theta\partial\theta'}\log L(\boldsymbol{\theta}_0)\right](\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).$$

Note that $\frac{\partial}{\partial\theta}\log L(\boldsymbol{\theta}_0) = S(\boldsymbol{\theta}_0)$ and $\frac{\partial^2}{\partial\theta\partial\theta'}\log L(\boldsymbol{\theta}_0) = \frac{\partial}{\partial\theta'}S(\boldsymbol{\theta}_0)$. Hence and since $S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$,

$$\log L(\hat{\boldsymbol{\theta}}) \approx \log L(\boldsymbol{\theta}_0) - \left[S(\hat{\boldsymbol{\theta}}) - S(\boldsymbol{\theta}_0)\right]'(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) + \tfrac{1}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)'\left[\tfrac{\partial}{\partial\theta'}S(\boldsymbol{\theta}_0)\right](\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).$$

Also

$$S(\hat{\boldsymbol{\theta}}) - S(\boldsymbol{\theta}_0) \approx \left[\frac{\partial}{\partial\theta'}S(\boldsymbol{\theta}_0)\right](\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).$$

Therefore

$$-2\log\lambda(\boldsymbol{X}) = -2\log L(\boldsymbol{\theta}_0) + 2\log L(\hat{\boldsymbol{\theta}}) \approx [\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]'\left[-\frac{1}{N}\frac{\partial^2}{\partial\theta\partial\theta'}\log L(\boldsymbol{\theta}_0)\right][\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)].$$

Assuming H_0, we have that $\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1})$ and by the LLN, $-\frac{1}{N}\frac{\partial^2}{\partial\theta\partial\theta'}\log L(\boldsymbol{\theta}_0)$ converges in probability to $\boldsymbol{I}(\boldsymbol{\theta}_0)$. It follows that under H_0, the statistic $-2\log\lambda(\boldsymbol{X})$ converges in distribution to χ^2_m.

In general $-2\log\lambda(\boldsymbol{X}) \xrightarrow{\mathcal{D}} \chi^2_{q-k}$ under H_0, where $q = \dim\Theta$ and $k = \dim\Theta_0$.

Power of the LR test under local alternatives
Suppose the following so-called parameter drift (local alternatives) for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$: $\boldsymbol{\theta}_{0,N} = \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{b}$, where $\boldsymbol{b} \in \mathbb{R}^m$ is a fixed vector. Then

$$\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = \sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{0,N}) + \sqrt{N}(\boldsymbol{\theta}_{0,N}-\boldsymbol{\theta}_0) = \sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{0,N}) + \boldsymbol{b} \xrightarrow{\mathcal{D}} N(\boldsymbol{b}, \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}).$$

Hence

$$-2\log\lambda \approx [\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]'\boldsymbol{I}(\boldsymbol{\theta}_0)[\sqrt{N}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)]$$

converges in distribution to noncentral chi-square $\chi^2_m(\delta)$ with the noncentrality paprameter $\delta = \boldsymbol{b}'\boldsymbol{I}(\boldsymbol{\theta}_0)\boldsymbol{b}$.

7.2 Testing equality constraints

Consider testing $H_0 : \boldsymbol{a}(\boldsymbol{\theta}) = (a_1(\boldsymbol{\theta}), ..., a_q(\boldsymbol{\theta}))' = \mathbf{0}$ against $H_1 : \boldsymbol{a}(\boldsymbol{\theta}) \neq \mathbf{0}$. Let

$$\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \mathbb{R}^m} L(\boldsymbol{\theta}) \text{ and } \tilde{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{a}(\boldsymbol{\theta})=\mathbf{0}} L(\boldsymbol{\theta})$$

be the respective unrestricted and restricted ML estimators. We have here that the 2log Likelihood Ratio Test (LRT) statistic is $2[\log L(\hat{\boldsymbol{\theta}}) - L(\tilde{\boldsymbol{\theta}})]$. Under H_0 (and corresponding regularity conditions) this test statistic converges in distribution to χ^2_q.

Wald test statistic for testing (linear[1]) equality constraints $H_0 : \boldsymbol{A\theta} = \boldsymbol{c}$ against $H_1 : \boldsymbol{A\theta} \neq \boldsymbol{c}$ is

$$N(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{c})'(\boldsymbol{AI}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{A}')^{-1}(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{c}).$$

Here $\boldsymbol{A}$ is $q \times m$ matrix of rank q. We have that under H_0,

$$\sqrt{N}\boldsymbol{A}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \sqrt{N}(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{c}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{AI}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{A}').$$

It follows that under H_0 the Wald test statistic converges in distribution to χ^2_q.

Note that the LRT

$$2[\log L(\hat{\boldsymbol{\theta}}) - L(\tilde{\boldsymbol{\theta}})] \approx \inf_{\boldsymbol{A\theta}=\boldsymbol{c}} [\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]' \left[-\frac{1}{N} \frac{\partial^2}{\partial\theta\partial\theta'} \log L(\boldsymbol{\theta}_0) \right] [\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]$$

and under H_0, $-\frac{1}{N}\frac{\partial^2}{\partial\theta\partial\theta'} \log L(\boldsymbol{\theta}_0) \approx \boldsymbol{I}(\boldsymbol{\theta}_0)$. Therefore under H_0, the LR and Wald test statistics are asymptotically equivalent.

Score function test statistic for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$:

$$N^{-1}S(\boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}S(\boldsymbol{\theta}_0).$$

Recall that $\mathbb{E}_{\boldsymbol{\theta}}[S(\boldsymbol{\theta})] = \mathbf{0}$ and $N^{-1/2}S(\theta) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}))$. It follows that under H_0

$$N^{-1}S(\boldsymbol{\theta}_0)'\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}S(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \chi^2_m.$$

In general, under H_0,

$$N^{-1}S(\tilde{\boldsymbol{\theta}})'\boldsymbol{I}(\tilde{\boldsymbol{\theta}})^{-1}S(\tilde{\boldsymbol{\theta}}) \xrightarrow{\mathcal{D}} \chi^2_q$$

when testing q equality constraints.

8 Multinomial distribution

Consider $\boldsymbol{Y} = (Y_1, ..., Y_m)'$ with $Y_1 + ... + Y_m = N$ and

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{N!}{y_1! \times \cdots \times y_m!} \prod_{i=1}^{m} p_i^{y_i},$$

where $p_i > 0$, $i = 1, ..., m$, and $p_1 + ... + p_m = 1$. We denote this as $\boldsymbol{Y} \sim Mult(N, \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, ..., p_m)'$. The ML estimation

$$\max_{p \geq 0} \sum_{i=1}^{m} y_i \log p_i \text{ s.t. } p_1 + ... + p_m = 1.$$

[1]For nonlinear constraints we can use $\boldsymbol{A} = \partial \boldsymbol{a}(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}'$.

It follows that the ML estimators are $\hat{p}_i = Y_i/N$, $i = 1, ..., m$.

If $\boldsymbol{Y} \sim Mult(N, \boldsymbol{p})$, then the covariance matrix $\mathrm{Cov}(\boldsymbol{Y}) = N\boldsymbol{C}$, where $c_{ii} = p_i(1-p_i)$, $i = 1, ..., m$ and $c_{ij} = -p_i p_j$, $i \neq j$.

Indeed Y_i has binomial distribution with probability of success p_i and hence $\mathrm{Var}(Y_i) = Np_i(1-p_i)$. Now $Y_i + Y_j$, $i \neq j$, has binomial distribution with probability of success $p_i + p_j$ and hence

$$\mathrm{Var}(Y_i + Y_j) = N(p_i + p_j)(1 - p_i - p_j) = N(p_i - p_i^2 + p_j - p_j^2 - 2p_i p_j).$$

On the other hand

$$\mathrm{Var}(Y_i + Y_j) = \mathrm{Var}(Y_i) + \mathrm{Var}(Y_j) + 2\mathrm{Cov}(Y_i, Y_j).$$

It follows that $\mathrm{Cov}(Y_i, Y_j) = -Np_i p_j$.

This can be written as $\boldsymbol{C} = \boldsymbol{P} - \boldsymbol{p}\boldsymbol{p}'$, where $\boldsymbol{P} = \mathrm{diag}(p_1, ..., p_m)$ and $\boldsymbol{p} = (p_1, ..., p_m)'$. Note that $\boldsymbol{C}\boldsymbol{1}_m = \boldsymbol{0}$ and $\mathrm{rank}(\boldsymbol{C}) = m - 1$.

Consider testing $H_0 : \boldsymbol{p} = \boldsymbol{p}^*$ against $H_1 : \boldsymbol{p} \neq \boldsymbol{p}^*$. The corresponding log LRT statistic is

$$\log \lambda = \sum_{i=1}^m Y_i \log p_i^* - \sum_{i=1}^m Y_i \log \hat{p}_i.$$

Since $\hat{p}_i = Y_i/N$ we can write

$$\log \lambda = \sum_{i=1}^m Y_i \log p_i^* - \sum_{i=1}^m Y_i \log Y_i/N = \sum_{i=1}^m Y_i \log \frac{Np_i^*}{Y_i}.$$

Note that (second order Taylor approximation of $\log x$ at $x = 1$)

$$\log x = x - 1 - \tfrac{1}{2}(x-1)^2 + o(x-1)^2.$$

Under H_0 values $\hat{p}_i$ are close to p_i^* and hence

$$\sum_{i=1}^m Y_i \log \frac{p_i^*}{\hat{p}_i} \approx \sum_{i=1}^m Y_i \left(\frac{p_i^*}{\hat{p}_i} - 1\right) - \tfrac{1}{2}\sum_{i=1}^m Y_i \left(\frac{p_i^*}{\hat{p}_i} - 1\right)^2.$$

Moreover

$$\sum_{i=1}^m Y_i \left(\frac{p_i^*}{\hat{p}_i} - 1\right) = \sum_{i=1}^m (Np_i^* - Y_i) = 0,$$

since $\sum_{i=1}^m p_i^* = 1$ and $\sum_{i=1}^m Y_i = N$. Hence under H_0,

$$-2\log \lambda \approx \sum_{i=1}^m \frac{(Y_i - Np_i^*)^2}{Y_i} \approx \sum_{i=1}^m \frac{(Y_i - Np_i^*)^2}{Np_i^*}.$$

Values Y_i are called observed frequencies, Np_i^* are called expected frequencies, and $\sum_{i=1}^m \frac{(Y_i - Np_i^*)^2}{Np_i^*}$ is the famous Pearson's chi-square test statistic.

We can write this statistic as

$$\sum_{i=1}^m \frac{(Y_i - Np_i^*)^2}{Np_i^*} = N(\hat{\boldsymbol{p}} - \boldsymbol{p}^*)'\boldsymbol{Q}(\hat{\boldsymbol{p}} - \boldsymbol{p}^*),$$

where $\hat{\boldsymbol{p}} = (Y_1/N, ..., Y_m/N)'$ and $\boldsymbol{Q} = \mathrm{diag}(1/p_1^*, ..., 1/p_m^*)$. Recall that $\sqrt{N}(\hat{\boldsymbol{p}} - \boldsymbol{p}^*) \xrightarrow{\mathcal{D}} N(\boldsymbol{0}, \boldsymbol{C})$. Consider $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{C})$, and let us show that $\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}$ has χ^2_{m-1} distribution. For $\boldsymbol{W} = \boldsymbol{Q}^{1/2}\boldsymbol{Z}$ we have that $\boldsymbol{W} \sim N\big(\boldsymbol{0}, \boldsymbol{I}_m - (\boldsymbol{p}^{*1/2})(\boldsymbol{p}^{*1/2})'\big)$ and $\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z} = \boldsymbol{W}'\boldsymbol{W}$. Note that $\boldsymbol{I}_m - (\boldsymbol{p}^{*1/2})(\boldsymbol{p}^{*1/2})'$ is a projection matrix of rank

$$\mathrm{rank}\big(\boldsymbol{I}_m - (\boldsymbol{p}^{*1/2})(\boldsymbol{p}^{*1/2})'\big) = \mathrm{tr}\big(\boldsymbol{I}_m - (\boldsymbol{p}^{*1/2})(\boldsymbol{p}^{*1/2})'\big) = m - (\boldsymbol{p}^{*1/2})'(\boldsymbol{p}^{*1/2}) = m - \sum_{i=1}^{m} p_i^* = m - 1.$$

It follows that $N(\hat{\boldsymbol{p}} - \boldsymbol{p}^*)'\boldsymbol{Q}(\hat{\boldsymbol{p}} - \boldsymbol{p}^*) \xrightarrow{\mathcal{D}} \chi^2_{m-1}$.

General model: $\boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^q$, with $\sum_{i=1}^m p_i(\boldsymbol{\theta}) = 1$. The ML estimator of $\boldsymbol{\theta}^*$ is solution of the optimization problem

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} Y_i \log p_i(\boldsymbol{\theta}).$$

Asymptotically, assuming that the model is correct, this is equivalent to

$$\min_{\boldsymbol{\theta}} (\hat{\boldsymbol{p}} - \boldsymbol{p}(\boldsymbol{\theta}))'\tilde{\boldsymbol{Q}}(\hat{\boldsymbol{p}} - \boldsymbol{p}(\boldsymbol{\theta})),$$

where $\hat{p}_i = Y_i/N$, $i = 1, ..., N$ and $\tilde{\boldsymbol{Q}} = \mathrm{diag}(1/\tilde{p}_1, ..., 1/\tilde{p}_m)$ with $\tilde{\boldsymbol{p}}$ being a consistent estimator of the true value of $\boldsymbol{p}$ (for example take $\tilde{p}_i = \hat{p}_i$). Then $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\mathcal{D}} N(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\theta})^{-1})$ with $\boldsymbol{I}(\boldsymbol{\theta}) = \boldsymbol{P}(\boldsymbol{\theta})'\boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})$, $\boldsymbol{P}(\boldsymbol{\theta}) = \partial \log \boldsymbol{p}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$ is $m \times q$ matrix and $\boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{P}(\boldsymbol{\theta}) - \boldsymbol{p}(\boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{\theta})'$.

The LR test for $H_0 : \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$ against unrestricted alternative is

$$-\log \lambda = \sum_{i=1}^{m} Y_i \log \frac{Y_i}{N p_i(\tilde{\boldsymbol{\theta}})},$$

where $\tilde{\boldsymbol{\theta}}$ is the MLE under H_0. Asymptotics of the LR test is $-2\log\lambda \xrightarrow{\mathcal{D}} \chi^2_{m-1-q}$.

9 Logistic regression

Let $Y_1, ..., Y_n$ be independent random variables such that Y_i has the binomial distribution $B(m_i, \pi_i)$. Consider the logit model:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}, \quad i = 1, ..., n. \tag{9.1}$$

That is

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki},\ i = 1, ..., n,$$

where $\frac{\pi_i}{1-\pi_i}$ is called the odds ratio.

We have that

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i}(1 - \pi_i)^{m_i - y_i}.$$

It follows that the likelihood function here is

$$L(\boldsymbol{\pi}; \boldsymbol{y}) = c \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{m_i - y_i},$$

where the constant $c = \prod_{i=1}^{n} \binom{m_i}{y_i}$ is independent of $\boldsymbol{\pi}$. Hence up to the constant $\log c$ independent of $\boldsymbol{\pi}$, the log likelihood function $\log L(\boldsymbol{\pi}; \boldsymbol{y})$ can be written as

$$l(\boldsymbol{\pi}; \boldsymbol{y}) = \sum_{i=1}^{n} \big[y_i \log \pi_i + (m_i - y_i) \log(1 - \pi_i) \big]$$

(note that, by definition, $0 \log 0 = 0$).

