Notes for ISyE 7401

Alexander Shapiro

Contents

1 Some matrix calculus

For an $m \times n$ matrix $\boldsymbol{A}$ we denote by $\boldsymbol{A}'$ its transpose $n \times m$ matrix. Unless stated otherwise vectors $\boldsymbol{a} = (a_1, ..., a_m)'$ are assumed to be column vectors. 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., $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$. Trace of a square $m \times m$ matrix $\boldsymbol{A}$ 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

$$\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA}). \tag{1.1}$$

In particular, if $\boldsymbol{a} = (a_1, ..., a_m)'$ is an $m \times 1$ vector, then $\boldsymbol{aa}'$ is an $m \times m$ matrix, its trace $\mathrm{tr}(\boldsymbol{aa}') = \sum_{i=1}^m a_i^2 = \boldsymbol{a}'\boldsymbol{a}$.

Let $\boldsymbol{A}$ be an $m \times m$ matrix. We denote by $|\boldsymbol{A}|$ the determinant of $\boldsymbol{A}$. Matrix $\boldsymbol{A}$ is nonsingular (invertible) if and only if (iff) $|\boldsymbol{A}| \neq 0$. It is said that λ is an eigenvalue of $\boldsymbol{A}$ if there is an $m \times 1$ vector $\boldsymbol{e} \neq \boldsymbol{0}$ such that $\boldsymbol{Ae} = \lambda \boldsymbol{e}$. It follows that $(\boldsymbol{A} - \lambda \boldsymbol{I}_m)\boldsymbol{e} = \boldsymbol{0}$, where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. Thus matrix $(\boldsymbol{A} - \lambda \boldsymbol{I}_m)$ is singular, and hence its determinant $|\boldsymbol{A} - \lambda \boldsymbol{I}_m| = 0$. Consider[1] $p(\lambda) := |\boldsymbol{A} - \lambda \boldsymbol{I}_m|$. This is a polynomial of degree m and hence has m roots which are eigenvalues of matrix $\boldsymbol{A}$. Therefore matrix $\boldsymbol{A}$ has m eigenvalues some of which can be complex numbers. Suppose now that matrix $\boldsymbol{A}$ is symmetric, i.e., $\boldsymbol{A}' = \boldsymbol{A}$. 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

$$\boldsymbol{Ae}_i = \lambda_i \boldsymbol{e}_i, \ i = 1, ..., m. \tag{1.2}$$

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. In that case we say the eigenvectors are *orthonormal*.

Consider the $m \times m$ 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 $\boldsymbol{TT}' = \boldsymbol{I}_m$. Such matrices are called *orthogonal*. Equations (1.2) can be written in the form $\boldsymbol{AT} = \boldsymbol{T\Lambda}$, where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_m)$ is the diagonal matrix. By multiplying both sides of this matrix equation by $\boldsymbol{T}'$ we obtain

$$\boldsymbol{A} = \boldsymbol{T\Lambda T}' = \sum_{i=1}^m \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i'. \tag{1.3}$$

The representation (1.3) is called *spectral decomposition* of matrix $\boldsymbol{A}$. It also follows that $\boldsymbol{T}'\boldsymbol{AT} = \boldsymbol{\Lambda}$, and that $\mathrm{tr}(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \lambda_1 + \cdots + \lambda_m$, and $\boldsymbol{A}^{-1} = \boldsymbol{T\Lambda}^{-1}\boldsymbol{T}'$, provided that all $\lambda_i \neq 0$, $i = 1, ..., m$.

It is said that matrix $\boldsymbol{A}$ is *positive semidefinite* if $\boldsymbol{x}'\boldsymbol{Ax} \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^m$, and it is said that $\boldsymbol{A}$ is *positive definite* if $\boldsymbol{x}'\boldsymbol{Ax} > 0$ for any $\boldsymbol{x} \neq \boldsymbol{0}$. By using (1.3) we can write

$$\boldsymbol{x}'\boldsymbol{Ax} = \boldsymbol{x}'\boldsymbol{T\Lambda T}'\boldsymbol{x} = \boldsymbol{y}'\boldsymbol{\Lambda y} = \sum_{i=1}^m \lambda_i y_i^2,$$

where $\boldsymbol{y} = \boldsymbol{T}'\boldsymbol{x}$. Note that $\boldsymbol{y}'\boldsymbol{y} = \boldsymbol{x}'\boldsymbol{TT}'\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{x}$. It follows that matrix $\boldsymbol{A}$ is positive semidefinite iff all its eigenvalues are nonnegative, and is positive definite iff all its eigenvalues are positive.

We can define a function of symmetric matrix $\boldsymbol{A}$ by considering a function of its eigenvalues. For example if matrix $\boldsymbol{A}$ is positive semidefinite and hence all its eigenvalues are

[1] Sometimes we write ':=' meaning 'equal by definition'.

nonnegative we can define $\boldsymbol{A}^{1/2} = \boldsymbol{T}\boldsymbol{\Lambda}^{1/2}\boldsymbol{T}'$, where $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\lambda_1^{1/2}, ..., \lambda_m^{1/2})$. The so defined matrix $\boldsymbol{A}^{1/2}$ is symmetric positive semidefinite and $(\boldsymbol{A}^{1/2})^2 = \boldsymbol{T}\boldsymbol{\Lambda}^{1/2}\boldsymbol{T}'\boldsymbol{T}\boldsymbol{\Lambda}^{1/2}\boldsymbol{T}' = \boldsymbol{A}$, since $\boldsymbol{T}'\boldsymbol{T} = \boldsymbol{I}_m$. Similarly if $\boldsymbol{A}$ is positive definite and hence all its eigenvalues are positive, we can define $\boldsymbol{A}^{-1/2} = \boldsymbol{T}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}'$. Matrix $\boldsymbol{A}^{-1/2}$ is symmetric positive definite and $(\boldsymbol{A}^{-1/2})^2 = \boldsymbol{T}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}'\boldsymbol{T}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{T}' = \boldsymbol{A}^{-1}$.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two $m \times m$ symmetric matrices. Then $(\boldsymbol{A}\boldsymbol{B})' = \boldsymbol{B}\boldsymbol{A}$, so the product matrix is not symmetric unless $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$. Suppose that $\boldsymbol{A}$ is positive semidefinite, then matrix $\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}$ is symmetric. Let $\boldsymbol{e}$ be an eigenvector and λ the corresponding eigenvalue of $\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}$, i.e., $\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}\boldsymbol{e} = \lambda\boldsymbol{e}$. Multiplying both sides of this equation by $\boldsymbol{A}^{1/2}$ we obtain $\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^{1/2}\boldsymbol{e} = \lambda\boldsymbol{A}^{1/2}\boldsymbol{e}$. That is, $\boldsymbol{A}^{1/2}\boldsymbol{e}$ is the corresponding eigenvector and λ is the eigenvalue of matrix $\boldsymbol{A}\boldsymbol{B}$. This shows that although $\boldsymbol{A}\boldsymbol{B}$ is not symmetric, it has real valued eigenvectors and eigenvalues. Moreover, if $\boldsymbol{B}$ is positive semidefinite, then $\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}$ is positive semidefinite and hence all eigenvalues of $\boldsymbol{A}\boldsymbol{B}$ are nonnegative, and if both $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive definite matrices, then $\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}$ is positive definite and hence all eigenvalues of $\boldsymbol{A}\boldsymbol{B}$ are positive.

Random vectors. Consider an $m \times 1$ 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. Similarly expectation of a random matrix is taken componentwise. Sometimes we write $\boldsymbol{\mu}_X$ to emphasize that this is mean vector of $\boldsymbol{X}$. The $m \times m$ covariance matrix of $\boldsymbol{X}$ is

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

The (i, j)-component of $\boldsymbol{\Sigma}$ is the covariance $\text{Cov}(X_i, X_j)$, $i, j = 1, ..., m$.

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 $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$. Then

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

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

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

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

$$\boldsymbol{\Sigma}_Y = \boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}'. \tag{1.4}$$

In particular, if $\boldsymbol{A} = \boldsymbol{a}' = (a_1, ..., a_m)'$ is a row vector, then

$$\text{Var}(\boldsymbol{a}'\boldsymbol{X}) = \boldsymbol{a}'\boldsymbol{\Sigma}\boldsymbol{a} = \sum_{i,j=1}^{m} \sigma_{ij}a_ia_j, \tag{1.5}$$

where $\sigma_{ij} = \text{Cov}(X_i, X_j)$ is the (i, j)-component of covariance 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 $\boldsymbol{a}$. Therefore covariance matrix $\boldsymbol{\Sigma}$ is positive semidefinite. If moreover $\boldsymbol{\Sigma}$ is nonsingular (invertible), then it is positive definite.

Recall that 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 $\text{Var}(\boldsymbol{a}'\boldsymbol{X}) = 0$ and hence $Y = \boldsymbol{a}'\boldsymbol{X}$ is constant. In turn this means that random variables $X_1 - \mu_1, ..., X_m - \mu_m$ are linearly dependent. Therefore $\boldsymbol{\Sigma}$ is positive

definite iff variables $X_1 - \mu_1, ..., X_m - \mu_m$ are linearly independent.

As an example let us compute expectation of $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} = \sum_{i,j=1}^m a_{ij}X_iX_j$, where $\boldsymbol{A}$ is an $m \times m$ matrix. Note that using property (1.1) we can write $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X} = \text{tr}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \text{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}')$. Also $\mathbb{E}[\text{tr}(\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}')] = \text{tr}(\mathbb{E}[\boldsymbol{A}\boldsymbol{X}\boldsymbol{X}']$ and hence

$$\mathbb{E}[\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}] = \text{tr}(\boldsymbol{A}\mathbb{E}[\boldsymbol{X}\boldsymbol{X}']) = \text{tr}(\boldsymbol{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')) = \text{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \text{tr}(\boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}') = \text{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\boldsymbol{A}\boldsymbol{\mu}. \quad (1.6)$$

2 Multivariate normal distribution

Recall that a random variable X has normal distribution with mean μ and variance σ^2, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, 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[2] of standard normal variables, i.e., $X_i \sim \mathcal{N}(0, 1)$, $i = 1, ..., m$, 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

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

Recall that $|\boldsymbol{A}|$ denotes determinant of (square) matrix $\boldsymbol{A}$. We used the following properties in the above derivations: $|\boldsymbol{A}^{-1}| = |\boldsymbol{A}|^{-1}$, $\boldsymbol{A}^{'-1}\boldsymbol{A}^{-1} = (\boldsymbol{A}\boldsymbol{A}')^{-1}$, 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

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

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 $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sometimes we write this as $\mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to emphasize dimension m of random vector $\boldsymbol{Y}$. Note that $\boldsymbol{\mu}$ is the $m \times 1$ mean vector and $\boldsymbol{\Sigma}$ is the $m \times m$ covariance matrix of $\boldsymbol{Y}$.