Fisher's information matrix, for $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$, can be written in the form $\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$, where $\boldsymbol{W}$ is a diagonal matrix given by

$$\boldsymbol{W} = \mathrm{diag}\{m_i \pi_i (1 - \pi_i)\}.$$

Indeed, we have that

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - m_i \pi_i}{\pi_i (1 - \pi_i)},$$

and hence

$$\frac{\partial l}{\partial \beta_s} = \sum_{i=1}^{n} \frac{y_i - m_i \pi_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_s},$$

where

$$\frac{\partial \pi_i}{\partial \beta_s} = \pi_i (1 - \pi_i) X_{si}.$$

Consequently the st-element of Fisher's information matrix is

$$\mathbb{E}\Big[\frac{\partial l}{\partial \beta_s} \frac{\partial l}{\partial \beta_t} \Big] = \mathbb{E} \left[\sum_{i,j} \Big(\frac{Y_i - m_i \pi_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_s} \Big) \Big(\frac{Y_j - m_j \pi_j}{\pi_j (1 - \pi_j)} \frac{\partial \pi_j}{\partial \beta_t} \Big) \right], \quad s, t = 0, ..., k.$$

Moreover, $\mathbb{E}[Y_i] = m_i \pi_i$, and hence (by independence)

$$\mathbb{E}\big[(Y_i - m_i \pi_i)(Y_j - m_j \pi_j) \big] = 0, \ \text{ if } i \neq j,$$

and

$$\mathbb{E}\big[(Y_i - m_i \pi_i)^2 \big] = \mathrm{Var}[Y_i] = m_i \pi_i (1 - \pi_i).$$

It follows that

$$\mathbb{E}\Big[\frac{\partial l}{\partial \beta_s} \frac{\partial l}{\partial \beta_t} \Big] = \sum_{i=1}^{n} \frac{m_i}{\pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \beta_s} \frac{\partial \pi_i}{\partial \beta_t} = \sum_{i=1}^{n} m_i \pi_i (1 - \pi_i) X_{si} X_{ti}.$$

The maximum likelihood (ML) equations are

$$\sum_{i=1}^{n} (y_i - m_i \pi_i) X_{si} = 0, \ \ s = 0, ..., k.$$

Since the function $g(x) = \log(1 + e^x)$ is strictly convex, it follows that $l(\cdot, \boldsymbol{y})$ is strictly concave, and hence the ML equations for estimating $\boldsymbol{\beta}$ have *unique* solution $\hat{\boldsymbol{\beta}}$.

Consider

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki})}, \quad i = 1, ..., n,$$

and the following so-called deviance function,

$$\Lambda = -2 l(\tilde{\boldsymbol{\pi}}; \boldsymbol{y}) + 2 l(\hat{\boldsymbol{\pi}}; \boldsymbol{y}),$$

where $\tilde{\boldsymbol{\pi}}$ is the ML estimate of $\boldsymbol{\pi}$ under a specified H_0. That is, Λ is the log-likelihood ratio test statistic $-2\log\lambda$ for testing H_0. In particular, for $H_0 : \beta_1 = ... = \beta_k = 0$ we have that $\tilde{\pi}_i = \tilde{\pi}$, $i = 1, ..., n$, where $\tilde{\pi} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n m_i}$.

If $m_i = 1$, $i = 1, ..., n$, then $Y_1, ..., Y_n$ become Bernoulli random variables with $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$. In that case

$$l(\boldsymbol{\pi}; \boldsymbol{y}) = \sum_{i=1}^{n} \big[y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i) \big].$$

For $H_0 : \beta_1 = ... = \beta_k = 0$ we have that $\tilde{\pi}_i = \tilde{\pi}$, $i = 1, ..., n$, where $\tilde{\pi} = \frac{\sum_{i=1}^n y_i}{n}$, and hence $l(\boldsymbol{\pi}; \boldsymbol{y}) = (\sum_{i=1}^n y_i) \log \tilde{\pi}$.

10 Exponential family of distributions

It is said that $\boldsymbol{X}$ is distributed according to the *exponential family* (in the canonical form) if its probability density function is of the form

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \exp \left[\sum_{i=1}^{q} \theta_i T_i(\boldsymbol{x}) - A(\boldsymbol{\theta}) \right] h(\boldsymbol{x}),$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_q)' \in \Theta$ is vector of parameters with

$$\Theta = \left\{ \boldsymbol{\theta} : \int \exp \left[\sum_{i=1}^{q} \theta_i T_i(\boldsymbol{x}) \right] h(\boldsymbol{x}) d\boldsymbol{x} < \infty \right\}.$$

Let us show that

$$\mathbb{E}_\theta(T_j) = \frac{\partial}{\partial \theta_j} A(\boldsymbol{\theta}), \tag{10.1}$$

$$\mathrm{Cov}(T_j, T_k) = \frac{\partial^2}{\partial \theta_j \partial \theta_k} A(\boldsymbol{\theta}). \tag{10.2}$$

Indeed, we have that $\int f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x} = 1$ for all $\boldsymbol{\theta} \in \Theta$. Let $\boldsymbol{\theta}$ be an interior point of Θ, and hence the expectation and differentiation can be interchanged. We have that $\frac{\partial}{\partial \theta_j} \int f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x} = 0$ and

$$\frac{\partial}{\partial \theta_j} f(\boldsymbol{x}; \theta) = \left[T_j(\boldsymbol{x}) - \frac{\partial}{\partial \theta_j} A(\boldsymbol{\theta}) \right] f(\boldsymbol{x}; \boldsymbol{\theta}),$$

and hence

$$\begin{aligned} 0 = \tfrac{\partial}{\partial \theta_j} \textstyle\int f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x} &= \textstyle\int \tfrac{\partial}{\partial \theta_j} f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x} \\ &= \mathbb{E}_\theta \left[T_j(\boldsymbol{X}) - \tfrac{\partial}{\partial \theta_j} A(\boldsymbol{\theta}) \right] \\ &= \mathbb{E}_\theta(T_j) - \tfrac{\partial}{\partial \theta_j} A(\boldsymbol{\theta}). \end{aligned}$$

It follows that $\mathbb{E}_\theta(T_j) = \frac{\partial}{\partial \theta_j} A(\boldsymbol{\theta})$. The other equation follows in a similar way from $\frac{\partial^2}{\partial \theta_j \partial \theta_k} \int f(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{x} = 0$.

11 Generalized linear models

Let $\boldsymbol{Y} = (Y_1, ..., Y_n)'$ be a vector of responses whose components are independently distributed with means $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)'$, i.e., $\mu_i = \mathbb{E}[Y_i]$, $i = 1, ..., n$. The linear model assumes that $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$, $i = 1, ..., n$, where $\boldsymbol{\beta}$ is $p \times 1$ vector of parameters and $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})'$ are observed values of the predictors. That is, the conditional expectation $\mathbb{E}[Y_i|\boldsymbol{X}_i = \boldsymbol{x}_i] = \boldsymbol{x}_i'\boldsymbol{\beta}$, $i = 1, ..., n$.

This can be generalized in the following way. Let us introduce a linear predictor

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta},\ i = 1, ..., n. \tag{11.1}$$

The new symbol η is related to μ by the equation $\eta = g(\mu)$, where $g(\cdot)$ is a specified function called the *link function*. That is

$$\eta_i = g(\mu_i),\ i = 1, ..., n, \tag{11.2}$$

and

$$\mu_i = g^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}),\ i = 1, ..., n. \tag{11.3}$$

For example, in the linear case $\boldsymbol{\eta} = \boldsymbol{\mu}$, i.e., $g(\mu) = \mu$. In the logistic regression $g(\pi) = \log \frac{\pi}{1-\pi}$ is the logit link function.

Suppose now that each component Y_i of the response vector has a distribution in the exponential family with pdf of the form

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{11.4}$$

for some specified functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. The parameter θ is called the natural parameter, and the parameter ϕ the dispersion parameter. For example for the normal distribution $N(\mu, \sigma^2)$ we can write the corresponding density

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

in the form (11.4) with $\theta = \mu$, $\phi = \sigma$ and

$$a(\phi) = \phi^2,\ b(\theta) = \theta^2/2,\ c(y, \phi) = -\tfrac{1}{2}\{y^2/\phi^2 + \log(2\pi\phi^2)\}.$$

If ϕ is known, then $a(\phi)$ is viewed as a constant, $c(y, \phi) = c(y)$, and (11.4) becomes an exponential family in the canonical form with canonical parameter θ.

Consider

$$l(y; \theta, \phi) = \log f_Y(y; \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi). \tag{11.5}$$

By the standard theory of the ML we have that

$$\mathbb{E}[\partial l/\partial\theta] = 0, \tag{11.6}$$

$$\mathbb{E}[\partial^2 l/\partial\theta^2] = -\mathbb{E}[(\partial l/\partial\theta)^2]. \tag{11.7}$$

Also by (11.5)

$$\partial l/\partial\theta = \frac{y - b'(\theta)}{a(\phi)}$$

and because of (11.6), $\mathbb{E}[Y - b'(\theta)] = 0$. Thus $\mathbb{E}[Y] = b'(\theta)$, that is (compare with (10.1))

$$\mu = b'(\theta).$$

Moreover

$$\partial^2 l/\partial\theta^2 == -b''(\theta)/a(\phi)$$

and hence $b''(\theta)/a(\phi) = \mathrm{Var}(Y)/a^2(\phi)$ and thus (compare with (10.2))

$$\mathrm{Var}(Y) = b''(\theta)a(\phi).$$

For binomial distribution $B(m,\pi)/m$ the corresponding distribution function is

$$P(Y=y) = \binom{m}{my} \pi^{my}(1-\pi)^{m(1-y)},\ y = 0, 1/m, ..., 1.$$

Let us set $\theta = \log\frac{\pi}{1-\pi}$ as the natural parameter, and hence $\pi = \frac{e^\theta}{1+e^\theta}$. Here $\mu = \pi$ and thus $\mu = \frac{e^\theta}{1+e^\theta}$. Assume that m is known and set $\phi = 1/m$, $a(\phi) = \phi$, $b(\theta) = \log(1+e^\theta)$, $c(y,\phi) = \log\binom{m}{my}$. Note that $0\log 0 = 0$, and hence for $m = 1$ we have that $\phi = 1$ and $c(y,\phi) = 0$. The link function here is logit $g(\pi) = \log\frac{\pi}{1-\pi}$.

For Poisson distribution

$$P(Y=y) = \frac{1}{y!}e^{-\mu}\mu^y,\ y = 0, 1, 2, ...,$$

with parameter $\mu > 0$. Note that $\mu = \mathbb{E}[Y]$ here. This can be written as

$$P(Y=y) = \exp\{y\log\mu - \mu - \log(y!)\},\ y = 0, 1, 2, ...$$

We have here that $\mu = \mathbb{E}[Y]$ and $\theta = \log\mu$ is the natural parameter with $b(\theta) = e^\theta$, $a(\phi) = 1$ and $c(y) = -\log(y!)$. The link function here is $g(\mu) = \log\mu$.

In the canonical case (when ϕ is known) the model is $\theta_i = \eta_i$, $i = 1, ..., n$, with η_i being linear predictors specified in equation (11.1). In order to compute the ML estimate of $\boldsymbol{\beta}$ we need to maximize the corresponding log-likelihood function (given in (11.5)), that is to solve the problem

$$\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} Y_i \boldsymbol{x}_i'\boldsymbol{\beta} - b(\boldsymbol{x}_i'\boldsymbol{\beta}). \tag{11.8}$$

When $b(\cdot)$ is a convex function, the above problem (11.8) is convex. For the binomial and Poisson distributions the corresponding functions $b(\cdot)$ are convex.

12 Classification problem

Consider an $m \times 1$ random vector $\boldsymbol{X}$ of measurements. We want to classify $\boldsymbol{X}$ into one of two population π_1 or π_2. Let $p_1(\boldsymbol{x})$ and $p_2(\boldsymbol{x})$ be respective densities of populations π_1 and π_2. Suppose that the probability that an observation comes from population π_i is q_i, $i = 1, 2$. Consider regions $R_1 \subset \mathbb{R}^m$ and $R_2 = \mathbb{R}^m \setminus R_1$. If $\boldsymbol{X} \in R_1$ we classify $\boldsymbol{X}$ as from π_1, and if $\boldsymbol{X} \in R_2$ we classify $\boldsymbol{X}$ as from π_2. Then the probability of misclassification of an observation from π_1 is

$$\mathsf{Prob}(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2} p_1(\boldsymbol{x})d\boldsymbol{x}.$$

Similarly the probability of misclassification of an observation from π_2 is $\int_{R_1} p_2(\boldsymbol{x})d\boldsymbol{x}$. The expected loss of misclassification is

$$c_1 q_1 \int_{R_2} p_1(\boldsymbol{x})d\boldsymbol{x} + c_2 q_2 \int_{R_1} p_2(\boldsymbol{x})d\boldsymbol{x},$$

where c_i is the cost of misclassification of an observation from π_i, $i = 1, 2$.

Note that

$$\int_{R_1} p_2(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^m \setminus R_2} p_2(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathbb{R}^m} p_2(\boldsymbol{x})d\boldsymbol{x} - \int_{R_2} p_2(\boldsymbol{x})d\boldsymbol{x}.$$

Suppose that $c_1 = c_2 = 1$. Then the probability (expected loss) of misclassification is

$$q_1 \int_{R_2} p_1(\boldsymbol{x})d\boldsymbol{x} + q_2 \int_{R_1} p_2(\boldsymbol{x})d\boldsymbol{x} = \int_{R_2} [q_1 p_1(\boldsymbol{x}) - q_2 p_2(\boldsymbol{x})]d\boldsymbol{x} + q_2 \int_{\mathbb{R}^m} p_2(\boldsymbol{x})d\boldsymbol{x}.$$

Since $p_2(\cdot)$ is a probability density, we have that $\int_{\mathbb{R}^m} p_2(\boldsymbol{x})d\boldsymbol{x} = 1$, and hence

$$q_1 \int_{R_2} p_1(\boldsymbol{x})d\boldsymbol{x} + q_2 \int_{R_1} p_2(\boldsymbol{x})d\boldsymbol{x} = \int_{R_2} [q_1 p_1(\boldsymbol{x}) - q_2 p_2(\boldsymbol{x})]d\boldsymbol{x} + q_2.$$

It follows that the expected loss is minimized if

$$R_2 = \{\boldsymbol{x} \in \mathbb{R}^m : q_1 p_1(\boldsymbol{x}) - q_2 p_2(\boldsymbol{x}) < 0\}.$$

Or equivalently

$$R_1 = \left\{\boldsymbol{x} \in \mathbb{R}^m : p_1(\boldsymbol{x}) \geq \frac{q_2}{q_1} p_2(\boldsymbol{x})\right\}$$

and

$$R_2 = \left\{\boldsymbol{x} \in \mathbb{R}^m : p_1(\boldsymbol{x}) < \frac{q_2}{q_1} p_2(\boldsymbol{x})\right\}.$$

If the costs c_1 and c_2 are unequal, then the optimal regions are

$$R_1 = \left\{\boldsymbol{x} \in \mathbb{R}^m : p_1(\boldsymbol{x}) \geq \frac{c_2 q_2}{c_1 q_1} p_2(\boldsymbol{x})\right\}$$

and

$$R_2 = \left\{\boldsymbol{x} \in \mathbb{R}^m : p_1(\boldsymbol{x}) < \frac{c_2 q_2}{c_1 q_1} p_2(\boldsymbol{x})\right\}.$$

When

$$\mathsf{Prob}\left\{\boldsymbol{x} \in \mathbb{R}^m : p_1(\boldsymbol{x}) = \frac{c_2 q_2}{c_1 q_1} p_2(\boldsymbol{x}) | \pi_i\right\} = 0,\ i = 1, 2,$$

the optimal procedure is unique except for sets of probability zero.

12.1 Classification with normally distributed populations

Suppose that the populations π_1 and π_2 have multivariate normal distributions with equal covariance matrices, i.e., $\pi_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$. Then

$$p_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\big\{-(\boldsymbol{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)/2\big\},$$

and

$$\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})} = \exp\left\{-\tfrac{1}{2}\big[(\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)\big]\right\}.$$

Hence the optimal region is

$$R_1 = \big\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \leq -2\kappa\big\},$$

where $\kappa = \log(c_2 q_2 / c_1 q_1)$. Equivalently

$$R_1 = \left\{ \boldsymbol{x} : \boldsymbol{x}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq \tfrac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \kappa \right\}. \tag{12.1}$$

Note that if $\boldsymbol{X} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, then $\boldsymbol{X}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ has normal distribution with mean $\boldsymbol{\mu}_i' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and variance $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. The function $\boldsymbol{X}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is called *Fisher's discriminant function* and $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is called *Mahalanobis' distance* between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

12.2 Fisher discriminant analysis

Suppose that distribution of population π_i has mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$. Consider the following problem

$$\max_{\boldsymbol{d} \in \mathbb{R}^m} \left\{ g(\boldsymbol{d}) = \frac{(\boldsymbol{d}' \boldsymbol{\mu}_1 - \boldsymbol{d}' \boldsymbol{\mu}_2)^2}{\boldsymbol{d}' \boldsymbol{\Sigma}_1 \boldsymbol{d} + \boldsymbol{d}' \boldsymbol{\Sigma}_2 \boldsymbol{d}} \right\}. \tag{12.2}$$

Note that $\boldsymbol{d}' \boldsymbol{\mu}_i$ is the expected value and $\boldsymbol{d}' \boldsymbol{\Sigma}_i \boldsymbol{d}$ is the variance of $\boldsymbol{d}' \boldsymbol{X}$ for population π_i.

We can write function $g(\boldsymbol{d})$ as

$$g(\boldsymbol{d}) = \frac{\boldsymbol{d}' (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{d}}{\boldsymbol{d}' (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \boldsymbol{d}}.$$

Hence the optimal solution $\bar{\boldsymbol{d}}$ of problem (12.2) is defined by the equation

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{d} = \lambda (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2) \boldsymbol{d},$$

where λ is the maximal eigenvalue of the matrix $(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'$. This solution can be written as

$$\bar{\boldsymbol{d}} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{12.3}$$

In particular if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, then $\bar{\boldsymbol{d}} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

12.3 Several populations

Suppose that there are r populations $\pi_1, ..., \pi_r$ with respective means $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_r$ and covariance matrices $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_r$. Let q_i be the probability that the measurements vector $\boldsymbol{X}$ comes from population π_i, $i = 1, ..., r$ (we assume that $q_i > 0$, $i = 1, ..., r$). Consider

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] = q_1 \boldsymbol{\mu}_1 + ... + q_r \boldsymbol{\mu}_r$$

and

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{X}) &= \mathbb{E}[\boldsymbol{X} \boldsymbol{X}'] - \boldsymbol{\mu} \boldsymbol{\mu}' = q_1 (\boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1 \boldsymbol{\mu}_1') + ... + q_r (\boldsymbol{\Sigma}_r + \boldsymbol{\mu}_r \boldsymbol{\mu}_r') - \boldsymbol{\mu} \boldsymbol{\mu}' \\ &= \sum_{i=1}^{r} q_i \boldsymbol{\Sigma}_i + \sum_{i=1}^{r} q_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' = \boldsymbol{\Omega} + \boldsymbol{M}, \end{aligned}$$

where

$$\boldsymbol{\Omega} = \sum_{i=1}^{r} q_i \boldsymbol{\Sigma}_i$$

and

$$\boldsymbol{M} = \sum_{i=1}^{r} q_i (\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'.$$

Consider the following optimization problem

$$\max_{\boldsymbol{d}\in\mathbb{R}^m}\left\{g(\boldsymbol{d})=\frac{\boldsymbol{d}'\boldsymbol{M}\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{\Omega}\boldsymbol{d}}\right\}. \tag{12.4}$$

The optimal solution of the above problem (12.4) is defined by the equation

$$\boldsymbol{M}\boldsymbol{d}=\lambda_1\boldsymbol{\Omega}\boldsymbol{d},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ are eigenvalues of $\boldsymbol{\Omega}^{-1}\boldsymbol{M}$. That is, the optimal solution $\boldsymbol{d}_1$ of problem (12.4) is the eigenvector of $\boldsymbol{\Omega}^{-1}\boldsymbol{M}$ corresponding to its largest eigenvalue.

Next maximize $g(\boldsymbol{d})$ subject to $\boldsymbol{d}'\boldsymbol{\Omega}\boldsymbol{d}_1 = 0$. The solution of of this problem is given by eigenvector $\boldsymbol{d}_2$ of $\boldsymbol{\Omega}^{-1}\boldsymbol{M}$ corresponding to the eigenvalue λ_2. By continuing this process we obtain discriminant functions $\boldsymbol{d}_i'\boldsymbol{X}$, $i = 1, ..., r-1$. Note that $\text{rank}(\boldsymbol{M}) \leq r-1$, and hence $\lambda_r = ... = \lambda_m = 0$.

12.3.1 Mahalanobis distance

Mahalanobis distance between two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$, with respect to covariance matrix $\boldsymbol{\Sigma}$, is defined as

$$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$$

Assuming that covariance matrices $\boldsymbol{\Sigma}_1 = .. = \boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}$ are equal to each other, classify $\boldsymbol{X}$ in π_i if $d(\boldsymbol{X}, \boldsymbol{\mu}_i) < d(\boldsymbol{X}, \boldsymbol{\mu}_j)$ for all $j \neq i$.

12.4 Bayes and Logistic Regression classifies

Suppose that we have two populations π_1 and π_2. We consider $(Y, \boldsymbol{X})$ with $Y = 1$ if $\boldsymbol{X} \sim \pi_1$ and $Y = -1$ if $\boldsymbol{X} \sim \pi_2$. By Bayes formula we have that

$$\mathsf{Prob}(Y=1|\boldsymbol{X}=\boldsymbol{x})=\frac{p_1(\boldsymbol{x})q_1}{p_1(\boldsymbol{x})q_1+p_2(\boldsymbol{x})q_2},$$

where $q_1 = \mathsf{Prob}(Y = 1)$ and $q_2 = \mathsf{Prob}(Y = -1)$. We classify $\boldsymbol{X}$ in π_1 if $\mathsf{Prob}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > \mathsf{Prob}(Y = -1|\boldsymbol{X} = \boldsymbol{x})$, which is equivalent to $p_1(\boldsymbol{x})q_1 > p_2(\boldsymbol{x})q_2$.

Logistic regression approach. The ratio $\text{odd}(\boldsymbol{x}) = \frac{\mathsf{Prob}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathsf{Prob}(Y=-1|\boldsymbol{X}=\boldsymbol{x})}$ is called odds ratio. Logistic regression model:

$$\log \text{odd}(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}. \tag{12.5}$$

We classify $\boldsymbol{X}$ in π_1 if $\text{odd}(\boldsymbol{x}) > 1$. This is equivalent to $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x} > 0$.

Note that

$$\frac{\mathsf{Prob}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathsf{Prob}(Y=-1|\boldsymbol{X}=\boldsymbol{x})}=\frac{p_1(\boldsymbol{x})q_1}{p_2(\boldsymbol{x})q_2}.$$

In case of normal distributions with the same covariance matrix $\boldsymbol{\Sigma}$ we have that (see section 12.1)

$$\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}=\exp\left\{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+const\right\}.$$

In that case (assuming $q_1 = q_2$) equation (12.5) holds with $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

13 Support Vector Machines

Suppose that we have two populations π_1 and π_2 . Suppose further that we have training data $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i = 1$ if $\boldsymbol{x}_i \sim \pi_1$ and $y_i = -1$ if $\boldsymbol{x}_i \sim \pi_2$. We want to separate these populations by a hyperplane $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x} = 0$. That is, we classify an observation $\boldsymbol{x}$ according to the sign of $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}$, i.e., we classify $\boldsymbol{x} \sim \pi_1$ if $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x} > 0$, and $\boldsymbol{x} \sim \pi_2$ if $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x} < 0$.

The data sets are separable if there exist β_0 and $\boldsymbol{\beta}$ such that $y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) > 0$ for all $i = 1, ..., n$. Then the largest margin of separation can be obtained by solving the following problem

$$\max_{\beta_0, \boldsymbol{\beta}, \|\boldsymbol{\beta}\|=1} c \tag{13.6}$$

$$\text{subject to} \quad y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq c, \ i = 1, ..., n. \tag{13.7}$$

Alternatively, by making change of variables $c = 1/\|\boldsymbol{\beta}\|$ we can write this problem as

$$\min_{\beta_0, \boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2 \tag{13.8}$$

$$\text{subject to} \quad y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq 1, \ i = 1, ..., n. \tag{13.9}$$

If the data sets (classes) overlap we can proceed in a similar way allowing some points to be on the wrong side of the margin. By introducing slack variables $\xi_1, ..., \xi_n$ we can modify the constraints $y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq c$ as

$$y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq c - \xi_i, \ i = 1, ..., n, \tag{13.10}$$

or

$$y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq c(1 - \xi_i), \ i = 1, ..., n, \tag{13.11}$$

where $\xi_i \geq 0$, $i = 1, ..., n$, and $\sum_{i=1}^n \xi_i \leq const.$ Similar to (13.8)–(13.9) formulation (13.11) leads to the following optimization problem

$$\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \|\boldsymbol{\beta}\|^2 \tag{13.12}$$

$$\text{subject to} \quad y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq 1 - \xi_i, \ i = 1, ..., n, \tag{13.13}$$

$$\xi_i \geq 0, \ i = 1, ..., n, \tag{13.14}$$

$$\textstyle\sum_{i=1}^n \xi_i \leq const. \tag{13.15}$$

We can look at this from the following point of view. Suppose that we want to find the hyperplane such that the number of misclassified points is minimal. Note that a point $(y_i, \boldsymbol{x}_i)$ is misclassified if $y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) < 0$. That is we would like to solve the following problem

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \delta\big(-y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i)\big), \tag{13.16}$$

where $\delta(t) = 1$ if $t \geq 0$, and $\delta(t) = 0$ if $t < 0$. Problem (13.16) is a difficult combinatorial problem. Note that $\delta(t) \leq [1 + t]_+$, where $[a]_+ = \max\{0, a\}$. Therefore we can approximate problem (13.16) by the following *convex* problem

$$\min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n [1 - y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i)]_+ + c\|\boldsymbol{\beta}\|^2. \tag{13.17}$$

Equivalently we can formulate problem (13.17) as

$$\min_{\beta_0,\boldsymbol{\beta},\boldsymbol{\xi}} \quad \tfrac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \textstyle\sum_{i=1}^n \xi_i \tag{13.18}$$

$$\text{s.t.} \quad y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq 1 - \xi_i,\ i = 1, ..., n, \tag{13.19}$$

$$\xi_i \geq 0,\ i = 1, ..., n, \tag{13.20}$$

where $\gamma = c^{-1}$.

The Lagrangian of the above problem is

$$L(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \tfrac{1}{2}\|\boldsymbol{\beta}\|^2 + \gamma \sum_{i=1}^n \xi_i - \sum_{i=1}^n \lambda_i[y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) - (1 - \xi_i)] - \sum_{i=1}^n \mu_i\xi_i.$$

The Lagrangian dual of problem (13.18)– (13.20) is the problem

$$\max_{\boldsymbol{\lambda}\geq 0, \boldsymbol{\mu}\geq 0} \min_{\beta_0,\boldsymbol{\beta},\boldsymbol{\xi}\geq 0} L(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{13.21}$$

The corresponding Lagrangian-Wolfe dual is obtained by employing optimality conditions for the problem of minimization of $L(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ in (13.21). That is, by setting derivatives of the Lagrangian to zero, with respect to $\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}$, we have

$$\boldsymbol{\beta} = \sum_{i=1}^n \lambda_i y_i \boldsymbol{x}_i \tag{13.22}$$

$$0 = \sum_{i=1}^n \lambda_i y_i \tag{13.23}$$

$$\lambda_i = \gamma - \mu_i,\ i = 1, ..., n, \tag{13.24}$$

By substituting these equations into the Lagrangian we obtain the Lagrangian-Wolfe dual:

$$\max_{\boldsymbol{\lambda}} \quad \sum_{i=1}^n \lambda_i - \tfrac{1}{2}\sum_{i=1}^n \sum_{j=1}^n \lambda_i\lambda_j y_i y_j \boldsymbol{x}_i'\boldsymbol{x}_j \tag{13.25}$$

$$\text{s.t.} \quad 0 \leq \lambda_i \leq \gamma,\ i = 1, ..., n, \tag{13.26}$$

$$\sum_{i=1}^n \lambda_i y_i = 0. \tag{13.27}$$

We also have the following complementarity conditions for problem (13.18)– (13.20):

$$\lambda_i[y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) - (1 - \xi_i)] = 0,\ i = 1, ..., n, \tag{13.28}$$

$$\mu_i\xi_i = 0,\ i = 1, ..., n. \tag{13.29}$$

Given solution $\bar{\boldsymbol{\lambda}}$ of problem (13.25)–(13.27) the optimal $\boldsymbol{\beta}$ can be computed using equation (13.22), that is

$$\bar{\boldsymbol{\beta}} = \sum_{i=1}^n \bar{\lambda}_i y_i \boldsymbol{x}_i. \tag{13.30}$$

The complementarity conditions (13.29) mean that $\xi_i = 0$ if $\mu_i > 0$, and similarly for the complementarity conditions (13.28) . By (13.24) we have that $\mu_i > 0$ if $\lambda_i < \gamma$. Therefore by using equation (13.28), for $0 < \bar{\lambda}_i < \gamma$ the optimal β_0 can be computed by solving $y_i f(\boldsymbol{x}_i) = 1$, where $f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}$.

Suppose now that we want to make classification by using feature vectors $\boldsymbol{h}(\boldsymbol{x}_i)$, $i = 1, ..., n$, where $\boldsymbol{h}(\cdot) = (h_1(\cdot), ..., h_q(\cdot))' : \mathbb{R}^p \to \mathbb{R}^q$. We can approach this by solving the corresponding dual problem with replacing $\boldsymbol{x}_i$ with $\boldsymbol{h}(\boldsymbol{x}_i)$, $i = 1, ..., n$. That is the objective function in (13.25) is replaced by

$$\sum_{i=1}^{n} \lambda_i - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j \boldsymbol{h}(\boldsymbol{x}_i)' \boldsymbol{h}(\boldsymbol{x}_j). \tag{13.31}$$

Consequently, by using $\boldsymbol{\beta} = \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{h}(\boldsymbol{x}_i)$ (see (13.30)), the classification is performed according to the sign of

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}' \boldsymbol{h}(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \lambda_i y_i \boldsymbol{h}(\boldsymbol{x})' \boldsymbol{h}(\boldsymbol{x}). \tag{13.32}$$

Both expressions (13.31) and (13.32) are defined by the so-called kernel function

$$K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{x})' \boldsymbol{h}(\boldsymbol{z}) = \sum_{s=1}^{q} h_s(\boldsymbol{x}) h_s(\boldsymbol{z}). \tag{13.33}$$

In terms of the kernel function the objective function (13.31) can be written as

$$\sum_{i=1}^{n} \lambda_i - \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{13.34}$$

and the classifier (13.32) as

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \lambda_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i). \tag{13.35}$$

For example

$$K(\boldsymbol{x}, \boldsymbol{z}) = (1 + \boldsymbol{x}'\boldsymbol{z})^2 = \left(1 + \sum_{i=1}^{p} x_i z_i\right)^2$$

defines a quadratic separation.

Kernel function should be symmetric, i.e., $K(\boldsymbol{x}, \boldsymbol{z}) = K(\boldsymbol{z}, \boldsymbol{x})$, and positive definite, i.e., for any $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ the matrix $\boldsymbol{A} = [a_{ij}]$ with components $a_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ should be positive semidefinite, or in other words $\sum_{i,j=1}^{m} \lambda_i \lambda_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ should be nonnegative for any $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ and $\lambda_1, ..., \lambda_m$. Popular examples of kernels:

- Polynomial $K(\boldsymbol{x}, \boldsymbol{z}) = (1 + \boldsymbol{x}'\boldsymbol{z})^d$.
- Radial basis $K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{z}\|^2)$, $\gamma > 0$.
- Hyperbolic tangent $K(\boldsymbol{x}, \boldsymbol{z}) = \tanh(c_1 + c_2 \boldsymbol{x}'\boldsymbol{z})$, $c_1 < 0$, $c_2 > 0$, where $\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, $\sinh x = -i \sin(ix) = \frac{e^x - e^{-x}}{2}$, $\cosh(x) = \cos(ix) = \frac{e^x + e^{-x}}{2}$.

14 Principal components analysis

Consider an $m \times 1$ random vector $\boldsymbol{X}$ with $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$ and $\boldsymbol{\Sigma} = \mathrm{Cov}[\boldsymbol{X}]$. Let $\lambda_1 \geq \cdots \geq \lambda_m$ be the eigenvalues and $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$ be corresponding eigenvectors of $\boldsymbol{\Sigma}$, i.e., $\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$, $i = 1, ..., m$. We assume that the eigenvectors are normalized such that $\|\boldsymbol{e}_i\|^2 = \boldsymbol{e}_i' \boldsymbol{e}_i = 1$, $i = 1, ..., m$, and $\boldsymbol{e}_i' \boldsymbol{e}_j = 0$ for $i \neq j$. Recall that then (spectral decomposition)

$$\boldsymbol{\Sigma} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}' = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1' + ... + \lambda_m \boldsymbol{e}_m \boldsymbol{e}_m', \tag{14.1}$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_m)$ is diagonal matrix and $\boldsymbol{E} = [\boldsymbol{e}_1, ..., \boldsymbol{e}_m]$ is orthogonal matrix.