Suppose that $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is partitioned $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}$, where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are subvectors of $\boldsymbol{X}$ of the respective dimensions $m_1 \times 1$ and $m_2 \times 1$, with $m_1 + m_2 = m$. The corresponding partitioning of $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$. Note that $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}'$ since $\boldsymbol{\Sigma}$ is symmetric.

[2] A sequence $X_1, ..., X_m$ of random variables is said to be iid (independent identically distributed), if these random variables are independent of each other and have the same probability distribution.

Suppose further that $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ and hence $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}' = \mathbf{0}$, i.e., matrix $\boldsymbol{\Sigma}$ is block diagonal. Then $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}|\,|\boldsymbol{\Sigma}_{22}|$ and $\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix}$, and hence

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

where $f_X(\cdot)$ is the pdf of $\boldsymbol{X}$ and $f_{X_1}(\cdot)$ and $f_{X_2}(\cdot)$ are pdfs of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. 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 $M(t) := \mathbb{E}[e^{tX}]$. Since $e^0 = 1$, it follows that $M(0) = 1$. Note that it can happen that $M(t) = +\infty$ for any $t \neq 0$. 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 that neighborhood.

Similarly moment generating function of a random vector $\boldsymbol{X} = (X_1, ..., X_m)'$ is defined as

$$M(\boldsymbol{t}) := \mathbb{E}[e^{t_1 X_1 + ... + t_m X_m}] = \mathbb{E}[\exp(\boldsymbol{t}'\boldsymbol{X})].$$

If $M_X(\boldsymbol{t})$ is finite valued in a neighborhood of $\mathbf{0} \in \mathbb{R}^m$, then it is differentiable in that neighborhood. Consider $m \times 1$ vector $\partial M_X(\boldsymbol{t})/\partial \boldsymbol{t} = (\partial M_X(\boldsymbol{t})/\partial t_1, ..., \partial M_X(\boldsymbol{t})/\partial t_m)'$ of first order partial derivatives, and $m \times m$ matrix of second order[3] partial derivatives $\partial^2 M_X(\boldsymbol{t})/\partial \boldsymbol{t}\partial \boldsymbol{t}'$ with (i, j) - element $\partial^2 M_X(\boldsymbol{t})/\partial t_i \partial t_j$, $i, j = 1, ..., m$. Then the expectation and differentiation operations can be interchanged (see Remark 8.1) and

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

Similarly the Hessian matrix

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

Let us compute the moment generating function of $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{X})$. For standard normal random variable $X \sim \mathcal{N}(0, 1)$ 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 $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ and hence components $X_i \sim \mathcal{N}(0, 1)$ of $\boldsymbol{X}$ are independent. Thus

$$M(\boldsymbol{t}) = \mathbb{E}[e^{t_1 X_1 + ... + t_m X_m}] = \mathbb{E}[e^{t_1 X_1} \times \cdots \times e^{t_m X_m}] = \prod_{i=1}^{m} \mathbb{E}[e^{t_i X_i}] = \prod_{i=1}^{m} e^{t_i^2/2} = \exp(\boldsymbol{t}'\boldsymbol{t}/2).$$

Consider now $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$. Since $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m)$ we have that $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}$ and the covariance matrix of $\boldsymbol{Y}$ 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 $\boldsymbol{Y} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ its moment generating function is finite valued for any $m \times 1$ vector $\boldsymbol{t}$, and

$$M_Y(\boldsymbol{t}) = \exp(\boldsymbol{t}'\boldsymbol{\mu} + \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2). \tag{2.2}$$

[3] Matrix of second order partial derivatives is called Hessian matrix.

Now let $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}_X)$ and $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\eta}$, where $\boldsymbol{A}$ is a $k \times m$ matrix and $\boldsymbol{\eta}$ is $k \times 1$ vector. Then

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

That is, the moment generating function of $\boldsymbol{Y}$ is the same as the moment generating function of multivariate normal with mean $\boldsymbol{A}\boldsymbol{X} + \boldsymbol{\eta}$ and covariance matrix $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$. It follows that $\boldsymbol{Y}$ has multivariate normal distribution with mean $\boldsymbol{\mu}_Y = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Sigma}_Y = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$. In particular, marginal distribution of every subvector of $\boldsymbol{X}$ is multivariate normal.

A delicate point of the above result is that the covariance matrix $\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}'$ of $\boldsymbol{Y}$ should be non-singular, i.e. *positive* definite, in order for its density function $f_Y(\boldsymbol{y})$ to be well defined. Since the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$ is positive definite, the matrix $\boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}'$ is nonsingular iff the $k \times m$ matrix $\boldsymbol{A}$ has rank k. For example, if $k > m$, then $\operatorname{rank}(\boldsymbol{A}) \leq m < k$ and hence $\boldsymbol{A}\boldsymbol{\Sigma}_X\boldsymbol{A}'$ is singular.

It follows that random vector $\boldsymbol{X}$ has multivariate normal distribution iff $Y = \boldsymbol{a}'\boldsymbol{X}$ is normally distributed for any vector $\boldsymbol{a} \neq \boldsymbol{0}$. Indeed, if $\boldsymbol{X}$ has normal distribution, then as it was shown above $\boldsymbol{a}'\boldsymbol{X}$ is normally distributed. Conversely, suppose that $\boldsymbol{a}'\boldsymbol{X}$ is normally distributed for any $\boldsymbol{a} \neq \boldsymbol{0}$. Consider $Y := \boldsymbol{t}'\boldsymbol{X}$ for $\boldsymbol{t} \neq \boldsymbol{0}$. We have that $\mu_Y = \boldsymbol{t}'\boldsymbol{\mu}_X$ and $\sigma_Y^2 = \boldsymbol{t}'\boldsymbol{\Sigma}_X\boldsymbol{t}$. Moreover since Y has normal distribution its moment generating function $M_Y(t) = \exp(\mu_Y t + \sigma_Y^2 t^2/2)$. It follows that

$$M_X(\boldsymbol{t}) = \mathbb{E}[\exp(\boldsymbol{t}'\boldsymbol{X})] = M_Y(1) = \exp(\mu_Y + \sigma_Y^2/2) = \exp(\boldsymbol{t}'\boldsymbol{\mu}_X + \boldsymbol{t}'\boldsymbol{\Sigma}_X\boldsymbol{t}/2).$$

That is, the moment generating function of $\boldsymbol{X}$ has the form of normal distribution (see equation (2.2)). It follows that $\boldsymbol{X}$ has normal distribution. □

Conditional normal distribution. Suppose that $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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

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

It follows that $\boldsymbol{Y}$ and $\boldsymbol{X}_2$ are uncorrelated and hence 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

$$\mathcal{N}_{m_1}\big(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\big)\,. \tag{2.3}$$

Note that the conditional covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ is given by the Schur complement of $\boldsymbol{\Sigma}$ (see equation (2.5) below).

Chi-square, t and F distributions. There are three important distributions derived from the normal distribution. Note that if $X \sim \mathcal{N}(0,1)$, then $\mathbb{E}[X^2] = \mathrm{Var}(X) = 1$ and, using integration by parts,

$$\mathbb{E}[X^4] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^4 e^{-x^2/2}dx = \frac{3}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^2 e^{-x^2/2}dx = 3\mathrm{Var}(X) = 3. \tag{2.4}$$

Hence $\mathrm{Var}(X^2) = \mathbb{E}[X^4] - 1 = 2$.

Let $Z_1, ..., Z_m$ be an iid sequence of standard normal random variables. Then $Y := Z_1^2 + ... + Z_m^2$ has chi-square distribution with m degrees of freedom, denoted $Y \sim \chi_m^2$. The expected value of Y is $\mathbb{E}[Y] = \mathbb{E}[Z_1^2] + ... + \mathbb{E}[Z_m^2] = m$ and variance $\mathrm{Var}(Y) = \mathrm{Var}(Z_1^2) + ... + \mathrm{Var}(Z_m^2) = 2m$. By the Law of Large Numbers, Y/m tends in probability to 1, and by the Central Limit Theorem, $m^{-1/2}(Y - m)$ tends in distribution to normal $N(0, 2)$, as $m \to \infty$.

The t distribution with m degrees of freedom is defined as distribution of $T = \frac{Z}{\sqrt{W/m}}$, where $Z \sim \mathcal{N}(0,1)$ and $W \sim \chi_m^2$ are independent random variables, denoted $T \sim t_m$. Since for large m, W/m becomes close to one, critical values of t-statistic are close to the respective standard normal critical values when the degrees of freedom are large.

The F distribution with k and m degrees of freedom is defined as distribution of $F = \frac{V/k}{W/m}$, where $V \sim \chi_k^2$ and $W \sim \chi_m^2$ are independent random variables, denoted $F \sim F_{k,m}$. It follows from the above definitions that if $T \sim t_m$, then $T^2 \sim F_{1,m}$.

2.1 Schur complement

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

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

where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are matrices of respective dimensions $n \times n$, $n \times m$, $m \times n$, $m \times m$. Suppose that $\boldsymbol{D}$ is invertible (nonsingular). Then

$$\begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ -\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{I}_m \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix}. \tag{2.5}$$

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

Note that

$$\begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_m \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_m \end{bmatrix}$$

and

$$\begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ -\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{I}_m \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I}_n & 0 \\ \boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{I}_m \end{bmatrix},$$

and determinants of these matrices equal one. Hence it follows from (2.5) that

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_m \end{bmatrix} \begin{bmatrix} \boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{I}_m \end{bmatrix}.$$

This implies the following formula for the determinant of matrix $\boldsymbol{M}$,

$$|\boldsymbol{M}| = |\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}|\,|\boldsymbol{D}|. \tag{2.6}$$

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

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ -\boldsymbol{D}^{-1}\boldsymbol{C} & \boldsymbol{I}_m \end{bmatrix} \begin{bmatrix} (\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{B}\boldsymbol{D}^{-1} \\ \boldsymbol{0} & \boldsymbol{I}_m \end{bmatrix}.$$

Using the above equation it is possible to compute

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & -(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & \boldsymbol{D}^{-1}+\boldsymbol{D}^{-1}\boldsymbol{C}(\boldsymbol{A}-\boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix}. \tag{2.7}$$

3 Quadratic forms

In this section we discuss distribution of quadratic forms $Q = \boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$, where $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{A}$ is an $m \times m$ symmetric (deterministic) matrix. Recall that the expected value of $\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}$ was computed in equation (1.6). Let us first consider simple case where $\boldsymbol{A} = \boldsymbol{I}_m$ and $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$. Then $Q = X_1^2 + ... + X_m^2$ has chi-square distribution with m degrees of freedom, $Q \sim \chi^2_m$.

Theorem 3.1 *Let $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{\Sigma})$. 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 $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m)$. 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 $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} = Y_1^2 + ... + Y_m^2 \sim \chi^2_m$. □

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. Indeed if λ is an eigenvalue of $\boldsymbol{P}$ and $\boldsymbol{e}$ the corresponding eigenvector, then $\boldsymbol{P}^2\boldsymbol{e} = \boldsymbol{P}(\boldsymbol{P}\boldsymbol{e}) = \lambda^2\boldsymbol{e}$. On the other hand since $\boldsymbol{P}^2 = \boldsymbol{P}$, $\boldsymbol{P}^2\boldsymbol{e} = \lambda\boldsymbol{e}$. It follows that $\lambda^2 = \lambda$, and hence $\lambda = 1$ or $\lambda = 0$.

Moreover, suppose that $\boldsymbol{P}$ is symmetric. Then for any $\boldsymbol{x} \in \mathbb{R}^m$,

$$(\boldsymbol{x}-\boldsymbol{P}\boldsymbol{x})'\boldsymbol{P}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{P}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{P}\boldsymbol{x} - \boldsymbol{x}'\boldsymbol{P}^2\boldsymbol{x} = 0.$$

That is, $\boldsymbol{P}$ makes orthogonal projection of vector $\boldsymbol{x}$ onto the linear space $\{\boldsymbol{y} : \boldsymbol{y} = \boldsymbol{P}\boldsymbol{x},\ \boldsymbol{x} \in \mathbb{R}^m\}$.

Also by the spectral decomposition, $\boldsymbol{P} = \boldsymbol{T}_1\boldsymbol{T}_1'$, where $\boldsymbol{T}_1$ is the $m \times r$ matrix whose columns are orthonormal eigenvectors corresponding to eigenvalues 1, i.e., $\boldsymbol{T}_1'\boldsymbol{T}_1 = \boldsymbol{I}_r$. Then $\operatorname{rank}(\boldsymbol{P}) = r = \operatorname{tr}(\boldsymbol{P})$.

Theorem 3.2 *Let $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{0}, \boldsymbol{I}_m)$ 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 $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{I}_m)$ 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 Theorems 3.1 and 3.2 it is possible to show the following.

Theorem 3.3 *If $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} \sim \chi^2_m(\delta)$ with the noncentrality parameter $\delta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$. If $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{I}_m)$ 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

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

Denote by $\boldsymbol{Y} = (Y_1, ..., Y_N)'$ vector of responses, $\boldsymbol{X}_j = (x_{1j}, ..., x_{Nj})'$, $j = 1, ..., k$, predictors (regressors), $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_N)'$ vector of errors and $\boldsymbol{1}_N = (1, ..., 1)'$ vector of ones. Then we can write model (4.1) as $\boldsymbol{Y} = \beta_0\boldsymbol{1}_N + \beta_1\boldsymbol{X}_1 + ... + \beta_k\boldsymbol{X}_k + \boldsymbol{\varepsilon}$, or equivalently in matrix form as

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ is vector of parameters and $\boldsymbol{X} = [\boldsymbol{1}_N, \boldsymbol{X}_1, ..., \boldsymbol{X}_k]$ is $N \times p$, $p = k + 1$, so called design matrix. Note that the first column of $\boldsymbol{X}$ is formed by ones. Unless stated otherwise, it will be assumed that $\boldsymbol{X}$ has full column rank p, i.e., column vectors $\boldsymbol{1}_N, \boldsymbol{X}_1, ..., \boldsymbol{X}_k$ of the design matrix are linearly independent.

Note that the design matrix $\boldsymbol{X}$ is assumed to be *deterministic*. This is justified when values x_{ij} of the predictors (regressors) are observed without error. If x_{ij} are modelled as random, the analysis below can be pushed through by conditional arguments.

The Least Squares Estimator (LSE) $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is solution of the problem

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}). \tag{4.3}$$

By Pythagoras Theorem vector of residuals $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is orthogonal to the linear space generated by columns of the design matrix $\boldsymbol{X}$. That is $\boldsymbol{e}'\boldsymbol{X} = \boldsymbol{0}$ or equivalently $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. It follows that $(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$. Since it is assumed that matrix $\boldsymbol{X}$ has full column rank, the $p \times p$ matrix $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular (invertible). Thus the LSE can be written as $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$.

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 $\boldsymbol{e} = (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{Y}$. Matrix $\boldsymbol{H}$

is the orthogonal projection matrix onto the space generated by columns of $\boldsymbol{X}$, i.e., $\boldsymbol{HY} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ and $\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{HY}) = \boldsymbol{0}$; and matrix $\boldsymbol{I}_N - \boldsymbol{H}$ is the orthogonal projection matrix onto the space orthogonal to the space generated by columns of $\boldsymbol{X}$.

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) $\boldsymbol{HX} = \boldsymbol{X}$ 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

$$\mathrm{Cov}(\boldsymbol{e}) = \mathrm{Cov}[(\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{Y}] = \sigma^2(\boldsymbol{I}_N - \boldsymbol{H})^2 = \sigma^2(\boldsymbol{I}_N - \boldsymbol{H}).$$

Moreover $\mathbb{E}[\boldsymbol{e}] = (\boldsymbol{I}_N - \boldsymbol{H})\mathbb{E}[\boldsymbol{Y}] = (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0}$ and hence

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

That is,

$$S^2 := \tfrac{1}{N-p} \textstyle\sum_{i=1}^{N} e_i^2$$

is an unbiased estimator of σ^2.

Since the first column of $\boldsymbol{X}$ is vector $\boldsymbol{1}_N = (1, ..., 1)'$ of ones and $\boldsymbol{e}'\boldsymbol{X} = \boldsymbol{0}$, it follows that $\boldsymbol{e}'\boldsymbol{1}_N = 0$, that is $\sum_{i=1}^{N} e_i = 0$. In a similar way we have $\boldsymbol{e}'\hat{\boldsymbol{Y}} = \boldsymbol{Y}'(\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{H}\boldsymbol{Y} = 0$. That is, residuals $\boldsymbol{e}$ and fitted values $\hat{\boldsymbol{Y}}$ are uncorrelated.

Consider $\bar{Y} = N^{-1}\sum_{i=1}^{N} Y_i$. Since $\sum_{i=1}^{N} e_i = 0$ we have that $\bar{Y} = N^{-1}\sum_{i=1}^{N} \hat{Y}_i$ as well. Note that

$$\sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})e_i = \boldsymbol{e}'\hat{\boldsymbol{Y}} - \bar{Y}\sum_{i=1}^{N} e_i = 0,$$

and hence

$$\sum_{i=1}^{N} (Y_i - \bar{Y})^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2.$$

That is

$$S_{YY} = SS_R + SS_E, \tag{4.4}$$

where

$$S_{YY} := \sum_{i=1}^{N} (Y_i - \bar{Y})^2,\ SS_R := \sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2,\ SS_E := \sum_{i=1}^{N} e_i^2. \tag{4.5}$$

The so-called coefficient of determination is defined as

$$R^2 := \frac{S_{YY}}{SS_R} = 1 - \frac{SS_E}{SS_R}.$$

It is interpreted as proportion of the total variation S_{YY} (corrected for the average) explained by variation SS_R due to regression. Another interpretation is that $R^2 = r^2$, where r is the sample correlation coefficient between Y_i and $\hat{Y}_i$. Indeed

$$\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})(Y_i - \bar{Y}) = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}) = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2,$$

and hence

$$r^2 = \left[\frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}}\right]^2 = \frac{\left[\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2\right]^2}{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}.$$

In case of one predictor, i.e., $Y_i = \beta_0 + \beta X_i + \varepsilon_i$, $i = 1, ..., N$, the sample correlation between Y_i and $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta} X_i$ is the same as the sample correlation between Y_i and X_i, $i = 1, ..., N$.

Theorem 4.1 (Gauss - Markov) *Suppose that* $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$ *and* $\mathrm{Cov}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}_N$. *Then the LSE* $\hat{\boldsymbol{\beta}}$ *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*

$$\mathrm{Var}(\boldsymbol{a}'\tilde{\boldsymbol{\beta}}) \geq \mathrm{Var}(\boldsymbol{a}'\hat{\boldsymbol{\beta}}) \tag{4.6}$$

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 $(\boldsymbol{I}_p - \boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta} = \mathbf{0}$ 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

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

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

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 $\boldsymbol{X}'\boldsymbol{B} = \mathbf{0}$ 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 $\boldsymbol{a}'\boldsymbol{B}'\boldsymbol{B}\boldsymbol{a} = (\boldsymbol{B}\boldsymbol{a})'\boldsymbol{B}\boldsymbol{a} \geq 0$. □

The LSE $\hat{\boldsymbol{\beta}}$ is the solution of the system of linear equations $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$. It can happen that small changes in values of the design matrix $\boldsymbol{X}$ result in big changes in the solution of that system of equations. In numerical analysis such problems are called ill-conditioned. In regression this is called multicollinearity problem, when columns of the design matrix are 'almost' linearly

dependent. Ill conditioning of a system of linear equations is measured by the so-called condition number (see below).

In regression the multicollinearity problem is measured by the so-called Variance Inflation Factor, VIF_i, which is a measure of collinearity of regressor (predictor) $\boldsymbol{X}_i$ with the other regressors, $i = 1, ..., k$. It is defined as $VIF_i := 1/(1 - R_i^2)$, where R_i^2 is the coefficient of determination of regression $\boldsymbol{X}_i$ on the other regressors. Let $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{X}}_1, ..., \tilde{\boldsymbol{X}}_k]$ be the $N \times k$ normalized design matrix, i.e., the averages are removed from each regressor so the sum of elements of each regressor $\tilde{\boldsymbol{X}}_i$ is zero, and all diagonal elements of $k \times k$ matrix $\boldsymbol{R} := \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}$ are equal to one, i.e., the sum of squared elements of each regressor $\tilde{\boldsymbol{X}}_i$ is equal to one.

Consider the following partitioning $\begin{bmatrix} 1 & \boldsymbol{r}_{12} \\ \boldsymbol{r}_{21} & \boldsymbol{R}_{11} \end{bmatrix}$ of matrix $\boldsymbol{R}$. Then by equation (2.7) we have that the first diagonal element of matrix $\boldsymbol{R}^{-1}$ is equal to $(1 - \boldsymbol{r}_{12}\boldsymbol{R}_{11}^{-1}\boldsymbol{r}_{21})^{-1}$. Consider now regression of the first regressor $\tilde{\boldsymbol{X}}_1$ on the other regressors $\tilde{\boldsymbol{X}}_2, ..., \tilde{\boldsymbol{X}}_k$. The corresponding coefficient of determination R_1^2 is equal to $\boldsymbol{r}_{12}\boldsymbol{R}_{11}^{-1}\boldsymbol{r}_{21}$. This can be applied to regression of every $\tilde{\boldsymbol{X}}_i$ on the other regressors. Therefore Variance Inflation Factors can be obtained as diagonal elements of the matrix $\boldsymbol{R}^{-1}$.

Condition number.
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 an $n \times 1$ 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 $\boldsymbol{b}$

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

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

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

Now let $\sigma_{\min}(\boldsymbol{A}) = \sqrt{\lambda_{\min}(\boldsymbol{A}'\boldsymbol{A})}$ and $\sigma_{\max}(\boldsymbol{A}) = \sqrt{\lambda_{\max}(\boldsymbol{A}'\boldsymbol{A})}$ be the minimal and maximal singular values of $\boldsymbol{A}$ (see section 15.4). Then

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

and

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

Therefore $\operatorname{cond}(\boldsymbol{A}) = \sigma_{\max}(\boldsymbol{A})/\sigma_{\min}(\boldsymbol{A})$.