Suppose that we want to find a linear combinations $\boldsymbol{w}'\boldsymbol{X} = w_1 X_1 + ... + w_m X_m$ with largest variance. That is we want to solve the problem

$$\max_{\|\boldsymbol{w}\|=1} \text{Var}(\boldsymbol{w}'\boldsymbol{X}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}. \tag{14.2}$$

By (14.1) we have that

$$\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} = \boldsymbol{w}'\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}'\boldsymbol{w} = \boldsymbol{v}'\boldsymbol{\Lambda}\boldsymbol{v} = \lambda_1 v_1^2 + ... + \lambda_m v_m^2.$$

where $\boldsymbol{v} = \boldsymbol{E}'\boldsymbol{w}$. Note that

$$v_1^2 + ... + v_m^2 = \boldsymbol{v}'\boldsymbol{v} = \boldsymbol{w}'\boldsymbol{E}\boldsymbol{E}'\boldsymbol{w} = \boldsymbol{w}'\boldsymbol{w} = 1.$$

It follows that $\boldsymbol{v}'\boldsymbol{\Lambda}\boldsymbol{v}$ is maximized when $\boldsymbol{v} = (1, 0, ..., 0)'$. Hence solution of problem (14.2) is given by eigenvector $\boldsymbol{e}_1$ corresponding to the largest eigenvalue. Note that

$$\text{Var}(\boldsymbol{e}_1'\boldsymbol{X}) = \boldsymbol{e}_1'\boldsymbol{\Sigma}\boldsymbol{e}_1 = \lambda_1 \boldsymbol{e}_1'\boldsymbol{e}_1 = \lambda_1.$$

Given the first principal component $Y_1 = \boldsymbol{e}_1'\boldsymbol{X}$, suppose that we want to find $Y_2 = \boldsymbol{w}'\boldsymbol{X}$, with $\|\boldsymbol{w}\| = 1$, such that $\text{Cov}(Y_1, Y_2) = 0$ and Y_2 has the largest possible variance. Since

$$\text{Cov}(Y_1, Y_2) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{e}_1 = \lambda_1 \boldsymbol{w}'\boldsymbol{e}_1,$$

this means that is we want to solve the problem

$$\max_{\|\boldsymbol{w}\|=1} \text{Var}(\boldsymbol{w}'\boldsymbol{X}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \text{ subject to } \boldsymbol{w}'\boldsymbol{e}_1 = 0. \tag{14.3}$$

Again by the spectral decomposition (14.1), solution of this problem is $\boldsymbol{e}_2$.

And so on, variables $Y_i = \boldsymbol{e}_i'\boldsymbol{X}$, $i = 1, ..., m$, are called principal components of the data vector $\boldsymbol{X}$. Note that $\text{Var}(Y_i) = \lambda_i$, $i = 1, ..., m$, $\text{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ and

$$\sum_{i=1}^{m} \text{Var}(Y_i) = \sum_{i=1}^{m} \text{Var}(X_i) = \sum_{i=1}^{m} \lambda_i = \text{tr}(\boldsymbol{\Sigma}).$$

The true (population) covariance matrix $\boldsymbol{\Sigma}$ is unknown. Therefore the PC analysis usually performed on the sample covariance matrix $\boldsymbol{S}$. Let $\ell_1 \geq \cdots \geq \ell_m$ be the eigenvalues and $\boldsymbol{q}_1, ..., \boldsymbol{q}_m$ be corresponding orthonormal eigenvectors of $\boldsymbol{S}$, considered as estimates of the respective true eigenvalues and eigenvectors. What are statistical properties of these estimates?

14.1 Derivatives of eigenvalues and eigenvectors

Consider the linear space of symmetric $m \times m$ matrices, denoted $\mathbb{S}^{m\times m}$. Consider $\boldsymbol{A} \in \mathbb{S}^{m\times m}$ and its eigenvalues $\lambda_1 \geq ... \geq \lambda_m$ and the corresponding orthonormal eigenvectors $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$. Suppose that eigenvalue λ_i is *simple*, i.e., $\lambda_{i-1} > \lambda_i > \lambda_{i+1}$. Then $\lambda_i(\cdot)$ is continuous at $\boldsymbol{A}$. For $d\boldsymbol{A} \in \mathbb{S}^{m\times m}$,

$$(\boldsymbol{A} + d\boldsymbol{A})(\boldsymbol{e}_i + d\boldsymbol{e}_i) = (\lambda_i + d\lambda_i)(\boldsymbol{e}_i + d\boldsymbol{e}_i) \tag{14.4}$$

and

$$(\boldsymbol{A} + d\boldsymbol{A})(\boldsymbol{e}_i + d\boldsymbol{e}_i) = \boldsymbol{A}\boldsymbol{e}_i + (d\boldsymbol{A})\boldsymbol{e}_i + \boldsymbol{A}d\boldsymbol{e}_i + (d\boldsymbol{A})d\boldsymbol{e}_i. \tag{14.5}$$

By disregarding the high order terms $(d\lambda_i)d\boldsymbol{e}_i$ and $(d\boldsymbol{A})d\boldsymbol{e}_i$ in (14.4) and (14.5), and since $\boldsymbol{A}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$, we can write

$$(d\boldsymbol{A})\boldsymbol{e}_i + \boldsymbol{A}d\boldsymbol{e}_i = (d\lambda_i)\boldsymbol{e}_i + \lambda_i d\boldsymbol{e}_i. \tag{14.6}$$

Furthermore up to high order terms

$$(\boldsymbol{e}_i + d\boldsymbol{e}_i)'(\boldsymbol{e}_j + d\boldsymbol{e}_j) = (d\boldsymbol{e}_i)'\boldsymbol{e}_j + \boldsymbol{e}_i' d\boldsymbol{e}_j + \boldsymbol{e}_i'\boldsymbol{e}_j. \tag{14.7}$$

It follows that for $i = j$

$$\boldsymbol{e}_i' d\boldsymbol{e}_i = 0, \tag{14.8}$$

and for $i \neq j$, since $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$ and $(\boldsymbol{e}_i + d\boldsymbol{e}_i)'(\boldsymbol{e}_j + d\boldsymbol{e}_j) = 0$,

$$(d\boldsymbol{e}_i)'\boldsymbol{e}_j + \boldsymbol{e}_i' d\boldsymbol{e}_j = 0. \tag{14.9}$$

Consequently by multiplying both sides of (14.6) by $\boldsymbol{e}_i'$ and noting that $\boldsymbol{e}_i'\boldsymbol{e}_i = 1$, $\boldsymbol{e}_i' d\boldsymbol{e}_i = 0$ and $\boldsymbol{e}_i'\boldsymbol{A}d\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i' d\boldsymbol{e}_i = 0$, we obtain

$$d\lambda_i = \boldsymbol{e}_i'(d\boldsymbol{A})\boldsymbol{e}_i. \tag{14.10}$$

It is also possible to write (14.10) as

$$d\lambda_i = \mathrm{tr}(\boldsymbol{e}_i \boldsymbol{e}_i' d\boldsymbol{A}). \tag{14.11}$$

Now let us compute $d\boldsymbol{e}_i$. Since eigenvectors $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$ are orthonormal, they form a basis and hence we can write $d\boldsymbol{e}_i$ as linear combination $d\boldsymbol{e}_i = c_1\boldsymbol{e}_1 + ... + c_m\boldsymbol{e}_m$ with $c_j = \boldsymbol{e}_j' d\boldsymbol{e}_i$. For $i \neq j$ we have by (14.6) and since $\boldsymbol{e}_j'\boldsymbol{e}_i = 0$ that

$$\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i + \boldsymbol{e}_j'\boldsymbol{A}d\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_j' d\boldsymbol{e}_i, \tag{14.12}$$

and since $\boldsymbol{e}_j'\boldsymbol{A}d\boldsymbol{e}_i = \lambda_j \boldsymbol{e}_j' d\boldsymbol{e}_i$ it follows that

$$\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i = (\lambda_i - \lambda_j)\boldsymbol{e}_j' d\boldsymbol{e}_i. \tag{14.13}$$

This implies that

$$c_j = (\lambda_i - \lambda_j)^{-1}\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i,\ j \neq i. \tag{14.14}$$

For $j = i$ we have $c_i = \boldsymbol{e}_i' d\boldsymbol{e}_i = 0$. We obtain the following formula for the differential of $\boldsymbol{e}_i$:

$$d\boldsymbol{e}_i = \sum_{\substack{j=1 \\ j\neq i}}^{m} \left[\frac{\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i}{\lambda_i - \lambda_j} \right] \boldsymbol{e}_j. \tag{14.15}$$

14.2 Elements of matrix calculus

Kronecker product of matrices $\boldsymbol{A} = [a_{ij}]$ and $\boldsymbol{B} = [b_{ij}]$, of respective orders $p \times q$ and $r \times s$, is the $pr \times qs$ matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1q}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2q}\boldsymbol{B} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1}\boldsymbol{B} & a_{p2}\boldsymbol{B} & \cdots & a_{pq}\boldsymbol{B} \end{bmatrix}.$$

Vec-operator of $p \times q$ matrix $\boldsymbol{A}$ is $pq \times 1$ vector

$$\mathsf{vec}(\boldsymbol{A}) = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{a}_q \end{bmatrix},$$

where $\boldsymbol{a}_1, ..., \boldsymbol{a}_q$ are columns of $\boldsymbol{A}$.

Note the following matrix identities

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD}) \tag{14.16}$$

and

$$\mathsf{vec}(\boldsymbol{BXC}) = (\boldsymbol{C}' \otimes \boldsymbol{B})\mathsf{vec}(\boldsymbol{X}), \tag{14.17}$$

where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{X}$ are matrices of appropriate order. Also for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same order $p \times q$, and vectors $\boldsymbol{a} = \mathsf{vec}(\boldsymbol{A})$ and $\boldsymbol{b} = \mathsf{vec}(\boldsymbol{B})$,

$$\mathrm{tr}(\boldsymbol{A}'\boldsymbol{B}) = \sum_{i,j} a_{ij} b_{ij} = \boldsymbol{a}'\boldsymbol{b}. \tag{14.18}$$

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid sample. Assume that the distribution of $\boldsymbol{X}_i$ has finite fourth order moments. Let $\boldsymbol{s} = \mathrm{vec}(\boldsymbol{S})$ and $\boldsymbol{\sigma}_0 = \mathrm{vec}(\boldsymbol{\Sigma}_0)$, where $\boldsymbol{\Sigma}_0 = [\sigma_{ij}]$ is the population covariance matrix. Then by the CLT, $N^{1/2}(\boldsymbol{s} - \boldsymbol{\sigma}_0)$ converges in distribution to normal with zero mean vector and a covariance matrix $\boldsymbol{\Gamma}$ of order $m^2 \times m^2$. Note that $\mathrm{rank}(\boldsymbol{\Gamma}) \leq m(m+1)/2$. The typical element of matrix $\boldsymbol{\Gamma}$ is

$$\begin{aligned} [\boldsymbol{\Gamma}]_{ij,kl} &= \mathbb{E}\big\{[(X_i - \mu_i)(X_j - \mu_j) - \sigma_{ij}][(X_k - \mu_k)(X_l - \mu_l) - \sigma_{kl}]\big\} \\ &= \mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)(X_l - \mu_l)\big] - \sigma_{ij}\sigma_{kl}. \end{aligned}$$

In particular if $\boldsymbol{X}_i$ have normal distribution, then

$$\mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)(X_l - \mu_l)\big] = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk},$$

and hence

$$[\boldsymbol{\Gamma}]_{ij,kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}. \tag{14.19}$$

In a matrix form equations (14.19) can be written as

$$\boldsymbol{\Gamma} = 2\boldsymbol{M}_m(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0), \tag{14.20}$$

where $\boldsymbol{M}_m$ is the $m^2 \times m^2$ matrix given by

$$\boldsymbol{M}_m = \frac{1}{2}\left[\boldsymbol{I}_{m^2} + \sum_{i,j=1}^{m} (\boldsymbol{H}_{ij} \otimes \boldsymbol{H}'_{ij})\right],$$

with $\boldsymbol{H}_{ij}$ being $m \times m$ matrix with $h_{ij} = 1$ and all other elements zero. The matrix $\boldsymbol{M}_m$ has the following properties: (i) $\mathrm{rank}\boldsymbol{M}_m = m(m+1)/2$, (ii) $\boldsymbol{M}_m^2 = \boldsymbol{M}_m$, (iii) for any symmetric matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{M}_m(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma}) = (\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\boldsymbol{M}_m \text{ and } \boldsymbol{M}_m\mathrm{vec}(\boldsymbol{\Sigma}) = \mathrm{vec}(\boldsymbol{\Sigma}).$$

It follows that

$$\boldsymbol{\Gamma} = 2\boldsymbol{M}_m(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0)\boldsymbol{M}_m. \tag{14.21}$$

14.3 Asymptotics of PCA

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid sample from $N_m(\mu, \boldsymbol{\Sigma})$ and $\boldsymbol{S}$ be the corresponding sample covariance matrix. Let $\lambda_1 \geq \cdots \geq \lambda_m$ be the eigenvalues and $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$ be a corresponding set of orthonormal eigenvectors of $\boldsymbol{\Sigma}$, and $\ell_1 \geq \cdots \geq \ell_m$ be the eigenvalues and $\boldsymbol{q}_1, ..., \boldsymbol{q}_m$ be a corresponding set of orthonormal eigenvectors of $\boldsymbol{S}$.

Suppose that λ_i has multiplicity one. Let us show that $N^{1/2}(\ell_i - \lambda_i)$ and $N^{1/2}(\boldsymbol{q}_i - \boldsymbol{e}_i)$ are asymptotically normal (with mean zero) and asymptotically independent of each other, and that the asymptotic variance of $N^{1/2}(\ell_i - \lambda_i)$ is $2\lambda_i^2$ and the asymptotic covariance matrix of $N^{1/2}(\boldsymbol{q}_i - \boldsymbol{e}_i)$ is

$$\sum_{j=1,\, j\neq i}^{m} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \boldsymbol{e}_j \boldsymbol{e}_j'. \tag{14.22}$$

By the Delta Theorem and (14.10) we have that

$$N^{1/2}(\ell_i - \lambda_i) = \boldsymbol{e}_i'[N^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma})]\boldsymbol{e}_i + o_p(1),$$

and hence $N^{1/2}(\ell_i - \lambda_i)$ converges in distribution to $N(0, \sigma^2)$, where σ^2 can be calculated as follows. We have that

$$\boldsymbol{e}_i'[N^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma})]\boldsymbol{e}_i = \operatorname{tr}\left[N^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma})\boldsymbol{e}_i\boldsymbol{e}_i'\right] = [\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')]'[N^{1/2}(\boldsymbol{s} - \boldsymbol{\sigma})]$$

and hence

$$\sigma^2 = [\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')]'\boldsymbol{\Gamma}[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')] = 2[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')]'\boldsymbol{M}_m(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\boldsymbol{M}_m[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')].$$

Moreover, $\boldsymbol{M}_m[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')] = \operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')$ and

$$[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')]'(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')] = \operatorname{tr}[(\boldsymbol{e}_i\boldsymbol{e}_i')\boldsymbol{\Sigma}(\boldsymbol{e}_i\boldsymbol{e}_i')\boldsymbol{\Sigma}] = (\boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_i)(\boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_i) = \lambda_i^2.$$

Similarly, by (14.15)

$$N^{1/2}(\boldsymbol{q}_i - \boldsymbol{e}_i) = \sum_{j\neq i} a_j \boldsymbol{e}_j + o_p(1),$$

where

$$a_j = \frac{\boldsymbol{e}_j'[N^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma})]\boldsymbol{e}_i}{\lambda_i - \lambda_j} = (\lambda_i - \lambda_j)^{-1}[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_j')]'[N^{1/2}(\boldsymbol{s} - \boldsymbol{\sigma})].$$

The asymptotic covariance between a_j and a_k (for $j \neq i$ and $k \neq i$) is

$$\frac{[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_j')]'\boldsymbol{\Gamma}[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_k')]}{(\lambda_i - \lambda_j)^2} = \frac{2\operatorname{tr}[(\boldsymbol{e}_i\boldsymbol{e}_j')\boldsymbol{M}_m\boldsymbol{\Sigma}(\boldsymbol{e}_i\boldsymbol{e}_k')\boldsymbol{M}_m\boldsymbol{\Sigma}]}{(\lambda_i - \lambda_j)^2}. \tag{14.23}$$

Also $\boldsymbol{M}_m(\boldsymbol{e}_i\boldsymbol{e}_j') = \frac{1}{2}[(\boldsymbol{e}_i\boldsymbol{e}_j') + (\boldsymbol{e}_j\boldsymbol{e}_i')]$ and $\boldsymbol{M}_m(\boldsymbol{e}_i\boldsymbol{e}_k') = \frac{1}{2}[(\boldsymbol{e}_i\boldsymbol{e}_k') + (\boldsymbol{e}_k\boldsymbol{e}_i')]$. It follows that the right hand side of (14.23) is equal to

$$\frac{\operatorname{tr}[((\boldsymbol{e}_i\boldsymbol{e}_j') + (\boldsymbol{e}_j\boldsymbol{e}_i'))\boldsymbol{\Sigma}((\boldsymbol{e}_i\boldsymbol{e}_k') + (\boldsymbol{e}_k\boldsymbol{e}_i'))\boldsymbol{\Sigma}]}{2(\lambda_i - \lambda_j)^2}. \tag{14.24}$$

Moreover, we have that $\boldsymbol{e}_j'\boldsymbol{\Sigma}\boldsymbol{e}_k = \lambda_j\boldsymbol{e}_j'\boldsymbol{e}_k$ equals 0 if $j \neq k$, and λ_j if $j = k$. Therefore the expression in (14.24) equals 0 if $j \neq k$, and $\lambda_i\lambda_j/(\lambda_i - \lambda_j)^2$ if $j = k$. We obtain that the asymptotic covariance matrix of $a_j\boldsymbol{e}_j$ is $\frac{\lambda_i\lambda_j}{(\lambda_i - \lambda_j)^2}\boldsymbol{e}_j\boldsymbol{e}_j'$, and $a_j\boldsymbol{e}_j$ is asymptotically uncorrelated with $a_k\boldsymbol{e}_k$ for $j \neq k$. Formula (14.22) follows.

Finally, the asymptotic covariance between $N^{1/2}(\ell_i - \lambda_i)$ and a_j, $j \neq i$, is proportional to

$$[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_i')]\boldsymbol{\Gamma}[\operatorname{vec}(\boldsymbol{e}_i\boldsymbol{e}_j')] = \operatorname{tr}[(\boldsymbol{e}_i\boldsymbol{e}_i')\boldsymbol{\Sigma}(\boldsymbol{e}_i\boldsymbol{e}_j' + \boldsymbol{e}_j\boldsymbol{e}_i')\boldsymbol{\Sigma}] = 0,$$

and hence $N^{1/2}(\ell_i - \lambda_i)$ and $N^{1/2}(\boldsymbol{q}_i - \boldsymbol{e}_i)$ are asymptotically independent.

14.4 Singular value decomposition

Let $\boldsymbol{X}$ be an $m \times N$ matrix of rank r (note that $r \leq \min\{m, N\}$). Its singular value decomposition is

$$\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{W}' = \sigma_1 \boldsymbol{v}_1 \boldsymbol{w}_1' + ... + \sigma_r \boldsymbol{v}_r \boldsymbol{w}_r',$$

where $\boldsymbol{V} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_r]$ and $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_r]$ are matrices of order $m \times r$ and $N \times r$, respectively, such that $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_r$ and $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{I}_r$, and $\boldsymbol{D} = \text{diag}(\sigma_1, ..., \sigma_r)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Note that

$$\boldsymbol{X}\boldsymbol{X}' = \boldsymbol{V}\boldsymbol{D}\boldsymbol{W}'\boldsymbol{W}\boldsymbol{D}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}',$$

i.e., $\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}'$ is the spectral decomosition of the (symmetric positive semidefinite) $m \times m$ matrix $\boldsymbol{X}\boldsymbol{X}'$. It follows that σ_i^2 are the nonzero eigenvalues of $\boldsymbol{X}\boldsymbol{X}'$. Similarly $\boldsymbol{W}\boldsymbol{D}^2\boldsymbol{W}'$ is the spectral decomosition of the $N \times N$ matrix $\boldsymbol{X}'\boldsymbol{X}$.

For $\rho < r$ consider the (truncated) matrix $\boldsymbol{X}_\rho = \boldsymbol{V}_\rho \boldsymbol{D}_\rho \boldsymbol{W}_\rho'$, where $\boldsymbol{V}_\rho = [\boldsymbol{v}_1, ..., \boldsymbol{v}_\rho]$, $\boldsymbol{W}_\rho = [\boldsymbol{w}_1, ..., \boldsymbol{w}_\rho]$ and $\boldsymbol{D}_\rho = \text{diag}(\sigma_1, ..., \sigma_\rho)$. The matrix $\boldsymbol{X}_\rho$ is the nearest matrix of rank ρ to the original matrix $\boldsymbol{X}$, in the sense of the difference between the two having the smallest possible Frobenius norm (EckartYoung theorem). That is, solution of the minimization problem

$$\min_{Z \in \mathbb{R}^{m \times N}} \|\boldsymbol{X} - \boldsymbol{Z}\|_F \text{ s.t. } \text{rank}(\boldsymbol{Z}) \leq \rho$$

is $\bar{\boldsymbol{Z}} = \boldsymbol{X}_\rho$. Recall that Frobenius norm of a matrix $\boldsymbol{A}$ is

$$\|\boldsymbol{A}\|_F = \sqrt{\text{tr}(\boldsymbol{A}\boldsymbol{A}')} = \sqrt{\text{tr}(\boldsymbol{A}'\boldsymbol{A})} = \sqrt{\textstyle\sum_{i,j} a_{ij}^2}.$$

14.5 Factor analysis model

Consider an $m \times 1$ random vector $\boldsymbol{X}$ (of measurements) with $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}_i]$ and $\text{Cov}(\boldsymbol{X}_i) = \boldsymbol{\Sigma}$. The factor analysis model assumes that

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{u}, \tag{14.25}$$

where $\boldsymbol{\Lambda}$ is an $m \times k$ matrix (of factor loadings), $\boldsymbol{f}$ is a $k \times 1$ random vector (of factors) and $\boldsymbol{u}$ is an $m \times 1$ random vector (errors). It is assumed that: (i) $\mathbb{E}[\boldsymbol{f}] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{u}] = \boldsymbol{0}$, (ii) the errors are uncorrelated, i.e., $\text{Cov}(\boldsymbol{u})$ is diagonal, (iii) the factors and errors are uncorrelated, i.e., $\mathbb{E}[\boldsymbol{f}\boldsymbol{u}'] = \boldsymbol{0}$.

It follows then that

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{u})(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{u})'] = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{14.26}$$

where $\boldsymbol{\Phi} = \text{Cov}(\boldsymbol{f})$ and $\boldsymbol{\Psi} = \text{Cov}(\boldsymbol{u})$. In this model elements of the $m \times k$ matrix $\boldsymbol{\Lambda}$, elements of $k \times k$ matrix $\boldsymbol{\Phi}$ and diagonal elements of matrix $\boldsymbol{\Psi}$ are parameters. It is required that matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ should be positive semidefinite. Often it is assumed that $\boldsymbol{\Phi} = \boldsymbol{I}_k$, i.e., the factors are standardized. Then the model becomes

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{14.27}$$

Note that $\text{rank}(\boldsymbol{\Sigma} - \boldsymbol{\Psi}) = \text{rank}(\boldsymbol{\Lambda}) \leq k$.

14.6 Kernel PCA

Given data (sample) $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$, suppose that we want to represent data in terms of vectors $\boldsymbol{Z}_i = \boldsymbol{h}(\boldsymbol{X}_i)$, $i = 1, ..., N$, where $\boldsymbol{h}(\cdot) = (h_1(\cdot), ..., h_q(\cdot))' : \mathbb{R}^m \to \mathbb{R}^q$ is a (nonlinear) mapping. Suppose for the moment that $\bar{\boldsymbol{Z}} = N^{-1}\sum_{i=1}^N \boldsymbol{Z}_i = N^{-1}\sum_{i=1}^N \boldsymbol{h}(\boldsymbol{X}_i)$ is $\boldsymbol{0}$. Consider the corresponding estimator of the covariance matrix in the new feature space

$$\boldsymbol{C} = N^{-1}\sum_{i=1}^N \boldsymbol{Z}_i\boldsymbol{Z}_i' = N^{-1}\sum_{i=1}^N \boldsymbol{h}(\boldsymbol{X}_i)\boldsymbol{h}(\boldsymbol{X}_i)'.$$

Let $\lambda_1 \geq \cdots \geq \lambda_q$ be eigenvalues and $\boldsymbol{e}_1, ..., \boldsymbol{e}_q$ be corresponding orthonormal eigenvectors of the $q \times q$ matrix $\boldsymbol{C}$, i.e., $\boldsymbol{C}\boldsymbol{e}_s = \lambda_s\boldsymbol{e}_s$, $s = 1, ..., q$. We have that

$$\lambda_s\boldsymbol{e}_s = \boldsymbol{C}\boldsymbol{e}_s = N^{-1}\sum_{i=1}^N \boldsymbol{Z}_i\boldsymbol{Z}_i'\boldsymbol{e}_s,$$

and hence (for $\lambda_s \neq 0$)

$$\boldsymbol{e}_s = \frac{1}{\lambda_s N}\sum_{i=1}^N \alpha_{is}\boldsymbol{Z}_i, \tag{14.28}$$

where $\alpha_{is} = \boldsymbol{Z}_i'\boldsymbol{e}_s$, $s = 1, ..., q$, $i = 1, ..., N$. It follows by (14.28) that

$$\alpha_{is} = \frac{1}{\lambda_s N}\boldsymbol{Z}_i'\left(\sum_{j=1}^N \alpha_{js}\boldsymbol{Z}_j\right) = \frac{1}{\lambda_s N}\sum_{j=1}^N \alpha_{js}\boldsymbol{Z}_i'\boldsymbol{Z}_j = \frac{1}{\lambda_s N}\sum_{j=1}^N \alpha_{js}\boldsymbol{h}(\boldsymbol{X}_i)'\boldsymbol{h}(\boldsymbol{X}_j). \tag{14.29}$$

Consider kernel function (compare with (13.33)) $K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{x})'\boldsymbol{h}(\boldsymbol{z})$. In terms of the kernel function equation (14.29) can be written as

$$\sum_{j=1}^N \alpha_{js}K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \lambda_s N\alpha_{is}. \tag{14.30}$$

Consider $N \times N$ matrix $\boldsymbol{K}$ with components $\boldsymbol{K}_{ij} = K(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $i, j = 1, ..., N$. Equation (14.30) can be written as

$$\boldsymbol{K}\boldsymbol{\alpha}_s = \lambda_s N\boldsymbol{\alpha}_s, \ s = 1, ..., q, \tag{14.31}$$

where $\boldsymbol{\alpha}_s = (\alpha_{1s}, ..., \alpha_{Ns})'$. That is, $\boldsymbol{\alpha}_s$ are eigenvectors of matrix $\boldsymbol{K}$. These eigenvectors can be normalized as follows

$$1 = \boldsymbol{e}_s'\boldsymbol{e}_s = \frac{1}{\lambda_s^2 N^2}\left(\sum_{i=1}^N \alpha_{is}\boldsymbol{Z}_i'\right)\left(\sum_{j=1}^N \alpha_{js}\boldsymbol{Z}_j'\right) = \frac{1}{\lambda_s^2 N^2}\sum_{i,j=1}^N \alpha_{is}\alpha_{js}\boldsymbol{Z}_i'\boldsymbol{Z}_j = \frac{1}{\lambda_s^2 N^2}\sum_{i,j=1}^N \alpha_{is}\alpha_{js}K(\boldsymbol{X}_i, \boldsymbol{X}_j).$$

That is $\boldsymbol{\alpha}_s'\boldsymbol{K}\boldsymbol{\alpha}_s = \lambda_s^2 N^2$. Because of (14.31) this implies that $\boldsymbol{\alpha}_s'\boldsymbol{\alpha}_s = \lambda_s N$.

In order to apply this PCA procedure we need to compute the eigenvectors of matrix $\boldsymbol{K}$ corresponding to its largest eigenvalues. This will give us vectors $\boldsymbol{\alpha}_s$ and numbers λ_s. For a data point $\boldsymbol{X} \in \mathbb{R}^m$ its s-PCA component is $\boldsymbol{e}_s'\boldsymbol{h}(\boldsymbol{X})$. By (14.28) we have

$$\boldsymbol{e}_s'\boldsymbol{h}(\boldsymbol{X}) = \frac{1}{\lambda_s N}\sum_{i=1}^N \alpha_{is}\boldsymbol{h}(\boldsymbol{X}_i)'\boldsymbol{h}(\boldsymbol{X}) = \frac{1}{\lambda_s N}\sum_{i=1}^N \alpha_{is}K(\boldsymbol{X}_i, \boldsymbol{X}).$$

When $N^{-1}\sum_{i=1}^N \boldsymbol{h}(X_i) \neq \mathbf{0}$ we can make the following correction to the matrix $\boldsymbol{K}$:

$$\begin{aligned}\tilde{\boldsymbol{K}}_{ij} &= \left[\boldsymbol{h}(\boldsymbol{X}_i) - N^{-1}\sum_{k=1}^N \boldsymbol{h}(\boldsymbol{X}_k)\right]' \left[\boldsymbol{h}(\boldsymbol{X}_j) - N^{-1}\sum_{k=1}^N \boldsymbol{h}(\boldsymbol{X}_k)\right] \\ &= \boldsymbol{K}_{ij} - N^{-1}\sum_{k=1}^N \boldsymbol{K}_{ki} - N^{-1}\sum_{k=1}^N \boldsymbol{K}_{kj} + N^{-2}\sum_{k=1}^N\sum_{l=1}^N \boldsymbol{K}_{kl}.\end{aligned}$$

15 Gaussian Mixture Models

Let $y_i \in \{1, ..., K\}$ be one of K possible labels for data point $\boldsymbol{X}_i$, $i = 1, ..., N$. Assume that the pdf of the data $f(\boldsymbol{x}_i, y_i) = f(\boldsymbol{x}_i|y_i)p(y_i)$, is defined as follows: $p(y_i = j) = \pi_j$, $j = 1, ..., K$, and the conditional distributions $f(\boldsymbol{x}_i|y_i = j) \sim \mathcal{N}_m(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ are normal. The corresponding log-likelihood function is

$$\ell_N(\boldsymbol{\theta}) = \sum_{i=1}^N \log\left(\sum_{j=1}^K \pi_j \phi(\boldsymbol{X}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right),$$

where

$$\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\big\{-(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})/2\big\},$$

and $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K)$.

EM (Expectation-Maximization) algorithm
Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, $k = 1, ..., K$.

The Expectation step (E-step) Given current estimates of the parameters $\pi_1, ..., \pi_K$, $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$, $\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K$, evaluate (by the Bayes rule) the corresponding posterior probabilities of data point $\boldsymbol{X}_i$ being in cluster $k \in \{1, ..., K\}$:

$$w_{ik} = \frac{\pi_k \phi(\boldsymbol{X}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \phi(\boldsymbol{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)},\ i = 1, ..., N.$$

Note that $\sum_{k=1}^K w_{ik} = 1$ for all i.

The Maximization step (M-step) For $k = 1, ..., K$, set $N_k = \sum_{i=1}^N w_{ik}$, and update $\pi_k^{new} = N_k/N$, $\boldsymbol{\mu}_k^{new} = N_k^{-1}\sum_{i=1}^N w_{ik}\boldsymbol{X}_i$, and

$$\boldsymbol{\Sigma}_k^{new} = N_k^{-1}\sum_{i=1}^N w_{ik}(\boldsymbol{X}_i - \boldsymbol{\mu}_k)(\boldsymbol{X}_i - \boldsymbol{\mu}_k)'.$$

Note that $\sum_{k=1}^K N_k = \sum_{i=1}^N \sum_{k=1}^K w_{ik} = N$.

16 Von Mises statistical functionals

Let $X_1, ..., X_N \sim F(\cdot)$ be an iid sample. With the sample is associated the so called empirical probability measure (distribution) $\hat{F}_N = N^{-1}\sum_{i=1}^N \delta_{X_i}$, where δ_x denotes probability measure of mass 1 at the point x. When $X_1, ..., X_N$ are scalars (numbers) the empirical cdf $\hat{F}_N(x) = \frac{\#(X_i \leq x)}{N}$. That is, if the sample is arranged in the increasing order $X_{(1)} \leq \cdots X_{(N)}$, then $\hat{F}_N(x) = 0$ for $x < X_{(1)}$, $\hat{F}_N(x) = 1/N$ for $X_{(1)} \leq x < X_{(2)}$, $\hat{F}_N(x) = 2/N$ for $X_{(2)} \leq x < X_{(3)}$, and so on.

Function $\theta = T(F)$ of the distribution F is called statistical functional. Its sample estimate is $\hat{\theta} = T(\hat{F}_N)$. Consider the following examples.