If matrix $\boldsymbol{A}$ is symmetric positive definite, then $\sigma_{\max}(\boldsymbol{A})$ and $\sigma_{\min}(\boldsymbol{A})$ are the largest and smallest eigenvalues of $\boldsymbol{A}$, respectively. For matrix $\boldsymbol{A} = \gamma\boldsymbol{I}_n$, $\gamma \neq 0$, its condition number $\operatorname{cond}(\gamma\boldsymbol{I}_n) = 1$. Otherwise the condition number is bigger than one.

4.1 Distribution theory

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

$$\sigma^{-2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi_p^2, \tag{4.7}$$

where $p = k + 1$. Recall that $S^2 = (N-p)^{-1}\boldsymbol{e}'\boldsymbol{e}$ is an unbiased estimator of σ^2.

Since $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and $(\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}$,

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

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

$$\begin{aligned}\operatorname{rank}(\boldsymbol{I}_N - \boldsymbol{H}) &= \operatorname{tr}(\boldsymbol{I}_N - \boldsymbol{H}) = \operatorname{tr}(\boldsymbol{I}_N) - \operatorname{tr}(\boldsymbol{H}) = N - \operatorname{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \\ &= N - \operatorname{tr}((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = N - p.\end{aligned}$$

By Theorem 3.2 it follows that

$$\frac{(N-p)S^2}{\sigma^2} \sim \chi^2_{N-p}. \tag{4.8}$$

Moreover

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

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

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/p}{S^2} \sim F_{p,N-p}. \tag{4.9}$$

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

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

Now consider S_{YY}, SS_R and SS_E and recall that $S_{YY} = SS_R + SS_E$ (see equation (4.4)). Also

$$(N-p)S^2 = SS_E = \boldsymbol{Y}'(\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{Y} \tag{4.10}$$

and

$$SS_R = (\hat{\boldsymbol{Y}} - \boldsymbol{1}_N\bar{Y})'(\hat{\boldsymbol{Y}} - \boldsymbol{1}_N\bar{Y}) = (\boldsymbol{H}\boldsymbol{Y} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N'\boldsymbol{Y})'(\boldsymbol{H}\boldsymbol{Y} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N'\boldsymbol{Y}) = \boldsymbol{Y}'(\boldsymbol{H} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N')^2\boldsymbol{Y}.$$

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

$$SS_R = \boldsymbol{Y}'(\boldsymbol{H} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N')\boldsymbol{Y}. \tag{4.11}$$

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

$$(\boldsymbol{I}_N - \boldsymbol{H})(\boldsymbol{H} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N') = (\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{H} - N^{-1}(\boldsymbol{I}_N - \boldsymbol{H})\boldsymbol{1}_N\boldsymbol{1}_N' = \boldsymbol{0},$$

and hence SS_E and SS_R are independent.

Consider the following so-called F-statistic, for testing $H_0 : \beta_1 = ... = \beta_k = 0$, against the alternative that at least one $\beta_i \neq 0$,

$$F = \frac{SS_R/k}{SS_E/(N-p)}. \tag{4.12}$$

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

$$(\boldsymbol{H} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N')\boldsymbol{Y} = \beta_0(\boldsymbol{H} - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N')\boldsymbol{1}_N = \beta_0(\boldsymbol{H}\boldsymbol{1}_N - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N'\boldsymbol{1}_N) = \boldsymbol{1}_N - \boldsymbol{1}_N = \boldsymbol{0}.$$

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

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

Therefore under the alternative, the F-statistic has noncentral F distribution with the noncentrality parameter δ for SS_R.

4.2 Estimation with linear constraints

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 full row rank q.

The respective constrained least squares estimator $\hat{\boldsymbol{\beta}}_H$ is obtained as a solution of the following optimization problem

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \text{ subject to } \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}. \tag{4.13}$$

Consider the Lagrangian of the above problem (4.13):

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

Problem (4.13) is a convex quadratic problem. Optimality conditions for problem (4.13) 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

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{A}' \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{Y} \\ \boldsymbol{c} \end{bmatrix}. \tag{4.14}$$

Note that the Schur complement of matrix $\begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{A}' \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix}$ with respect to matrix $\boldsymbol{X}'\boldsymbol{X}$ is $-\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$. Since it is assumed that matrix $\boldsymbol{X}$ has full column rank and hence matrix $\boldsymbol{X}'\boldsymbol{X}$ is nonsingular, and since matrix $\boldsymbol{A}$ has full row rank q, it follows that $-\boldsymbol{A}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{A}'$ is nonsingular, and hence matrix $\begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{A}' \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix}$ is invertible. Therefore the corresponding estimators are given by

$$\begin{bmatrix} \hat{\boldsymbol{\beta}}_H \\ \hat{\boldsymbol{\lambda}}_H \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{A}' \\ \boldsymbol{A} & \mathbf{0} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}'\boldsymbol{Y} \\ \boldsymbol{c} \end{bmatrix}. \tag{4.15}$$

By using formula (2.7) 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 estimator $\hat{\boldsymbol{\beta}}_H$ in the following form

$$\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}} + (\boldsymbol{X'X})^{-1}\boldsymbol{A'}[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\beta}}), \tag{4.16}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$ is the unconstrained LSE. It is possible to give the following geometrical interpretation. Consider $\hat{\boldsymbol{Y}}_H = \boldsymbol{X}\hat{\boldsymbol{\beta}}_H$. Recall that $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is orthogonal to the linear space generated by columns of matrix $\boldsymbol{X}$. Since $\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H = \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H)$, it follows that $\boldsymbol{Y} - \hat{\boldsymbol{Y}}$ is orthogonal to $\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H$. Hence (Pythagoras Theorem)

$$\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_H\|^2 = \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2 + \|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H\|^2, \tag{4.17}$$

where $\|\cdot\|$ is the Euclidean norm. Moreover

$$\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{c} - \boldsymbol{A}\hat{\boldsymbol{\beta}}). \tag{4.18}$$

The term $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}_H\|^2$ represents the sum of squares of residuals of the reduced (constrained) model, i.e., it is the optimal value of the least squares problem (4.13), and the term $\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2$ is the sum of squares of residuals of the full (unconstrained) model. By (4.18),

$$\|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_H\|^2 = (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c}). \tag{4.19}$$

The F-statistic for testing $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ is

$$F = \frac{(SS_E(H) - SS_E(F))/q}{SS_E(F)/(N-p)}, \tag{4.20}$$

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. By (4.17) and (4.19) we have

$$SS_E(H) - SS_E(F) = (\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'[\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}]^{-1}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c}). \tag{4.21}$$

Recall that $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1})$, and hence

$$\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c} \sim \mathcal{N}\big(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{c}, \sigma^2\boldsymbol{A}(\boldsymbol{X'X})^{-1}\boldsymbol{A'}\big).$$

It follows by Theorem 3.1 that under the $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$ hypothesis, $[SS_E(H) - SS_E(F)]/\sigma^2 \sim \chi^2_q$. Also $SS_E(F)/\sigma^2 \sim \chi^2_{N-p}$ and $SS_E(F)$ is independent of $\hat{\boldsymbol{\beta}}$, and hence $SS_E(H) - SS_E(F)$ and $SS_E(F)$ are independent. It follows that under the H_0 hypothesis, the F statistic (4.20) has $F_{q,N-p}$ distribution.

The F-statistic (4.12), for testing $H_0 : \beta_1 = ... = \beta_k = 0$, is a particular case of the F-statistic (4.20). Indeed in that case, under H_0, the LSE $\hat{\beta}_0 = \bar{Y}$ and hence $SS_E(H) = S_{YY}$. It follows that $SS_E(H) - SS_E(F) = SS_R$.

The statistical inference discussed in section 4.1 and this section is based on the assumption that the error vector $\boldsymbol{\varepsilon}$ has normal distribution. Without this assumption the inference is asymptotic. Recall that for large N, the $qF_{q,N-p}$ distribution becomes approximately like χ^2_q distribution.

4.3 Polynomial Regression

Consider the polynomial regression model (one predictor)

$$Y_i = \beta_0 + \beta_1 x_i + ...\beta_k x_i^k + \varepsilon_i,\ i = 1, ..., N. \tag{4.22}$$

We can formulate this as the linear multivariate model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with the 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 $[\boldsymbol{X}'\boldsymbol{X}]_{st} = \sum_{i=1}^N x_i^{s+t}$.

Note that (Riemann sum)

$$\int_0^1 x^{s+t} dx \approx \frac{1}{N} \sum_{i=1}^N x_i^{s+t},$$

where x_i is a point of the interval $[(i-1)/N, i/N]$, $i = 1, ..., N$. Therefore

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

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. Therefore polynomial regression of the form (4.22) typically has multicollinearity problem for $k \geq 3$. To a certain extend this can be dealt with by using orthogonal polynomials. A famous example of orthogonal polynomials is Chebishev polynomials. Even so, polynomial regression of degree larger than 2 usually is difficult to interpret.

Chebishev polynomials

$$T_m(x) = \cos[m(\arccos x)],\ -1 \leq x \leq 1.$$

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

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

can be used for recursive computation of Chebishev polynomials. For example

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

By using substitution $d \arccos x = \frac{1}{\sqrt{1-x^2}} dx$, we can compute the following integral

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

For $x_i = \cos(\pi/N)i$ and $\theta_i = (\pi/N)i$ we have

$$\sum_{i=0}^{N-1} T_k(x_i)T_\ell(x_i) = \sum_{i=0}^{N-1} \cos k\theta_i \cos \ell\theta_i = 0, \ k \neq \ell. \tag{4.23}$$

For the corresponding polynomial regression

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

the design matrix is

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

By (4.23) columns of this design matrix are orthogonal to each other, and hence matrix $\boldsymbol{X}'\boldsymbol{X}$ is diagonal.

5 Shrinkage Methods

A norm $\|\cdot\|$, on space $\mathbb{R}^m$, assigns a nonnegative number to vector $\boldsymbol{x} \in \mathbb{R}^m$. It should have the following properties: (i) $\|\boldsymbol{x}\| > 0$ for any $\boldsymbol{x} \neq \boldsymbol{0}$, (ii) $\|\lambda \boldsymbol{x}\| = |\lambda| \, \|\boldsymbol{x}\|$ for any $\lambda \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^m$, (iii) $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$. Properties (ii) and (iii) imply that function $f(\boldsymbol{x}) = \|\boldsymbol{x}\|$ is convex. Any two norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{R}^m$ are equivalent in the following sense: there is a constant $C > 0$ (depending on dimension m of the space $\mathbb{R}^m$) such that $\|\boldsymbol{x}\| \leq C\|\boldsymbol{x}\|'$ and $\|\boldsymbol{x}\|' \leq C\|\boldsymbol{x}\|$ for all $\boldsymbol{x} \in \mathbb{R}^m$.