- Expectation of a function:

$$T(F) = \mathbb{E}_F[h(X)] = \int h(x)dF(x).$$

  Its sample estimate

$$T(\hat{F}_N) = \mathbb{E}_{\hat{F}_N}[h(X)] = N^{-1}\sum_{i=1}^{N} h(X_i).$$

- Variance

$$T(F) = \mathrm{Var}(X) = \mathbb{E}_F[X^2] - (\mathbb{E}_F[X])^2.$$

  Its sample estimate

$$T(\hat{F}_N) = N^{-1}\sum_{i=1}^{N} X_i^2 - \bar{X}^2 = N^{-1}\sum_{i=1}^{N}(X_i - \bar{X})^2.$$

- Median

$$T(F) = F^{-1}(1/2).$$

  Its sample estimate

$$T(\hat{F}_N) = \hat{F}_N^{-1}(1/2).$$

- Solution of equation $\mathbb{E}_F[g(X,\theta)] = 0$. Its sample estimate is obtained as solution of equation $\mathbb{E}_{\hat{F}_N}[g(X,\hat{\theta})] = 0$, which is $\sum_{i=1}^{N} g(X_i,\hat{\theta}) = 0$.

It is known (Glivenko-Cantelli Theorem) that $\sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)|$ converges w.p.1 to 0 as N tends to infinity. If $T(\cdot)$ is continuous (in a certain sense), it follows then that $T(\hat{F}_N)$ converges to $T(F)$ w.p.1, i.e., $\hat{\theta} = T(\hat{F}_N)$ is a consistent estimator of $\theta = T(F)$.

Asymptotic normality For distributions F and G consider their convex combination

$$(1-t)F + tG = F + t(G-F), \quad t \in [0,1].$$

The directional derivative of $T(\cdot)$ at F in the direction $G - F$ is

$$T'(F, G-F) = \lim_{t\downarrow 0} \frac{T(F + t(G-F)) - T(F)}{t}.$$

That is, $T'(F, G-F)$ is the right side derivative of $F_t = (1-t)F + tG$ at $t = 0$. Suppose that $T'(F,\cdot)$ is linear (as a function of the direction) and let $G = \hat{F}_N$. Then

$$\begin{aligned} T'(F, \hat{F}_N - F) &= T'\left(F, N^{-1}\textstyle\sum_{i=1}^{N}\delta_{X_i} - F\right) = T'\left(F, N^{-1}\textstyle\sum_{i=1}^{N}[\delta_{X_i} - F]\right) \\ &= N^{-1}\textstyle\sum_{i=1}^{N} T'\left(F, \delta_{X_i} - F\right). \end{aligned}$$

Now

$$\hat{\theta} - \theta = T(\hat{F}_N) - T(F) \approx T'(F, \hat{F}_N - F) = N^{-1}\sum_{i=1}^{N} IC_{T,F}(X_i),$$

where

$$IC_{T,F}(x) = T'(F, \delta_x - F) = \lim_{t\downarrow 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t},$$

is the so called Influence Curve (or Influence Function).

Let us note that $\mathbb{E}_F\left[IC_{T,F}(X)\right]=0$. Indeed suppose for the moment that F has discrete distribution, i.e., $F=\sum_{i=1}^m p_i\delta_{x_i}$ for some x_i and probabilities $p_i>0$. Then

$$\mathbb{E}_F\left[IC_{T,F}(X)\right]=\sum_{i=1}^m p_i IC_{T,F}(x_i)=\sum_{i=1}^m p_i T'(F,\delta_{x_i}-F)=T'\Big(F,\sum_{i=1}^m p_i\delta_{x_i}-F\Big),$$

where the last equality holds by linearity of $T'(F,\cdot)$ and since $\sum_{i=1}^m p_i=1$. Since $\sum_{i=1}^m p_i\delta_{x_i}=F$ and $T'(F,F-F)=0$, it follows that $\mathbb{E}_F\left[IC_{T,F}(X)\right]=0$.

By the above

$$N^{1/2}\Big[T(\hat{F}_N)-T(F)\Big]\approx N^{-1/2}\sum_{i=1}^N IC_{T,F}(X_i).$$

Since $\mathbb{E}_F\left[IC_{T,F}(X_i)\right]=0$, we have by the CLT that $N^{-1/2}\sum_{i=1}^N IC_{T,F}(X_i)$ converges in distribution to normal with zero mean and variance

$$\sigma^2_{T,F}=\mathbb{E}_F\left[IC_{T,F}(X)^2\right]=\mathrm{Var}_F\left[IC_{T,F}(X)\right].$$

For example, consider the median functional $T(F)=F^{-1}(1/2)$ (here F is the cumulative distribution function). Let us compute its directional derivative $T'(F,G-F)$ for some cdf G. Let $F_t=(1-t)F+tG$ and consider $T(F_t)=F_t^{-1}(1/2)$. We have that $F_t(T(F_t))=1/2$, i.e.,

$$(1-t)F(T(F_t))+tG(T(F_t))=1/2.$$

Computing derivative of the above with respect to t gives

$$-F(T(F_t))+(1-t)\frac{dF(T(F_t))}{dt}+G(T(F_t))+t\frac{dG(T(F_t))}{dt}=0. \tag{16.1}$$

At $t=0$ we have that $F_0=F$ and

$$\frac{dF(T(F_t))}{dt}\Big|_{t=0}=f(\mathfrak{m})\frac{dT(F_t))}{dt}\Big|_{t=0}, \tag{16.2}$$

where $f(x)=dF(x)/dx$ is the density function (assuming it exists) and $\mathfrak{m}=T(F)=F^{-1}(1/2)$ is the (population) median. Equation (16.1) (for $t=0$) together with (16.1) imply that

$$-F(\mathfrak{m})+G(\mathfrak{m})+f(\mathfrak{m})\frac{dT(F_t))}{dt}\Big|_{t=0}=0,$$

and hence (since $F(\mathfrak{m})=1/2$)

$$T'(F,G-F)=\frac{dT(F_t))}{dt}\Big|_{t=0}=\frac{1/2-G(\mathfrak{m})}{f(\mathfrak{m})}$$

We obtain that

$$IC_{T,F}(x)=T'(F,\delta_x-F)=\frac{1/2-\delta_x(\mathfrak{m})}{f(\mathfrak{m})},$$

where δ_x is the cdf such that $\delta_x(t)=0$ for $t<x$, and $\delta_x(t)=1$ for $t\geq x$.

Note that $\mathbb{E}_F\left[IC_{T,F}(X)\right]=0$ (as it should be), since $\mathbb{E}_F\left[\delta_X(\mathfrak{m})\right]=P(X\leq\mathfrak{m})=1/2$. Also $\mathrm{Var}_F\left[\delta_X(\mathfrak{m})\right]=1/2-1/4=1/4$ and hence

$$\mathrm{Var}_F\left[IC_{T,F}(X)\right]=\frac{\mathrm{Var}_F\left[\delta_X(\mathfrak{m})\right]}{f(\mathfrak{m})^2}=\frac{1}{4f(\mathfrak{m})^2}.$$

We obtain that $N^{1/2}\Big[T(\hat{F}_N)-T(F)\Big]$ converges in distribution to normal $N\big(0,\frac{1}{4f(\mathfrak{m})^2}\big)$. That is the sample median has approximately normal distribution with variance $\frac{1}{4Nf(\mathfrak{m})^2}$, provided that the population median $\mathfrak{m}$ is defined uniquely and the distribution has density $f(\mathfrak{m})=F'(\mathfrak{m})$.

Finite sample interpretation of the influence curve. By adding one more observation X_{N+1}, we have that

$$\hat{F}_{N+1}(\cdot) = \frac{N}{N+1}\hat{F}_N(\cdot) + \frac{1}{N+1}\delta_{X_{N+1}}(\cdot) = (1-t)\hat{F}_N(\cdot) + t\delta_{X_{N+1}}(\cdot),$$

where $t = 1/(N+1)$. Hence we can write

$$\hat{\theta}_{N+1} \approx \hat{\theta}_N + \frac{1}{N+1} IC_{T,\hat{F}_N}(X_{N+1}).$$

17 Bootstrap

Let $\hat{\theta} = \hat{\theta}(X_1, ..., X_N)$ be an estimator which is a function of sample $X_1, ..., X_N$. Suppose that we want to evaluate statistical properties of that estimator without assuming a parametric model. For example we would like to construct two sided 95% confidence interval for this estimator. This means that we need to evaluate 2.5% and 97.5% quantiles of the distribution of $\hat{\theta}$. Note that both quantiles are functions of the true distribution F of the sample. If we knew the true distribution F we can proceed by using the so called Monte Carlo sampling techniques. That is, we generate a sample $X_1^*, ..., X_N^*$ from F and compute $\hat{\theta}^* = \hat{\theta}(X_1^*, ..., X_N^*)$. We repeat this procedure independently M times. In that way we generate M independent replications $\hat{\theta}_1^*, ..., \hat{\theta}_M^*$ of the random variable $\hat{\theta}$. In that way, for sufficiently large M, we can accurately reconstruct the true distribution of $\hat{\theta}$, and hence to evaluate the required quantiles, or some other parameters. For example we can evaluate variance of $\hat{\theta}$ as

$$\widehat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{M-1}\sum_{m=1}^{m}(\hat{\theta}_m^* - \bar{\theta}^*)^2,$$

where $\bar{\theta}^* = \frac{1}{M}\sum_{m=1}^{m}\hat{\theta}_m^*$.

Of course the true distribution F is not known. So we replace it by the empirical distribution $\hat{F}_N$.

18 Spatial statistics

Consider a (real valued) function $Z(\boldsymbol{x})$ of $x \in \mathbb{R}^d$. Given values (observations, measurements) of $Z(\cdot)$ at some points, we would like to evaluate (to estimate) value of $Z(\boldsymbol{x})$ at a given point $x = x^*$. As a modeling approach we view $Z(\boldsymbol{x})$ as a random process. It is said that $Z(\boldsymbol{x})$ is stationary if for any points $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \in \mathbb{R}^d$ and $\boldsymbol{h} \in \mathbb{R}^d$, random vector $(Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_m))$ has the same distribution as $(Z(\boldsymbol{x}_1 + \boldsymbol{h}), ..., Z(\boldsymbol{x}_m + \boldsymbol{h}))$. This definition of stationarity is too general for practical use. It is said that $Z(\boldsymbol{x})$ is second order (or weakly) stationary if its mean $\mathbb{E}[Z(\boldsymbol{x})]$ is constant (independent of $\boldsymbol{x}$), and its covariance function $c(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{y}))$ has the property that for any $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{h} \in \mathbb{R}^d$ it follows that $c(\boldsymbol{x} + \boldsymbol{h}, \boldsymbol{y} + \boldsymbol{h}) = c(\boldsymbol{x}, \boldsymbol{y})$. Of course any stationary process is second order stationary provided it has finite second order moments. By taking $\boldsymbol{h} = -\boldsymbol{y}$ we have then that $c(\boldsymbol{x}, \boldsymbol{y}) = c(\boldsymbol{x} - \boldsymbol{y}, \boldsymbol{0})$. That is, for the second order stationary process the covariance function depends on the difference $\boldsymbol{x} - \boldsymbol{y}$. So we use notation $c(\boldsymbol{x} - \boldsymbol{y}) = \mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{y}))$ for the (auto)covariance function.

The autocovariance function $c(\cdot)$ has the following properties. It is symmetric, i.e., $c(\boldsymbol{h}) = c(-\boldsymbol{h})$, this follows from that $\mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{y})) = \mathrm{Cov}(Z(\boldsymbol{y}), Z(\boldsymbol{x}))$. Since $c(\boldsymbol{0}) = \mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x})) = \mathrm{Var}(Z(\boldsymbol{x}))$, it follows that $c(\boldsymbol{0}) > 0$. We have that $|\mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{y}))| \leq \sqrt{\mathrm{Var}(Z(\boldsymbol{x}))}\sqrt{\mathrm{Var}(Z(\boldsymbol{y}))}$ and hence $|c(\boldsymbol{h})| \leq c(\boldsymbol{0})$ for all $\boldsymbol{h} \in \mathbb{R}^d$. The function $c(\cdot)$ should be positive definite. That is

for any $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \in \mathbb{R}^d$ the covariance matrix of $(Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_m))$ should be positive semidefinite, i.e., the $m \times m$ matrix with entries $a_{ij} = c(\boldsymbol{x}_i - \boldsymbol{x}_j)$, $i, j = 1, ..., m$, should be positive semidefinite.

The semivariogram of (stationary) process $Z(\boldsymbol{x})$ is defined as

$$\gamma(\boldsymbol{h}) = \tfrac{1}{2}\mathbb{E}\big[|Z(\boldsymbol{x}+\boldsymbol{h}) - Z(\boldsymbol{x})|^2\big].$$

Note that we can assume that $\mathbb{E}[Z(\boldsymbol{h})] = 0$ and hence $c(\boldsymbol{0}) = \mathrm{Var}(Z(\boldsymbol{h})) = \mathbb{E}\big[Z(\boldsymbol{h})^2\big]$, and thus

$$\gamma(\boldsymbol{h}) = \tfrac{1}{2}\mathbb{E}\big[Z(\boldsymbol{x}+\boldsymbol{h})^2 + Z(\boldsymbol{x})^2 - 2Z(\boldsymbol{x}+\boldsymbol{h})Z(\boldsymbol{x})\big] = c(\boldsymbol{0}) - c(\boldsymbol{h}).$$

Consider $m \times m$ matrix $\boldsymbol{\Gamma}$ with entries $\Gamma_{ij} = \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j)$, $i, j = 1, ..., m$. Note that $\Gamma_{ij} = c_0 - c_{ij}$, where $c_0 = c(\boldsymbol{0})$ and $c_{ij} = c(\boldsymbol{x}_i - \boldsymbol{x}_j)$. In matrix form this can be written as $\boldsymbol{\Gamma} = c_0 \boldsymbol{1}_m \boldsymbol{1}_m' - \boldsymbol{C}$, where $\boldsymbol{C}$ is $m \times m$ matrix with entries c_{ij}.

Given observations $Z(\boldsymbol{x}_1), ..., Z(\boldsymbol{x}_n)$ consider the linear predictor

$$\hat{Z}(\boldsymbol{x}) = \sum_{i=1}^n w_i Z(\boldsymbol{x}_i).$$

We have that

$$\mathbb{E}[\hat{Z}(\boldsymbol{x})] = \sum_{i=1}^n w_i \mathbb{E}[Z(\boldsymbol{x}_i)] = \mu \sum_{i=1}^n w_i,$$

where μ is the mean of the process. Therefore $\hat{Z}(\boldsymbol{x})$ is unbiased iff $\sum_{i=1}^n w_i = 1$. It is said that $\hat{Z}(\boldsymbol{x})$ is the Best Linear Unbiased Predictor (BLUP) if the weights w_i are chosen to minimize variance of the error $\hat{Z}(\boldsymbol{x}) - Z(\boldsymbol{x})$. Now (since $\sum_{i=1}^n w_i = 1$)

$$\mathrm{Var}(\hat{Z}(\boldsymbol{x}) - Z(\boldsymbol{x})) = \mathrm{Var}\left[\sum_{i=1}^n w_i (Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}))\right],$$

and

$$\mathrm{Cov}(Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}), Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}) = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)) - \mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x}_i)) - \mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x}_j)) + c(\boldsymbol{0}).$$

Moreover

$$\mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)) = c(\boldsymbol{0}) - \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j) = c_0 - \Gamma_{ij}.$$

In matrix form we can write this as

$$\mathrm{Var}(\hat{Z}(\boldsymbol{x}) - Z(\boldsymbol{x})) = -\boldsymbol{w}'\boldsymbol{\Gamma}\boldsymbol{w} + 2\boldsymbol{g}'\boldsymbol{w},$$

where $\Gamma_{ij} = \gamma(\boldsymbol{x}_i - \boldsymbol{x}_j)$ and $g_i = \gamma(\boldsymbol{x} - \boldsymbol{x}_i)$. The BLUP is solution of the problem

$$\min_{\boldsymbol{w}} -\boldsymbol{w}'\boldsymbol{\Gamma}\boldsymbol{w} + 2\boldsymbol{g}'\boldsymbol{w} \quad \text{subject to } \sum_{i=1}^n w_i = 1.$$

By using method of Lagrange multipliers this can be written as the following system of $n+1$ linear equations

$$\begin{bmatrix} \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_1) & \cdots & \gamma(\boldsymbol{x}_1 - \boldsymbol{x}_n) & 1 \\ \cdots & \cdots & \cdots & \\ \gamma(\boldsymbol{x}_n - \boldsymbol{x}_1) & \cdots & \gamma(\boldsymbol{x}_n - \boldsymbol{x}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ \cdots \\ w_n \\ \lambda \end{bmatrix} = \begin{bmatrix} \gamma(\boldsymbol{x} - \boldsymbol{x}_1) \\ \cdots \\ \gamma(\boldsymbol{x} - \boldsymbol{x}_n) \\ 1 \end{bmatrix}$$

with $n+1$ unknowns $w_1, ..., w_n, \lambda$.

It is said that the model is isotropic if $\gamma(\boldsymbol{h})$ is a function of $\|\boldsymbol{h}\|$. In that case the semivariogram $\gamma(h)$ becomes a function of one dimensional variable $h = \|\boldsymbol{h}\|$. The following are some popular parametric models of semivariograms.

Linear $\gamma(0) = 0$ and $\gamma(h) = c_0 + bh$ for $h > 0$, where $c_0 \geq 0$ and $b > 0$ are parameters. This model is valid for any dimension d. Note that here $\lim_{h \downarrow 0} \gamma(h) = c_0$ with c_0 could be strictly positive. Value $\lim_{h \downarrow 0} \gamma(h)$ is called the nugget effect.

Exponential model $\gamma(0) = 0$ and $\gamma(h) = c_0 + c_\ell(1 - e^{-h/a_\ell})$ for $h > 0$, where $c_0 \geq 0$, $c_\ell > 0$ and $a_\ell > 0$. This model is valid for any dimension d.

Note that both models have nugget c_0, and in the linear model the semivariogram is unbounded, while in the exponential model the semivariogram is bounded by $c_0 + c_\ell$.

19 Spherical and elliptical distributions

An $m \times 1$ random vector $\boldsymbol{X}$ is said to have *spherical* distribution if $\boldsymbol{X}$ and $\boldsymbol{TX}$ have the same distribution for any $m \times m$ orthogonal matrix $\boldsymbol{T}$.

Examples

(i) Normal distribution $\boldsymbol{X} \sim N_m(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m)$. The corresponding density function

$$f(\boldsymbol{x}) = \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(-\tfrac{1}{2}\sigma^{-2}\boldsymbol{x}'\boldsymbol{x}\right).$$

(ii) ε-contaminated normal distribution, with pdf $(1-\varepsilon)f_1(\boldsymbol{x}) + \varepsilon f_2(\boldsymbol{x})$, $\varepsilon \in [0,1]$, where $f_i(\cdot)$ is pdf of $N_m(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I}_m)$, $i = 1, 2$.

(iii) Multivariate t-distribution with n degrees of freedom. Its pdf is

$$f(\boldsymbol{x}) = \frac{\Gamma[\frac{1}{2}(n+m)]}{\Gamma(\frac{1}{2}n)(\pi n)^{m/2}} \frac{1}{(1 + n^{-1}\boldsymbol{x}'\boldsymbol{x})^{(n+m)/2}},$$

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

This is distribution of random vector $\boldsymbol{X} = Z^{-1/2}n^{1/2}\boldsymbol{Y}$, where $Z \sim \chi^2_n$ and $\boldsymbol{Y} \sim N_m(\boldsymbol{0}, \boldsymbol{I}_m)$, and Z and $\boldsymbol{Y}$ are independent.

Spherical distributions can be generated in the following way. Let $X_1, ..., X_m$ be random variables such that conditional on random variable $Z > 0$, $Z \sim G(\cdot)$, these variables are iid $N(0, Z)$. Then the pdf of random vector $\boldsymbol{X} = (X_1, ..., X_m)'$ is

$$f(\boldsymbol{x}) = \int_o^\infty (2\pi z)^{-m/2} \exp\left(-\tfrac{1}{2}z^{-1}\boldsymbol{x}'\boldsymbol{x}\right) dG(z).$$

This is scale mixture of normal distributions. In particular, if Z can have two possible values σ_1^2 and σ_2^2 with respective probabilities $1-\varepsilon$ and ε, then this is the ε-contaminated normal distribution. If $Z \sim n/\chi^2_n$, then $\boldsymbol{X}$ has m-variate t-distribution with n degrees of freedom.

It is said that an $m \times 1$ random vector $\boldsymbol{X}$ has *elliptical* distribution with parameters $\boldsymbol{\mu} \in \mathbb{R}^m$ and symmetric positive definite $m \times m$ matrix $\boldsymbol{V} = [v_{ij}]_{i,j=1,...,m}$ if its pdf is

$$f(\boldsymbol{x}) = c_m |\boldsymbol{V}|^{-1/2} h\left((\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

for some function $h : \mathbb{R} \to \mathbb{R}_+$. We use notation $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ for elliptical distributions. For $h(t) = e^{-t/2}$ this is normal $N(\boldsymbol{\mu}, \boldsymbol{V})$ distribution. Note that if $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$, then $\boldsymbol{Y} = \boldsymbol{V}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})$ has spherical distribution.

Recall that characteristic function of a random vector $\boldsymbol{X}$ is $\phi(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{X})]$, where $i^2 = -1$ and $e^{i\theta} = \cos\theta + i\sin\theta$. If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\phi(\boldsymbol{t}) = \exp(i\boldsymbol{t}'\boldsymbol{\mu} - \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2)$. Note that

$$\partial\phi(\boldsymbol{0})/\partial\boldsymbol{t} = i\mathbb{E}[\boldsymbol{X}] \tag{19.1}$$

and

$$\partial^2\phi(\boldsymbol{0})/\partial\boldsymbol{t}\partial\boldsymbol{t}' = -\mathbb{E}[\boldsymbol{X}\boldsymbol{X}'] = -\boldsymbol{\mu}\boldsymbol{\mu}' - \mathrm{Cov}(\boldsymbol{X}). \tag{19.2}$$

If $\boldsymbol{X}$ has spherical distribution, then for any orthogonal matrix $\boldsymbol{T}$ we have

$$\phi(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{X})] = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{T}\boldsymbol{X})] = \mathbb{E}[\exp(i(\boldsymbol{T}'\boldsymbol{t})'\boldsymbol{X})] = \phi(\boldsymbol{T}'\boldsymbol{t}).$$

It follows that $\phi(\boldsymbol{t})$ is a function of $\boldsymbol{t}'\boldsymbol{t}$, i.e.,

$$\phi(\boldsymbol{t}) = \psi(\boldsymbol{t}'\boldsymbol{t}) \tag{19.3}$$

for some function $\psi(\cdot)$.

If $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$, then $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{V}^{1/2}\boldsymbol{Y}$, where $\boldsymbol{Y}$ has spherical distribution. Hence the characteristic function of $\boldsymbol{X}$ can be written as

$$\phi(\boldsymbol{t}) = \mathbb{E}\left[\exp\left(i\boldsymbol{t}'(\boldsymbol{\mu} + \boldsymbol{V}^{1/2}\boldsymbol{Y})\right)\right] = \exp(i\boldsymbol{t}'\boldsymbol{\mu})\mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{V}^{1/2}\boldsymbol{Y})],$$

and since $\boldsymbol{Y}$ has spherical distribution we have by (19.3) that

$$\mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{V}^{1/2}\boldsymbol{Y}] = \psi((\boldsymbol{V}^{1/2}\boldsymbol{t})'(\boldsymbol{V}^{1/2}\boldsymbol{t})) = \psi(\boldsymbol{t}'\boldsymbol{V}\boldsymbol{t}).$$

That is, the characteristic function of $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ can be represented in the form

$$\phi(\boldsymbol{t}) = \exp(i\boldsymbol{t}'\boldsymbol{\mu})\psi(\boldsymbol{t}'\boldsymbol{V}\boldsymbol{t}) \tag{19.4}$$

for some function $\psi(\cdot)$. It follows from (19.4) together with (19.1) and (19.2), that if $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$, then $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \alpha\boldsymbol{V}$, where $\alpha = -2\psi'(0)$. In particular this implies that

$$\mathrm{Corr}(X_i, X_j) = \frac{v_{ij}}{\sqrt{v_{ii}v_{jj}}}. \tag{19.5}$$

Let $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ be partitioned $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$ with the corresponding partitioning of $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}$. Then it follows from (19.4) that the characteristic function of $\boldsymbol{X}_1$ is

$$\phi_1(\boldsymbol{t}_1) = \exp(i\boldsymbol{t}_1'\boldsymbol{\mu}_1)\psi(\boldsymbol{t}_1'\boldsymbol{V}_{11}\boldsymbol{t}_1). \tag{19.6}$$

Theorem 19.1 *Let $\boldsymbol{X} = (X_1, ..., X_m)' \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{V} = \mathrm{diag}(v_{11}, ..., v_{mm})$. If $X_1, ..., X_m$ are all independent, then $\boldsymbol{X}$ has multivariate normal distribution.*

Proof. Without loss of generality we can assume that $\boldsymbol{\mu} = \boldsymbol{0}$, and hence the the characteristic function of $\boldsymbol{X}_1$ is

$$\phi(\boldsymbol{t}) = \psi(\boldsymbol{t}'\boldsymbol{V}\boldsymbol{t}) = \psi\left(\sum_{i=1}^{m} t_i^2 v_{ii}\right).$$

Denote $u_i = t_i v_{ii}^{1/2}$. Since $X_1, ..., X_m$ are independent, we have that

$$\phi(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{X})] = \mathbb{E}\left[\prod_{i=1}^{m} e^{it_i X_i}\right] = \prod_{i=1}^{m} \mathbb{E}\left[e^{it_i X_i}\right] = \prod_{i=1}^{m} \phi_i(t_i),$$

where $\phi_i(t_i)$ is the characteristic function of X_i, $i = 1, ..., m$. It follows by (19.6) that

$$\psi\left(\sum_{i=1}^{m} u_i^2\right) = \prod_{i=1}^{m} \psi(u_i^2). \tag{19.7}$$

In turn equation (19.7) implies that $\psi(z) = e^{kz}$ for some k. That is $\phi(\boldsymbol{t}) = \exp(k\boldsymbol{t}'\boldsymbol{V}\boldsymbol{t})$. ∎

19.1 Multivariate cumulants

Consider a random variable X. Let

$$\log \mathbb{E}[e^{tX}] = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} \tag{19.8}$$

be Taylor expansion of its log-moments generating function (note that for $t = 0$ this function is 0). The coefficient κ_n is called n-th cumulant of X. Since $\mathbb{E}[e^{tX}]$ may not exist, it is preferable to define cumulants in terms of the characteristic function as

$$\log \mathbb{E}[e^{itX}] = \sum_{n=1}^{\infty} \kappa_n \frac{(it)^n}{n!}. \tag{19.9}$$

We have that

$$\frac{\partial \log \mathbb{E}[e^{tX}]}{\partial t} = \frac{1}{\mathbb{E}[e^{tX}]} \mathbb{E}[X e^{tX}],$$

and hence $\kappa_1 = \mathbb{E}[X]$. Denote $\mu_i = \mathbb{E}[X^i]$ the i-th moment of X. Then

$$\begin{aligned}
\kappa_1 &= \mu_1 = \mathbb{E}[X], \\
\kappa_2 &= \mu_2 - \mu_1^2 = \mathrm{Var}(X), \\
\kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3, \\
\kappa_4 &= \mu_4 - 4\mu_1\mu_3 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4.
\end{aligned}$$

Skewness of X is defined as

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}}, \tag{19.10}$$

kurtosis of X is defined as

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2}. \tag{19.11}$$

If distribution of X is symmetrical around its mean, then $\gamma_1 = 0$. If $X \sim N(\mu, \sigma^2)$, then $\mu_4 = 3\mu^2$, and hence its kurtosis $\gamma_2 = 0$.

If X and Y are two independent random variables, then

$$\log \mathbb{E}[e^{it(X+Y)}] = \log \mathbb{E}[e^{itX}] + \log \mathbb{E}[e^{itY}],$$

and hence cumulants of $X+Y$ are equal to the sum of the respective cumulants of X and Y. That is, for *independent* random variables $X_1, ..., X_n$, we know that $\mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i)$.

Similarly $\kappa_3(\sum_{i=1}^n X_i) = \sum_{i=1}^n \kappa_3(X_i)$ and $\kappa_4(\sum_{i=1}^n X_i) = \sum_{i=1}^n \kappa_4(X_i)$.

Consider now random vector $\boldsymbol{X} = (X_1, ..., X_m)'$. Let $\phi_j(t_j)$ be the characteristic function of X_j. The cumulants of X_j are defined by

$$\log \phi_j(t_j) = \sum_{n=1}^{\infty} \kappa_n^j \frac{(it_j)^n}{n!}.$$

Mixed cumulants:

$$\log \phi_{j\ell}(t_j, t_\ell) = \sum_{n_1=1, n_2=1}^{\infty} \kappa_{n_1 n_2}^{j\ell} \frac{(it_j)^{n_1}(it_\ell)^{n_2}}{n_1! n_2!},$$

and so on.

Suppose that $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ has elliptical distribution. Then marginal distributions of X_j have zero skewness and the same kurtosis

$$\gamma_2^j = \frac{3[\psi''(0) - \psi'(0)^2]}{\psi'(0)^2}.$$

Denote $\kappa = \gamma_2^j/3$. Forth order cumulants of $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ are

$$\kappa_{1111}^{ijk\ell} = \kappa(\sigma_{ij}\sigma_{k\ell} + \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}).$$

By the CLT we have that $\boldsymbol{U}_N = N^{1/2}(\boldsymbol{S} - \boldsymbol{\Sigma})$ converges in distribution to normal with zero mean and covariances

$$\mathrm{Cov}(u_{ij}, u_{k\ell}) = \kappa_{1111}^{ijk\ell} + \kappa_{11}^{ik}\kappa_{11}^{j\ell} + \kappa_{11}^{i\ell}\kappa_{11}^{jk}.$$

If $\boldsymbol{X}$ has normal distribution, then $\kappa = 0$ and

$$\mathrm{Cov}(u_{ij}, u_{k\ell}) = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}.$$

Denote by $\boldsymbol{\Gamma}_N$ the corresponding $m^2 \times m^2$ covariance matrix. For elliptical distribution, $N^{1/2}(\boldsymbol{s} - \boldsymbol{\sigma})$ converges in distribution to normal with zero mean and $m^2 \times m^2$ covariance matrix $\boldsymbol{\Gamma}$ with

$$\boldsymbol{\Gamma} = (1 + \kappa)\boldsymbol{\Gamma}_N + \kappa \boldsymbol{\sigma}\boldsymbol{\sigma}'.$$

20 Correlation analysis

20.1 Partial correlation

Let X, Y and Z be random variables. Without loss of generality we can assume that $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z] = 0$. Partial correlation between X and Y given Z is defined as the correlation between residuals of X and Y regressed on Z. That is, let us consider regression X on Z. This is obtained by solving

$$\min_{\beta} \mathbb{E}[(X - \beta Z)^2].$$

Solution of this problem is $\beta = \mathrm{Cov}(X, Z)/\mathrm{Var}(Z)$. Hence the partial correlation is

$$\mathrm{Corr}(X, Y|Z) = \mathrm{Corr}\left(X - \frac{\mathrm{Cov}(X, Z)}{\mathrm{Var}(Z)}Z, Y - \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Var}(Z)}Z\right).$$

Without loss of generality we can assume that $\mathrm{Var}(X) = \mathrm{Var}(Y) = \mathrm{Var}(Z)$. Then

$$\mathrm{Corr}(X, Y|Z) = \mathrm{Corr}\left(X - \rho_{XZ}Z, Y - \rho_{YZ}Z\right) = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}.$$

In similar way partial correlation between random variables X and Y given random variables $Z_1, Z_2, ..., Z_n$, is defined. That is, suppose that $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z_1] = ... = \mathbb{E}[Z_n] = 0$, and $\mathrm{Var}(X) = \mathrm{Var}(Y) = \mathrm{Var}(Z_1) = ... = \mathrm{Var}(Z_n) = 1$. Consider the problem

$$\min_{\boldsymbol{\beta}} \mathbb{E}[(X - \boldsymbol{\beta}'\boldsymbol{Z})^2].$$

Solution of this problem is $\boldsymbol{\beta} = \boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\Sigma}_{ZX}$. Hence

$$\mathrm{Corr}(X, Y|\boldsymbol{Z}) = \mathrm{Corr}(X - \Sigma_{XZ}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{Z}, Y - \Sigma_{YZ}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{Z}).$$

20.2 Canonical correlation analysis

Consider random vectors $\boldsymbol{X} = (X_1, ..., X_p)'$ and $\boldsymbol{Y} = (Y_1, ..., Y_q)'$. Let $\boldsymbol{\mu}_1 = \mathbb{E}[\boldsymbol{X}]$ and $\boldsymbol{\mu}_2 = \mathbb{E}[\boldsymbol{Y}]$, and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$ be the covariance matrix of $(\boldsymbol{X}', \boldsymbol{Y}')'$. Consider random variables $U = \boldsymbol{a}'\boldsymbol{X}$ and $V = \boldsymbol{b}'\boldsymbol{Y}$ for some vectors $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{b} \in \mathbb{R}^q$. We want to solve the problem

$$\max_{\boldsymbol{a},\boldsymbol{b}} \mathrm{Corr}(U, V). \tag{20.1}$$

Suppose for the moment that $\boldsymbol{\Sigma}_{11} = \boldsymbol{I}_p$ and $\boldsymbol{\Sigma}_{22} = \boldsymbol{I}_q$. Then $\mathrm{Cov}(U, V) = \boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}$ and $\mathrm{Var}(U) = \boldsymbol{a}'\boldsymbol{a}$, $\mathrm{Var}(V) = \boldsymbol{b}'\boldsymbol{b}$. Hence problem (20.1) becomes

$$\max_{\boldsymbol{a},\boldsymbol{b}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{b}}}. \tag{20.2}$$

Note that for a given vector $\boldsymbol{w}$, the maximum of $\boldsymbol{w}'\boldsymbol{b}$ subject to $\|\boldsymbol{b}\| = 1$ is attained at $\bar{\boldsymbol{b}} = \boldsymbol{w}/\|\boldsymbol{w}\|$. Therefore for given $\boldsymbol{a}$ the maximum in (20.2) is attained at $\boldsymbol{b} = \boldsymbol{\Sigma}_{21}\boldsymbol{a}$. Hence with respect to $\boldsymbol{a}$ problem (20.2) becomes

$$\max_{\boldsymbol{a}} \left\{ \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}{\sqrt{\boldsymbol{a}'\boldsymbol{a}}\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}} = \sqrt{\frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{a}}} \right\} \tag{20.3}$$

Optimal solution $\bar{\boldsymbol{a}}$ of problem (20.3) is given by the eigenvector of matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}$ corresponding to its largest eigenvalue λ_1, and the maximum in (20.1) is equal to $\sqrt{\lambda_1}$. Similar the optimal $\bar{\boldsymbol{b}}$ is given by the eigenvector of matrix $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}$ corresponding to its largest eigenvalue λ_1. Note that

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\bar{\boldsymbol{a}} = \lambda_1\boldsymbol{\Sigma}_{21}\bar{\boldsymbol{a}},$$

and hence $\bar{\boldsymbol{b}} = \boldsymbol{\Sigma}_{21}\bar{\boldsymbol{a}}$.