Important examples of norms are the ℓ_q, $q \geq 1$, norms defined as $\|\boldsymbol{x}\|_q = (|x_1|^q + ... + |x_m|^q)^{1/q}$. In particular, the ℓ_2 norm is the Euclidean norm $\|\boldsymbol{x}\|_2 = \sqrt{x_1^2 + ... + x_m^2}$, and ℓ_1 norm is $\|\boldsymbol{x}\|_1 = |x_1| + ... + |x_m|$. Note that function $\|\cdot\|_q$ is homogeneous (i.e., satisfies the above property (ii)) for any $q > 0$. However for $q \in (0, 1)$, $\|\cdot\|_q$ does not satisfy property (iii), i.e., it 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 the LSE $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. It is also possible to formulate problem (5.1) in the following form

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \text{ subject to } \|\boldsymbol{\beta}\|_2 \leq c, \tag{5.2}$$

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 $\|\hat{\boldsymbol{\beta}}\|_2 \leq c$, 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 LSE to the origin. In particular if columns of the design matrix $\boldsymbol{X}$ are orthogonal, i.e., matrix $\boldsymbol{X}'\boldsymbol{X} = \text{diag}(\lambda_1, ..., \lambda_p)$ is diagonal. Then

$$\boldsymbol{X}'\boldsymbol{X} + \varepsilon \boldsymbol{I}_p = \text{diag}(\lambda_1 + \varepsilon, ..., \lambda_p + \varepsilon)$$

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 $\lambda_1 \geq \cdots \geq \lambda_p > 0$ being the eigenvalues and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$. Then $\boldsymbol{X}'\boldsymbol{X} + \varepsilon \boldsymbol{I}_p = \boldsymbol{T}(\boldsymbol{\Lambda} + \varepsilon \boldsymbol{I}_p)\boldsymbol{T}'$.

Recall that 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 $\varepsilon > 0$ if the ratio ε/λ_p is large. 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 LSE of the corresponding linear model with $\boldsymbol{X}$ replaces by $\tilde{\boldsymbol{X}}$. If ε is much larger than λ_i, and hence the ratio ε/λ_i is large, then $\tilde{\gamma}_{\varepsilon,i}$ becomes small. In that sense this procedure removes from the design matrix $\tilde{\boldsymbol{X}}$ columns corresponding to small values of the eigenvalues λ_i, and in an implicit way is related to the Principal Components Analysis discussed in section 15.

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 (MSE) than the respective components of the LSE $\hat{\boldsymbol{\beta}}$. That is, let $\theta = \boldsymbol{a}'\boldsymbol{\beta}$ for some given vector $\boldsymbol{a} \neq \boldsymbol{0}$, and let $\tilde{\theta}_\varepsilon = \boldsymbol{a}'\tilde{\boldsymbol{\beta}}_\varepsilon$ and $\hat{\theta} = \boldsymbol{a}'\hat{\boldsymbol{\beta}}$ be estimators of θ. Note that $\boldsymbol{a}'\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{a}'\boldsymbol{\beta}$. 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

$$\mathbb{E}[\tilde{\theta}_\varepsilon] = \boldsymbol{a}'[\boldsymbol{I}_p + \varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}]^{-1}\boldsymbol{\beta}.$$

For a matrix $\boldsymbol{A}$ sufficiently small we have the following geometric series expansion

$$(\boldsymbol{I} + \boldsymbol{A})^{-1} = \boldsymbol{I} - \boldsymbol{A} + \boldsymbol{A}^2 - ... = \boldsymbol{I} - \boldsymbol{A} + o(\|\boldsymbol{A}\|),$$

where $\|\boldsymbol{A}\| := \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \|\boldsymbol{A}\boldsymbol{x}\|/\|\boldsymbol{x}\|$. By applying this to matrix $\boldsymbol{A} = \varepsilon(\boldsymbol{X}'\boldsymbol{X})^{-1}$ for $\varepsilon > 0$ small enough, we obtain

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

where $o(\varepsilon)/\varepsilon \to 0$ as $\varepsilon \to 0$. It follows that

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

and hence

$$\text{Bias}^2[\tilde{\theta}_\varepsilon] = \varepsilon^2[\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{\beta}]^2 + o(\varepsilon^2) = o(\varepsilon).$$

We also have that

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

Therefore

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

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

In particular this implies that for every $i \in \{1, ..., k\}$, there exists $\varepsilon > 0$ such that $MSE(\tilde{\boldsymbol{\beta}}_{\varepsilon,i}) < MSE(\hat{\beta}_i)$. However for different $i \in \{1, ..., k\}$ the corresponding ε can be different, and could be difficult to find. In practical applications the components of $\tilde{\boldsymbol{\beta}}_\varepsilon$ are plotted as a function of $\varepsilon > 0$ until they stabilize.

5.2 Lasso method

The Least Absolute Shrinkage and Selection Operator (Lasso) method is based on using regularization term of the form $\varepsilon\|\boldsymbol{\beta}\|_1$ for some $\varepsilon > 0$. 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

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

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

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 $S = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq c\}$ and the corresponding optimality conditions are

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

where $N_S(\boldsymbol{\beta}) := \{\boldsymbol{\gamma} : \boldsymbol{\gamma}'(\boldsymbol{\zeta} - \boldsymbol{\beta}) \leq 0,\ \forall \boldsymbol{\zeta} \in S\}$ is the normal cone to S at $\boldsymbol{\beta} \in S$.

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 $[\boldsymbol{X'Y}]_i$ of vector $\boldsymbol{X'Y}$, then $\tilde{\boldsymbol{\beta}}_\varepsilon = \mathbf{0}$. If $\boldsymbol{X'X}$ is diagonal, then $\tilde{\beta}_{\varepsilon,i} = 0$, when $\varepsilon > [\boldsymbol{X'Y}]_i$.

It is possible to look at Lasso estimation from 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

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

i.e., it is the least squares problem subject to the constraint that the number of used regressors is not larger than c. This is a difficult combinatorial problem. Problem (5.4) can be viewed as a convex approximation of problem (5.5). Problem (5.3) can be formulated as the following problem

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

and problem (5.4) as

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta},\boldsymbol{\xi}} & \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} & \beta_i \leq \xi_i, \ \ -\beta_i \leq \xi_i, \ i = 0, ..., k, \\ & \sum_{i=0}^k \xi_i \leq c. \end{array} \tag{5.7}$$

Both problems (5.6) and (5.7) are convex quadratic programming problems, and can be solved efficiently.

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

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

Convergence in probability can be also considered for a sequence $\boldsymbol{Y}_n \in \mathbb{R}^m$, $n = 1, ...$, of random vectors. That is, $\boldsymbol{Y}_n$ converges in probability to $\boldsymbol{a}$ if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathsf{Prob}\{\|\boldsymbol{Y}_n - \boldsymbol{a}\| \geq \varepsilon\} = 0.$$

It is straightforward to show that $\boldsymbol{Y}_n$ converges in probability to $\boldsymbol{a}$ iff its every component Y_{in} converges in probability to a_i, $i = 1, ..., m$.

Law of Large Numbers (LLN) can be proved by using Chebishev inequality. Let X be a nonnegative valued random variable. Then for any $\varepsilon > 0$ we have

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

where $\mathbf{1}_{[\varepsilon,\infty)}(x) = 0$ if $x < \varepsilon$ and $\mathbf{1}_{[\varepsilon,\infty)}(x) = 1$ if $x \geq \varepsilon$. The above inequality sometimes is called Markov inequality. Now let X be a random variable with finite second order moment, i.e., $\mathbb{E}[X^2] < \infty$. By taking $Y = (X - \mu)^2$, where $\mu = \mathbb{E}[X]$, we obtain from Markov inequality the following Chebishev inequality:

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

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 then for any $\varepsilon > 0$,

$$\mathsf{Prob}\{|Y_n - \mu| \geq \varepsilon\} \leq \varepsilon^{-2}\mathrm{Var}(Y_n) \to 0.$$

This implies that $Y_n \xrightarrow{p} \mu$. 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 $n \to \infty$.

> The convergence of $\bar{X}$ to μ in probability is referred to as the (weak) Law of Large Numbers (WLLN). The stronger version of LLN is that $\bar{X}$ converges to μ with probability one (w.p.1), provided the mean μ is well defined and finite. Note that convergence w.p.1 implies convergence in probability.

In Calculus the notation $y_n = o(x_n)$ is used to denote that if x_n and y_n are sequences of (deterministic) numbers, then y_n/x_n tends to zero as $n \to \infty$. The notation $y_n = O(x_n)$ means that there is a constant $C \geq 0$ such that $|y_n| \leq C|x_n|$ for all n. Now let X_n and Y_n be two sequences of random numbers. For random numbers counterparts of $o(\cdot)$ and $O(\cdot)$ are defined as follows. 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 $\mathsf{Prob}\{|Y_n| > c\} \leq \varepsilon$ for all n. The notation $Y_n = O_p(X_n)$ means that Y_n/X_n is bounded in probability. These notations $o_p(\cdot)$ and $O_p(\cdot)$ can be viewed as probabilistic analogues of their deterministic counterparts $o(\cdot)$ and $O(\cdot)$ and have similar properties. For example if $X_n = o_p(1)$ and $Y_n = O_p(1)$, then $X_nY_n = o_p(1)$.

Recall that X_n converges in distribution to a random variable X, denoted $X_n \rightsquigarrow X$, if for any number x such that $\mathsf{Prob}\{X = x\} = 0$ it follows that

$$\lim_{n\to\infty} \mathsf{Prob}\{X_n \leq x\} = \mathsf{Prob}\{X \leq x\}. \tag{6.1}$$

Note that condition $\mathsf{Prob}\{X = x\} = 0$ means that the cumulative distribution function (cdf) $F(x) = \mathsf{Prob}(X \leq x)$ of X is continuous at x, and condition (6.1) means that $\lim_{n\to\infty} F_n(x) = F(x)$, where $F_n(\cdot)$ is the cdf of X_n. That is, X_n converges in distribution to X if the cdf of X_n converges to the cdf of X at every point where the cdf of X is continuous.

A sequence of random vectors $\boldsymbol{X}_n \in \mathbb{R}^m$ converges in distribution to a random vector X if

$$\lim_{n\to\infty} \mathsf{Prob}\{\boldsymbol{X}_n \in A\} = \mathsf{Prob}\{\boldsymbol{X} \in A\}$$

for any rectangular set $A = \{\boldsymbol{x} : a_i \leq x_i \leq b_i,\ i = 1, ..., m\}$ such that probability of $\boldsymbol{X}$ to be on the boundary of A is zero.