In general let $\boldsymbol{c} = \boldsymbol{\Sigma}_{11}^{1/2}\boldsymbol{a}$ and $\boldsymbol{d} = \boldsymbol{\Sigma}_{22}^{1/2}\boldsymbol{b}$. Then

$$\mathrm{Corr}(U, V) = \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{\Sigma}_{22}\boldsymbol{b}}} = \frac{\boldsymbol{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{d}}{\sqrt{\boldsymbol{c}'\boldsymbol{c}}\sqrt{\boldsymbol{d}'\boldsymbol{d}}}.$$

Hence the maximum is attained at $\bar{\boldsymbol{c}}$ given by the eigenvector of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ corresponding to its largest eigenvalue λ_1, and at $\bar{\boldsymbol{d}}$ given by the eigenvector of $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{22}^{-1/2}$ corresponding to its largest eigenvalue λ_1, and

$$\bar{\boldsymbol{d}} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\bar{\boldsymbol{c}}.$$

We have that

$$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\bar{\boldsymbol{c}} = \lambda_1\bar{\boldsymbol{c}}$$

and $\bar{\boldsymbol{c}} = \boldsymbol{\Sigma}_{11}^{1/2}\bar{\boldsymbol{a}}$. Hence

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\bar{\boldsymbol{a}} = \lambda_1\bar{\boldsymbol{a}}, \tag{20.4}$$

and similarly

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\bar{\boldsymbol{b}} = \lambda_1\bar{\boldsymbol{b}}. \tag{20.5}$$

Let $\boldsymbol{a}_1 = \bar{\boldsymbol{a}}$ and $\boldsymbol{b}_1 = \bar{\boldsymbol{b}}$, and $U_1 = \boldsymbol{a}_1'\boldsymbol{X}$ and $V_1 = \boldsymbol{b}_1'\boldsymbol{Y}$. At the second stage we want to find $U_2 = \boldsymbol{a}_2'\boldsymbol{X}$ and $V_2 = \boldsymbol{b}_2'\boldsymbol{Y}$ such that $\mathrm{Cov}(U_2, U_1) = 0$, $\mathrm{Cov}(V_2, V_1) = 0$ and $\mathrm{Corr}(U_2, V_2)$ is maximized. Consider $\boldsymbol{c}_2 = \boldsymbol{\Sigma}_{11}^{1/2}\boldsymbol{a}_2$ and $\boldsymbol{d}_2 = \boldsymbol{\Sigma}_{22}^{1/2}\boldsymbol{b}_2$. Then

$$\mathrm{Cov}(U_2, U_1) = \boldsymbol{a}_2'\boldsymbol{\Sigma}_{11}\boldsymbol{a}_1 = \boldsymbol{c}_2'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{c}_1 = \boldsymbol{c}_2'\boldsymbol{c}_1.$$

Hence $\mathrm{Cov}(U_2, U_1) = 0$ iff $\boldsymbol{c}_2'\boldsymbol{c}_1 = 0$. Therefore the second stage problem is

$$\max_{\boldsymbol{c}} \frac{\boldsymbol{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{d}}{\sqrt{\boldsymbol{c}'\boldsymbol{c}}\sqrt{\boldsymbol{d}'\boldsymbol{d}}} \text{ s.t. } \boldsymbol{c}'\boldsymbol{c}_1 = 0.$$

The maximum is attained at $\bar{\boldsymbol{c}}$ given by the eigenvector of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ corresponding to its second largest eigenvalue λ_2. And so on.

21 Graphical Models

Directed graphical models Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a directed acyclic graph, where $\mathcal{V}$ are the nodes and $\mathcal{E}$ are the edges of the graph. Let $\{X_v : v \in \mathcal{V}\}$ be a collection of random variables indexed by the nodes of the graph. To each node $v \in \mathcal{V}$, let π_v denote the subset of indices of its parents, and X_{π_v} let be the vector of random variables indexed by the parents of $v \in \mathcal{V}$. Let $p(x_v|x_{\pi_v})$, $v \in \mathcal{V}$, be the corresponding conditional densities. Then

$$p(x_{\mathcal{V}}) = \prod_{v \in \mathcal{V}} p(x_v|x_{\pi_v}).$$

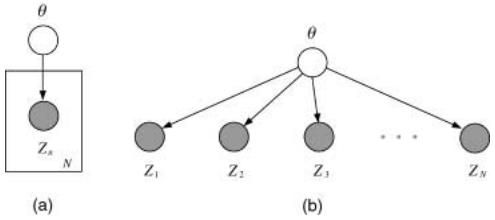

Figure 1: The diagram in (a) is shorthand for the graphical model in (b). This model asserts that the variables Z_n are conditionally independent and identically distributed given θ.

Undirected graphical models Given an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, let $\mathfrak{C}$ denote a collection of cliques of the graph (i.e., fully connected subsets of nodes). Associated with each clique $C \in \mathfrak{C}$, let $\psi_C(x_C)$ denote a nonnegative potential function. Define the joint probability $p(x_\mathcal{V})$ by taking the product over these potential functions and normalizing,

$$p(x_\mathcal{V}) = \frac{1}{F} \prod_{C \in \mathfrak{C}} \psi_C(x_C),$$

where F is a normalization factor obtained by integrating the product with respect to $x_\mathcal{V}$.

Probabilistic inference The problem of probabilistic inference is that of computing conditional probabilities $p(x_A|x_B)$ where A, B, C are disjoint subsets of $\mathcal{V}$ such that $\mathcal{V} = A \cup B \cup C$.

22 Discrete Choice Models

There are I decision makers that can choose between J alternatives. The utility U_{ij} that i-th decision maker obtains from alternative j is decomposed into two parts - a part labeled V_{ij} (systemic component of utility) that is known by the researcher up to some parameters, and an unknown part ε_{ij} that is treated as random. That is

$$U_{ij} = V_{ij} + \varepsilon_{ij}, \;\; i = 1, ..., I, \; j = 1, ..., J. \tag{22.1}$$

It is assumed that i-th decision maker chooses alternative j if $U_{ij} > U_{ik}$ for all $k \neq j$. The probability that decision maker i chooses alternative j is

$$\begin{aligned} p_{ij} &= \mathsf{Prob}\left(V_{ij} + \varepsilon_{ij} > V_{ik} + \varepsilon_{ik}, \forall\, k \neq j\right) \\ &= \mathsf{Prob}\left(\varepsilon_{ik} < \varepsilon_{ij} + V_{ij} - V_{ik}, \forall\, k \neq j\right). \end{aligned} \tag{22.2}$$

Assume that random variables ε_{ij} are independent of each other and have (standard) Gumbel distribution.

> It is said that random variable X has Gumbel (maximum) distribution with location parameter μ and scale parameter $\beta > 0$ if its cdf function has the form
>
> $$F(x) = e^{-e^{-\frac{x-\mu}{\beta}}}, \; -\infty < x < +\infty.$$
>
> Notation $X \sim \text{Gumbel}(\mu, \beta)$. In case $\mu = 0$ and $\beta = 1$ it is called standard Gumbel distribution.

We have that the conditional probability p_{ij} given ε_{ij} is

$$p_{ij}|\varepsilon_{ij} = \prod_{k \neq j} \mathsf{Prob}\left(\varepsilon_{ik} < \varepsilon_{ij} + V_{ij} - V_{ik}\right) = \prod_{k \neq j} e^{-e^{-(\varepsilon_{ij} + V_{ij} - V_{ik})}}. \tag{22.3}$$

Hence

$$p_{ij} = \int_{-\infty}^{+\infty} \prod_{k \neq j} e^{-e^{-(x + V_{ij} - V_{ik})}} dF(x),$$

where $F(x) = e^{-e^{-x}}$ is the cdf of ε_{ij}. Then by change of variables $z = F(x)$,

$$p_{ij} = \int_0^1 z^\alpha dz = (1 + \alpha)^{-1},$$

where $\alpha = \sum_{k \neq j} e^{-(V_{ij}-V_{ik})}$. That is probability that i-th decision maker chooses alternative j is

$$p_{ij} = \frac{1}{1 + \sum_{k \neq j} e^{-(V_{ij}-V_{ik})}} = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}, \tag{22.4}$$

which is the logit probability. Note that $p_{ij} > 0$ and $\sum_{j=1}^{J} p_{ij} = 1$.

Representative utility is usually specified to be linear in parameters $V_{ij} = \boldsymbol{\beta}' \boldsymbol{x}_{ij}$, where $\boldsymbol{x}_{ij}$ is a vector of observed variables relating to alternative j. Then

$$p_{ij} = \frac{e^{\boldsymbol{\beta}' \boldsymbol{x}_{ij}}}{\sum\limits_{k} e^{\boldsymbol{\beta}' \boldsymbol{x}_{ik}}}. \tag{22.5}$$

Estimation, the ML approach Suppose that a sample of N decision makers is obtained. The probability of person i choosing the alternative that he was actually observed to choose is $\prod_j p_{ij}^{y_{ij}}$, where $y_{ij} = 1$ if person i chose j and zero otherwise. Note that since $y_{ij} = 0$ for all nonchosen alternatives, this term is simply the probability of the chosen alternative.

Assuming that each decision maker's choice is independent of that of other decision makers, the probability of each person in the sample choosing the alternative that he was observed actually to choose is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} \prod_{j} p_{ij}^{y_{ij}},$$

where p_{ij} are of the form (22.5) with $\boldsymbol{\beta}$ being vector of the corresponding parameters. The ML approach estimate $\boldsymbol{\beta}$ by maximizing

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{N} \sum_{j} y_{ij} \log p_{ij}.$$

Gumbel distribution If $X_i \sim \text{Gumbel}(\mu_i, \beta)$, $i = 1, ..., m$, are independent, then

$$W = \max_{1 \leq i \leq m} X_i$$

has Gumbel (μ, β) distribution, with

$$\mu = \beta \log \left[\sum_{i=1}^{m} \exp(\mu_i / \beta) \right].$$

In particular, if X_i are standard Gumbel, then $W \sim \text{Gumbel}(\log m, 1)$.

Indeed, we have that the cdf function of W is

$$\begin{aligned} \mathsf{Prob}(W \leq x) &= \mathsf{Prob}(X_i \leq x,\ i = 1, ..., m) = \textstyle\prod_{i=1}^{m} \mathsf{Prob}(X_i \leq x) \\ &= \textstyle\prod_{i=1}^{m} e^{-e^{-\frac{x-\mu_i}{\beta}}} = \exp\left[-e^{-x/\beta} \sum_{i=1}^{m} e^{\mu_i/\beta}\right] = e^{-e^{-\frac{x-\mu}{\beta}}}. \end{aligned}$$

If X has standard Gumbel distribution, then its moment generating function $M_X(t) = \mathbb{E}[e^{tX}]$ is

$$M_X(t) = \Gamma(1 - t),$$

where $\Gamma(z) = \int_0^\infty \tau^{z-1} e^{-\tau} d\tau$, $z > 0$, is the Gamma function. Indeed we have that the pdf of X is $f(x) = e^{-x} e^{-e^{-x}}$, and hence

$$M(t) = \int_{-\infty}^{+\infty} e^{tx} e^{-x} e^{-e^{-x}} dx.$$

By change of variables $\tau = e^{-x}$ we can write this integral as

$$M(t) = \int_0^{+\infty} \tau^{-t} e^{-\tau} d\tau = \Gamma(1-t).$$

If $X \sim \text{Gumbel}(\mu, \beta)$, then $\mathbb{E}[X] = \mu + \gamma\beta$, where

$$\gamma = \lim_{n\to\infty} \left[\sum_{i=1}^{n} \frac{1}{i} - \log n \right] = -\int_0^\infty e^{-\tau} \log \tau d\tau \approx 0.577$$

is Euler's constant, and that $\text{Var}[X] = \pi^2\beta^2/6$. Indeed consider random variable $Y = (X-\mu)/\beta$. It has standard Gumbel distribution. We have that $\mathbb{E}[X] = \mu + \beta\mathbb{E}[Y]$ and

$$\mathbb{E}[Y] = M_Y'(0) = -\Gamma'(1).$$

Recall that $\Gamma'(z) = \int_0^\infty \tau^{z-1} e^{-\tau} \log \tau d\tau$, and hence $\Gamma'(1) = \int_0^\infty e^{-\tau} \log \tau d\tau = \gamma$.

Now $\text{Var}[X] = \beta^2 \text{Var}[Y]$ and

$$\mathbb{E}[Y^2] = M_Y''(0) = \Gamma''(1),$$

and hence

$$\text{Var}[Y] = \Gamma''(1) - [\Gamma'(1)]^2 = \pi^2/6.$$

The last equation follows from (Gauss formula)

$$\frac{d^2 \log \Gamma(z)}{dz^2} = \sum_{i=0}^{\infty} \frac{1}{(z+i)^2},$$

by taking $z = 1$ and noting that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}.$$

If $X_1 \sim \text{Gumbel}(\mu_1, \beta)$ and $X_2 \sim \text{Gumbel}(\mu_2, \beta)$ are independent random variables, then $V = X_1 - X_2$ has logistic distribution with cdf

$$F_V(v) = \frac{1}{1 + \exp[(\mu_2 - \mu_1 - v)/\beta]}.$$

Indeed, since $\frac{X_1 - X_2 + \mu_2 - \mu_1}{\beta} = Y_1 - Y_2$, where $Y_1 \sim \text{Gumbel}(0,1)$ and $Y_2 \sim \text{Gumbel}(0,1)$, it suffices to show this for standard Gumbel variables.

We have

$$\begin{aligned} F_V(v) &= \mathsf{Prob}(V \le v) = \mathsf{Prob}(X_1 - v \le X_2) \\ &= \int_{-\infty}^{+\infty} \mathsf{Prob}(X_1 - v \le x | X_2 = x) e^{-x} e^{-e^{-x}} dx \\ &= \int_{-\infty}^{+\infty} e^{-e^{-(x+v)}} e^{-x} e^{-e^{-x}} dx. \end{aligned}$$

By change of variables $t = e^{-x}$ we obtain

$$F_V(v) = \int_0^{+\infty} e^{-(k+1)t} dt = \tfrac{1}{1+k},$$

where $k = e^{-v}$.