Proposition 6.1 *If $X_n \rightsquigarrow X$, then $X_n = O_p(1)$, i.e., if X_n converges in distribution, then X_n is bounded in probability.*

Proof. Let $F_n(x) = \mathsf{Prob}(X_n \leq x)$ and $F(x) = \mathsf{Prob}(X \leq x)$ be cumulative distribution functions of X_n and X, respectively, and $\varepsilon > 0$. Recall that X_n converges in distribution to X iff $\lim_{n\to\infty} F_n(x) = F(x)$ for every $x \in \mathbb{R}$ such that $F(\cdot)$ is continuous at x. Therefore we have that $F_n(x) \to F(x)$ provided that F is continuous at x. Since $F(x) \to 1$ as $x \to +\infty$ and $F(x) \to 0$ as $x \to -\infty$, there exists a constant c_0 such that $F(c_0) > 1 - \varepsilon$ and $F(-c_0) < \varepsilon$. Moreover, since a monotonically nondecreasing function can have only a countable number of discontinuous points, we can choose this constant c_0 such that F is continuous at c_0. It follows that there exists N such that $F_n(c_0) \geq 1 - 2\varepsilon$ and $F_n(c_0) \leq 2\varepsilon$ for all $n \geq N$. That is, $\mathsf{Prob}(|X_n| \geq c_0) \leq 4\varepsilon$. Now for every k there is a constant c_k such that $\mathsf{Prob}(|X_k| \geq c_k) \leq \varepsilon$. Then for $c = \max\{c_0, c_1, ..., c_N\}$ we have that $\mathsf{Prob}(|X_n| \geq c) \leq 4\varepsilon$ for all $n \in \mathbb{N}$. That is, for any $\varepsilon > 0$ there is c such that $\mathsf{Prob}(|X_n| \geq c) \leq 4\varepsilon$ for all n. This shows that X_n is bounded in probability. □

Theorem 6.1 (Slutsky's theorem) *If $X_n \rightsquigarrow X$ and $Y_n \xrightarrow{p} 0$, then $X_n + Y_n \rightsquigarrow X$.*

Proof. Consider the cdf $F(x) = \mathsf{Prob}\{X \leq x\}$ of X. Let x be such that $F(\cdot)$ is continuous at x. We need to show that $\mathsf{Prob}\{X_n + Y_n \leq x\}$ tends to $F(x)$ as $n \to \infty$. For $\varepsilon > 0$ we have

$$\begin{aligned} \mathsf{Prob}(X_n + Y_n \leq x) &= \mathsf{Prob}(X_n + Y_n \leq x, |Y_n| \leq \varepsilon) + \mathsf{Prob}(X_n + Y_n \leq x, |Y_n| > \varepsilon) \\ &\leq \mathsf{Prob}(X_n \leq x + \varepsilon) + \mathsf{Prob}(|Y_n| > \varepsilon). \end{aligned}$$

Since $Y_n \xrightarrow{p} 0$ we have that $\mathsf{Prob}(|Y_n| > \varepsilon)$ tends to zero. Moreover let $\varepsilon > 0$ be such that $F(\cdot)$ is continuous at $x + \varepsilon$. Then since $X_n \rightsquigarrow X$, we have that $\mathsf{Prob}(X_n \leq x + \varepsilon)$ tends to $F(x + \varepsilon)$. It follows that

$$\limsup_{n\to\infty} \mathsf{Prob}(X_n + Y_n \leq x) \leq F(x + \varepsilon).$$

Note that since $F(x)$ is monotonically nondecreasing, the set of points where it is discontinuous is countable. Therefore we can choose a sequence $\varepsilon_n \downarrow 0$ such that $F(\cdot)$ is continuous at $x + \varepsilon_n$. By continuity of $F(\cdot)$ at x it follows that

$$\limsup_{n\to\infty} \mathsf{Prob}(X_n + Y_n \leq x) \leq F(x).$$

In a similar way it is possible to show that

$$\liminf_{n\to\infty} \mathsf{Prob}(X_n + Y_n \leq x) \geq F(x).$$

It follows that $\mathsf{Prob}\{X_n + Y_n \leq x\}$ tends to $F(x)$. □

The concept of 'bounded in probability' can be extended to a sequence $\boldsymbol{Y}_n \in \mathbb{R}^m$ of random vectors. That is, $\boldsymbol{Y}_n$ is bounded in probability if for any $\varepsilon > 0$ there is a bounded set $A \subset \mathbb{R}^m$ such that $\mathsf{Prob}\{\boldsymbol{Y}_n \notin A\} \leq \varepsilon$ for all n. It is not difficult to show that $\boldsymbol{Y}_n$ is bounded in probability iff its every component Y_{in} is bounded in probability.

Slutsky's theorem also can be extended to random vectors. That is, if $\boldsymbol{X}_n$ converges in distribution to $\boldsymbol{X}$ and $\boldsymbol{Y}_n$ converges in probability to $\boldsymbol{0}$, then $\boldsymbol{X}_n + \boldsymbol{Y}_n$ converges in distribution to $\boldsymbol{X}$.

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 $\lambda_n(\boldsymbol{X}_n - \boldsymbol{\mu}) \rightsquigarrow \boldsymbol{Z}$, where $\boldsymbol{\mu} \in \mathbb{R}^m$ and $\lambda_n \to \infty$, and that $\boldsymbol{g}(\cdot)$ is differentiable at $\boldsymbol{\mu}$ with $\nabla \boldsymbol{g}(\boldsymbol{\mu}) = \partial \boldsymbol{g}(\boldsymbol{\mu})/\partial \boldsymbol{x}$ being the $m \times k$ matrix of partial derivatives (Jacobian matrix). Then*

$$\lambda_n\big(g(\boldsymbol{X}_n) - g(\boldsymbol{\mu})\big) \rightsquigarrow [\nabla g(\boldsymbol{\mu})]' \boldsymbol{Z}. \tag{6.2}$$

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 $\boldsymbol{0}$ as $\boldsymbol{x} \to \boldsymbol{\mu}$. Hence

$$\lambda_n\big(\boldsymbol{g}(\boldsymbol{X}_n) - \boldsymbol{g}(\boldsymbol{\mu})\big) = [\nabla \boldsymbol{g}(\boldsymbol{\mu})]'[\lambda_n(\boldsymbol{X}_n - \boldsymbol{\mu})] + \boldsymbol{\varepsilon}(\boldsymbol{X}_n)\big[\lambda_n \|\boldsymbol{X}_n - \boldsymbol{\mu}\|\big]. \tag{6.3}$$

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 $\boldsymbol{\varepsilon}(\boldsymbol{X}_n) \xrightarrow{p} \boldsymbol{0}$, and thus

$\boldsymbol{\varepsilon}(\boldsymbol{X}_n)\big[\lambda_n\|\boldsymbol{X}_n-\boldsymbol{\mu}\|\big]\xrightarrow{p}\mathbf{0}$. By Slutsky's theorem the convergence (6.2) follows from (6.3). □

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 $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$, 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 $k=1$, i.e., when $g:\mathbb{R}^m\to\mathbb{R}$ is a real valued function, $\nabla g(\boldsymbol{\mu})$ becomes the gradient $\nabla g(\boldsymbol{\mu})=(\partial g(\boldsymbol{\mu})/\partial x_1,...,\partial g(\boldsymbol{\mu})/\partial x_m)'$ and $[\nabla g(\boldsymbol{\mu})]'\boldsymbol{\Sigma}[\nabla g(\boldsymbol{\mu})]$ becomes the asymptotic variance of $\sqrt{n}\big(g(\boldsymbol{X}_n)-g(\boldsymbol{\mu})\big)$.

Example 6.1 Let X_n and Y_n be two independent sequences of random variables such that $\sqrt{n}\big(X_n-\mu_x\big)\rightsquigarrow\mathcal{N}(0,\sigma_x^2)$ and $\sqrt{n}\big(Y_n-\mu_y\big)\rightsquigarrow\mathcal{N}(0,\sigma_y^2)$, $\mu_y\neq 0$. 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 $\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$ with

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

That is, elements of the asymptotic covariance matrix $\boldsymbol{\Sigma}$ are: $\sigma_{11}=\mu_y^2\sigma_x^2+\mu_x^2\sigma_y^2$, $\sigma_{22}=\sigma_x^2/\mu_y^2+\big(\mu_x^2/\mu_y^4\big)\,\sigma_y^2$, $\sigma_{12}=\sigma_x^2-(\mu_x/\mu_y)^2\sigma_y^2$. In particular, if $\mu_x=\mu_y$ and $\sigma_x^2=\sigma_y^2$, then $\sigma_{12}=0$. In that case $V_n=X_nY_n$ and $W_n=X_n/Y_n$ are asymptotically independent. □

Delta method can be extended to hight order terms. For example suppose that $g:\mathbb{R}^m\to\mathbb{R}$ is twice continuously differentiable and $\nabla g(\boldsymbol{\mu})=\mathbf{0}$. Then the right hand side of (6.2) degenerates to 0. Let $\boldsymbol{H}$ be the $m\times m$ Hessian matrix of second order partial derivatives at $\boldsymbol{x}=\boldsymbol{\mu}$, i.e., $H_{ij}=\frac{\partial^2 g(\boldsymbol{\mu})}{\partial x_i\partial x_j}$, $i,j=1,...,m$. The second order expansion of $g(\cdot)$ at $\boldsymbol{\mu}$ is

$$g(\boldsymbol{x})-g(\boldsymbol{\mu})=\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{H}(\boldsymbol{x}-\boldsymbol{\mu})+r(\boldsymbol{x}),$$

where the remainder $r(\boldsymbol{x})$ is of order $o(\|\boldsymbol{x}-\boldsymbol{\mu}\|^2)$, i.e., $r(\boldsymbol{x})/\|\boldsymbol{x}-\boldsymbol{\mu}\|^2$ tends to 0 as $\boldsymbol{x}\to\boldsymbol{\mu}$. Suppose further that $\sqrt{n}(\boldsymbol{X}_n-\boldsymbol{\mu})\rightsquigarrow\boldsymbol{Z}$. Then

$$n\big(g(\boldsymbol{X}_n)-g(\boldsymbol{\mu})\big)\rightsquigarrow\tfrac{1}{2}\boldsymbol{Z}'\boldsymbol{H}\boldsymbol{Z}. \tag{6.4}$$

That is, $2n\big(g(\boldsymbol{X}_n)-g(\boldsymbol{\mu})\big)$ converges in distribution to the quadratic form $Q=\boldsymbol{Z}'\boldsymbol{H}\boldsymbol{Z}$.

7 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 (pdf) is of the form

$$f(\boldsymbol{x},\boldsymbol{\theta})=\exp\Big[\textstyle\sum_{i=1}^k\theta_iT_i(\boldsymbol{x})-A(\boldsymbol{\theta})\Big]h(\boldsymbol{x}), \tag{7.1}$$

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

$$\Theta=\Big\{\boldsymbol{\theta}:\textstyle\int\exp\Big[\sum_{i=1}^k\theta_iT_i(\boldsymbol{x})\Big]h(\boldsymbol{x})d\boldsymbol{x}<\infty\Big\}.$$

Let us show that for $T_j=T_j(\boldsymbol{X})$ and $\mathbb{E}_\theta(T_j)=\int T_j(\boldsymbol{x})f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}$,

$$\mathbb{E}_\theta(T_j)=\frac{\partial}{\partial\theta_j}A(\boldsymbol{\theta}), \tag{7.2}$$

$$\mathrm{Cov}(T_j, T_\ell) = \frac{\partial^2}{\partial\theta_j \partial\theta_\ell} A(\boldsymbol{\theta}). \tag{7.3}$$

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

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

and hence

$$\begin{aligned} 0 = \frac{\partial}{\partial\theta_j} \int f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} &= \int \frac{\partial}{\partial\theta_j} f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} \\ &= \mathbb{E}_\theta \Big[T_j - \frac{\partial}{\partial\theta_j} A(\boldsymbol{\theta}) \Big] = \mathbb{E}_\theta(T_j) - \frac{\partial}{\partial\theta_j} A(\boldsymbol{\theta}). \end{aligned}$$

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 $\frac{\partial^2}{\partial\theta_j \partial\theta_\ell} \int f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} = 0$.

8 Point estimation

8.1 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. Since $\log x$ is monotonically increasing for $x > 0$, this can be written as

$$\hat{\boldsymbol{\theta}}_N \in \arg\max_{\boldsymbol{\theta}\in\Theta} \log L_N(\boldsymbol{\theta}). \tag{8.1}$$

Note that such maximizer may not exist or could be not unique. We assume that the random sample is an iid replication of random vector $\boldsymbol{X}$ having pdf $g(\boldsymbol{x})$, written $\boldsymbol{X} \sim g(\cdot)$, i.e., each $\boldsymbol{X}_i$ has pdf $g(\cdot)$. 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

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

provided this expectation is well defined and finite. The notation $\mathbb{E}_g$ emphasizes that the expectation is taken with respect to the distribution of the sample defined by the pdf $g(\cdot)$. It is natural then to expect that the ML estimator $\hat{\boldsymbol{\theta}}_N$ will converge w.p.1 to a maximizer of $\mathbb{E}_g\big[\log f(\boldsymbol{X}, \boldsymbol{\theta}) \big]$ 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 8.1 (Jensen inequality) *Let $\phi : \mathbb{R}^m \to \mathbb{R}$ be a convex function and $\boldsymbol{X}$ be an $m \times 1$ random vector having mean $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$. Then*

$$\mathbb{E}[\phi(\boldsymbol{X})] \geq \phi(\boldsymbol{\mu}). \tag{8.2}$$

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 (8.2) follows. □

Kullback-Leibler divergence of pdf $f(\cdot)$ from pdf $g(\cdot)$ is defined as

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

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

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

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

Since

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

we have that maximizing $\mathbb{E}_g[\log f(\boldsymbol{X},\boldsymbol{\theta})]$, over $\boldsymbol{\theta} \in \Theta$, is equivalent to minimizing the Kullback-Leibler 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 $\mathbb{E}_{\theta^*}[\log f(\boldsymbol{X},\boldsymbol{\theta})]$, over $\boldsymbol{\theta} \in \Theta$, where the notation $\mathbb{E}_{\theta^*}$ emphasizes that the expectation is taken with respect to the distribution $g(\cdot) = f(\cdot,\boldsymbol{\theta}^*)$. That is

$$\boldsymbol{\theta}^* \in \arg\max_{\theta\in\Theta} \left\{ \mathbb{E}_{\theta^*}[\log f(\boldsymbol{X},\boldsymbol{\theta})] = \int \big[\log f(\boldsymbol{x},\boldsymbol{\theta})\big] f(\boldsymbol{x},\boldsymbol{\theta}^*) d\boldsymbol{x} \right\}.$$

It follows that if the model is correctly specified and 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 $\boldsymbol{\theta}^*$.

8.1.1 Asymptotic distribution of the ML estimators

Let $\boldsymbol{X} \sim f(\boldsymbol{x},\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^k$, be a random vector. The following $k \times k$ matrix is called (Fisher) information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) := \mathbb{E}_\theta \left\{ \Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{X},\boldsymbol{\theta})\Big] \Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{X},\boldsymbol{\theta})\Big]' \right\}. \tag{8.3}$$

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

Let us show that

$$\boldsymbol{I}(\boldsymbol{\theta}) = -\mathbb{E}_\theta \left\{ \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(\boldsymbol{X},\boldsymbol{\theta}) \right\}. \tag{8.4}$$

We need to show that

$$\mathbb{E}_\theta\left\{\frac{\partial \log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial \log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_j}\right\} = -\mathbb{E}_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{X},\boldsymbol{\theta})\right\}, \tag{8.5}$$

$i,j=1,...,k$. We have that

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

Suppose now that

$$\int\frac{\partial f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}d\boldsymbol{x} = \frac{\partial}{\partial\theta_i}\int f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}, \tag{8.6}$$

i.e., the operations of differentiation with respect to θ_i and integration with respect to $\boldsymbol{x}$ can be interchanged. Then the right hand side of (8.6) is 0, since $\int f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}=1$ for all $\boldsymbol{\theta}$. It follows that

$$\mathbb{E}_\theta\left\{\frac{\partial \log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\right\} = \int\frac{\partial \log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x} = 0, \tag{8.7}$$

for all $\boldsymbol{\theta}$ and hence

$$\frac{\partial}{\partial\theta_j}\int\frac{\partial \log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x} = 0. \tag{8.8}$$

By taking the derivative, in the left hand side of (8.8), inside the integral we obtain

$$\begin{aligned}
0 &= \int\frac{\partial}{\partial\theta_j}\left[\frac{\partial \log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}f(\boldsymbol{x},\boldsymbol{\theta})\right]d\boldsymbol{x}\\
&= \int\frac{\partial^2 \log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x} + \int\frac{\partial \log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial \log f(\boldsymbol{x},\boldsymbol{\theta})}{\partial\theta_j}f(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{x}\\
&= \mathbb{E}_\theta\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{X},\boldsymbol{\theta})\right\} + \mathbb{E}_\theta\left\{\frac{\partial \log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\,\frac{\partial \log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_j}\right\},
\end{aligned} \tag{8.9}$$

and hence (8.5) follows. □

Remark 8.1 The above derivations are based on the interchangeability property that the operations of differentiation with respect to θ_i and integration with respect to $\boldsymbol{x}$ can be interchanged. We used it twice, in (8.6) and again in (8.9). As it is discussed below the interchangeability property (8.6) holds if $f(\boldsymbol{x},\cdot)$ is differentiable and there is nonnegative valued function $K(\boldsymbol{x})$ such that $\mathbb{E}[K(\boldsymbol{X})]<\infty$ and

$$|f(\boldsymbol{x},\boldsymbol{\theta}_1)-f(\boldsymbol{x},\boldsymbol{\theta}_2)| \le K(\boldsymbol{x})\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|,\ \ \boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathbb{R}^k,$$

i.e., if $f(\boldsymbol{x},\cdot)$ is Lipschitz continuous with integrable Lipschitz constant. Similar condition is needed for $\partial f(\boldsymbol{x},\boldsymbol{\theta})/\partial\theta_i$, $i=1,...,k$, in order to justify (8.9).

Let us discuss conditions ensuring that the expectation and differentiation can be interchanged. Let $g(x,\theta)$ be a real valued function of $x,\theta\in\mathbb{R}$. Suppose that $g(x,\theta)$ is differentiable in θ. We would like to verify that

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

where the expectation is with respect to distribution of random variable X. We have

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

In order to interchange the limit and the expectation (integration) we can use the Lebesgue Dominated Convergence Theorem: if $f_n, g : \Omega \to \mathbb{R}$ are such that $|f_n| \le g$, $\int_\Omega g dP < \infty$ and $f_n(\omega) \to f(\omega)$ for a.e. $\omega \in \Omega$, then $\int_\Omega f_n dP \to \int_\Omega f dP$.

That is, suppose that there is function $K(x) \ge 0$ such that $\mathbb{E}[K(X)] < \infty$ and for all h,

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

Then by the Lebesgue Dominated Convergence Theorem, the limit and the expectation (integration) can be interchanged and hence

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

□

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

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

where

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

and hence

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

It follows that

$$\boldsymbol{a}'\boldsymbol{I}(\boldsymbol{\theta})\boldsymbol{a} = \mathbb{E}_\theta\left\{\left[\sum_{i=1}^{k} a_i\frac{\partial \log f(\boldsymbol{X},\boldsymbol{\theta})}{\partial\theta_i}\right]^2\right\},$$

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

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 $\boldsymbol{\theta}^*$, i.e., $\hat{\boldsymbol{\theta}}_N$ converges w.p.1 to $\boldsymbol{\theta}^*$. Suppose further that $\boldsymbol{\theta}^*$ is an interior point of the set Θ. Since $\hat{\boldsymbol{\theta}}_N$ is a consistent estimator of $\boldsymbol{\theta}^*$, it follows that $\hat{\boldsymbol{\theta}}_N$ is in the interior of the set Θ for all N large enough. Then since $\hat{\boldsymbol{\theta}}_N$ is a maximizer of $\log L_N(\boldsymbol{\theta})$, the following optimality condition holds

$$\frac{\partial}{\partial\boldsymbol{\theta}}\left[\sum_{i=1}^{N}\log f(\boldsymbol{X}_i,\hat{\boldsymbol{\theta}}_N)\right] = \boldsymbol{0}. \tag{8.10}$$

By the Mean Value Theorem we can write

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

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

$$\begin{aligned}\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*) &= -\sqrt{N}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\sum_{i=1}^N \log f(\boldsymbol{X}_i, \tilde{\boldsymbol{\theta}}_N)\right]^{-1}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\sum_{i=1}^N \log f(\boldsymbol{X}_i, \boldsymbol{\theta}^*)\right] \\ &= -\left[\frac{1}{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\sum_{i=1}^N \log f(\boldsymbol{X}_i, \tilde{\boldsymbol{\theta}}_N)\right]^{-1}\left[\frac{1}{\sqrt{N}}\frac{\partial}{\partial\boldsymbol{\theta}}\sum_{i=1}^N \log f(\boldsymbol{X}_i, \boldsymbol{\theta}^*)\right]. \quad (8.11)\end{aligned}$$

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

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

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 (8.11) this implies that

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}^*) \rightsquigarrow \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}^*)^{-1}\right). \quad (8.12)$$

That is, the ML estimator $\hat{\boldsymbol{\theta}}_N$ has approximately normal distribution with mean $\boldsymbol{\theta}^*$ and covariance matrix $N^{-1}\boldsymbol{I}(\boldsymbol{\theta}^*)^{-1}$. □

Remark 8.2 The above derivations of the basic result (8.12) involve several assumptions (regularity conditions). The asymptotic result (8.12) is local, it is based on the second order approximation of the likelihood function at the true value θ^*. So the MLE estimator should be consistent in order to justify such approximations. In order to apply necessary condition (8.10), the MLE should be an interior point of the set Θ, i.e., should not be on the boundary of the set Θ. If θ^* is a boundary point of Θ, then the asymptotics of the MLE is different. We also needed the interchangeability property, that the operations of integration with respect to $\boldsymbol{x}$ and differentiation with respect to $\boldsymbol{\theta}$ can be interchanged (see Remark 8.1). □

Example 8.1 Suppose that $X_1, ..., X_N$ are iid having uniform distribution on the interval $[0, \theta]$, $\theta > 0$, with pdf $f(x, \theta) = 1/\theta$ for $x \in [0, \theta]$, and $f(x, \theta) = 0$ otherwise. Hence the likelihood function is $L_N(\theta) = 1/\theta^N$ for $X_{(N)} \leq \theta$, where $X_{(N)} = \max\{X_1, ..., X_N\}$. Since $L_N(\theta)$ is monotonically decreasing with increase of θ, the MLE is given by the smallest possible value of θ which is $X_{(N)}$. That is $X_{(N)}$ is the MLE of θ.

The cdf of X_i is $F(x) = x/\theta$ for $x \in [0, \theta]$. Then the cdf of $N\big[\theta - X_{(N)}\big]$, for $x \in [0, N\theta]$, is

$$\begin{aligned}\mathsf{Prob}\left(N\big[\theta - X_{(N)}\big] \leq x\right) &= \mathsf{Prob}\big(X_{(N)} \geq \theta - x/N\big) = 1 - \mathsf{Prob}\big(X_{(N)} < \theta - x/N\big) \\ &= 1 - \mathsf{Prob}\big(X_i < \theta - x/N,\ i = 1, ..., N\big) \\ &= 1 - \prod_{i=1}^N \mathsf{Prob}\big(X_i < \theta - x/N\big) \\ &= 1 - [F(\theta - x/N)]^N = 1 - (1 - x/(N\theta))^N.\end{aligned}$$

Furthermore

$$\lim_{N\to\infty}\big(1-x/(N\theta)\big)^N = e^{-x/\theta}.$$

It follows that the cdf of $N\big[\theta - X_{(N)}\big]$ converges to $1-e^{-x/\theta}$. This implies that $N\big[\theta - X_{(N)}\big]$ converges in distribution to exponential $\exp(\lambda)$ with $\lambda = 1/\theta$. Note that the situation here is not standard, the optimality equation (8.10) is not applicable here. Also the asymptotic variance is of order $O(N^{-2})$ rather than $O(N^{-1})$ as in the standard case. □

8.2 Cramér - Rao lower bound

Let $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$ be an iid sample from $f(\boldsymbol{x}, \theta)$, $\theta \in \mathbb{R}$, and $T(\boldsymbol{X})$ be a statistic, i.e., $T(\boldsymbol{X})$ is a function of $\boldsymbol{X}$. Note that $f(\boldsymbol{x}, \theta) = \prod_{j=1}^N f_j(\boldsymbol{x}_j, \theta)$ is the pdf of $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$, where $f_j(\boldsymbol{x}_j, \theta)$ is pdf of $\boldsymbol{X}_j$. Since the sample is iid, pdfs $f_j(\cdot, \theta)$ are the same for all $j = 1, ..., N$.

Then under some regularity conditions

$$\mathrm{Var}_\theta[T(\boldsymbol{X})] \geq i_X(\theta)^{-1}[\partial g(\theta)/\partial\theta]^2, \tag{8.13}$$

where $g(\theta) := \mathbb{E}_\theta[T(\boldsymbol{X})]$ and

$$i_X(\theta) = \mathbb{E}_\theta\left[\Big(\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}, \theta)\Big)^2\right] \tag{8.14}$$

is Fisher's information of $f(\boldsymbol{x}, \theta)$. In particular, if $\mathbb{E}_\theta[T(\boldsymbol{X})] = \theta$, i.e. $T(\boldsymbol{X})$ is an unbiased estimator of θ, then

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

Note that, by the independence of $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$,

$$i_X(\theta) = \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X}, \theta)\right] = \sum_{j=1}^N \mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f_j(\boldsymbol{X}_j, \theta)\right] = Ni(\theta),$$

where $i(\theta) = \mathbb{E}_\theta\left[(\frac{\partial}{\partial\theta}\log f_j(\boldsymbol{X}_j, \theta))^2\right]$ is the information number of individual $\boldsymbol{X}_j$.

Proof. 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 interchanged with the integral (see Remark 8.1). Then

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

That is,

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

Now by Cauchy inequality we have

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

Moreover

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

and hence the inequality (8.13) follows. □

This bound can be extended to a multivariate setting.

Theorem 8.2 (multivariate Cramér - Rao lower bound) *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid sample from $f(\boldsymbol{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^k$, and $T = T(\boldsymbol{X})$ be an unbiased estimator of $\boldsymbol{\theta}$, i.e., $\mathbb{E}_\theta[T(\boldsymbol{X})] = \boldsymbol{\theta}$. Suppose that the information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ is nonsingular and the interchangeability property holds. Then*

$$\mathrm{Var}_\theta\Big(\sum_{i=1}^k a_i T_i\Big) \geq \boldsymbol{a}' \boldsymbol{I}(\boldsymbol{\theta})^{-1} \boldsymbol{a} \tag{8.15}$$

for any $\boldsymbol{a} \in \mathbb{R}^k$.

Proof. For $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$ we have (by using the interchangeability property)

$$\mathrm{Cov}_\theta\Big(\sum_{i=1}^k a_i T_i, \sum_{j=1}^k b_j \frac{\partial}{\partial \theta_j} \log f(\boldsymbol{X}, \boldsymbol{\theta})\Big) = \mathrm{Cov}_\theta\Big(\sum_{i=1}^k a_i T_i, \sum_{j=1}^k b_j \frac{\partial f(\boldsymbol{X}, \boldsymbol{\theta})/\partial \theta_j}{f(\boldsymbol{X}, \boldsymbol{\theta})}\Big) =$$

$$\int \Big(\sum_{i=1}^k a_i T_i(\boldsymbol{x})\Big)\Big(\sum_{j=1}^k b_j \partial f(\boldsymbol{x}, \boldsymbol{\theta})/\partial \theta_j\Big) d\boldsymbol{x} = \sum_{j=1}^k b_j \frac{\partial}{\partial \theta_j} \int \Big(\sum_{i=1}^k a_i T_i(\boldsymbol{x})\Big) f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x} =$$

$$\sum_{j=1}^k b_j \frac{\partial}{\partial \theta_j} \mathbb{E}_\theta\Big[\sum_{i=1}^k a_i T_i\Big].$$

It follows by Cauchy inequality

$$\left(\sum_{j=1}^k b_j \frac{\partial}{\partial \theta_j} \mathbb{E}_\theta\Big[\sum_{i=1}^k a_i T_i\Big]\right)^2 \leq \mathrm{Var}_\theta\Big(\sum_{i=1}^k a_i T_i\Big) \mathbb{E}_\theta\left[\Big(\sum_{j=1}^k b_j \frac{\partial}{\partial \theta_j} \log f(\boldsymbol{X}, \boldsymbol{\theta})\Big)^2\right].$$

Since T is unbiased we have that $\mathbb{E}_\theta\Big[\sum_{i=1}^k a_i T_i\Big] = \sum_{i=1}^k a_i \theta_i$, and hence $\frac{\partial}{\partial \theta_j} \mathbb{E}_\theta\Big[\sum_{i=1}^k a_i T_i\Big] = a_j$. Also $\mathrm{Var}_\theta\Big(\sum_{i=1}^k a_i T_i\Big) = \boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a}$, where $\boldsymbol{\Sigma}$ is the covariance matrix of T, and

$$\mathbb{E}_\theta\left[\Big(\sum_{j=1}^k b_j \frac{\partial}{\partial \theta_j} \log f(\boldsymbol{X}, \boldsymbol{\theta})\Big)^2\right] = \boldsymbol{b}' \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{b}.$$

We obtain that

$$(\boldsymbol{a}' \boldsymbol{b})^2 \leq (\boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a})(\boldsymbol{b}' \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{b}).$$

It follows that

$$\boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a} \geq \max_{\boldsymbol{b} \neq 0} \frac{\boldsymbol{b}' (\boldsymbol{a}\boldsymbol{a}') \boldsymbol{b}}{\boldsymbol{b}' \boldsymbol{I}(\boldsymbol{\theta}) \boldsymbol{b}}.$$

The maximum in the right hand side of the above inequality is attained for $\boldsymbol{b} = \boldsymbol{I}(\boldsymbol{\theta})^{-1} \boldsymbol{a}$ (see section 13.1.1), and hence this maximum is equal to $\boldsymbol{a}' \boldsymbol{I}(\boldsymbol{\theta})^{-1} \boldsymbol{a}$. Therefore we obtain that

$$\boldsymbol{a}' \boldsymbol{\Sigma} \boldsymbol{a} \geq \boldsymbol{a}' \boldsymbol{I}(\boldsymbol{\theta})^{-1} \boldsymbol{a} \tag{8.16}$$

for any $\boldsymbol{a} \in \mathbb{R}^k$. □

Definition 8.1 *It is said that a sequence of estimators W_N is* asymptotically efficient *for $\boldsymbol{\theta}$ if $\sqrt{N}(W_N - \boldsymbol{\theta})$ converges in distribution to normal $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{I}(\boldsymbol{\theta})^{-1}$.*

The basic result (8.12), of asymptotic normality of the ML estimator $\hat{\boldsymbol{\theta}}_N$, shows that the MLE is asymptotically efficient. That is, in the standard case, under the corresponding regularity conditions, the MLE attains asymptotically the smallest possible variance. It could be noted that the bound (8.15) is not asymptotic. On the other hand, it assumes that the estimator T is unbiased, while the ML estimators often are biased. There are some other concepts of the "best possible" estimators. In the next section we briefly discuss some basic concepts.

8.3 Best unbiased estimators

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid random sample and $f(\boldsymbol{x}, \boldsymbol{\theta})$ be pdf of $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$. By writing P_θ it is emphasized that the probability distribution of $\boldsymbol{X}$ depends on the parameter vector $\boldsymbol{\theta}$.

Definition 8.2 *A statistic $T(\boldsymbol{X})$ is a* sufficient *statistic for $\boldsymbol{\theta}$ if the conditional distribution of sample $\boldsymbol{X}$ given $T(\boldsymbol{X})$ does not depend on $\boldsymbol{\theta}$. That is, $\mathsf{Prob}(\boldsymbol{X} \in A | T = t)$ is independent of $\boldsymbol{\theta}$ for all (measurable) sets A and t in the range of T.*

Note that a sufficient statistic always exists, take for example $T(\boldsymbol{X}) = \boldsymbol{X}$.

Theorem 8.3 (Fisher - Neyman factorization criterion) *Suppose that $\boldsymbol{X}$ has pdf $f(\boldsymbol{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Then $T = T(\boldsymbol{X})$ is sufficient for $\boldsymbol{\theta}$ iff $f(\boldsymbol{x}, \boldsymbol{\theta}) = g(T(\boldsymbol{x}), \boldsymbol{\theta}) h(\boldsymbol{x})$.*

Proof. (Sketch for discrete distribution)
Suppose that T is sufficient. Then Since $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$ has discrete distribution, the weight function $f(\boldsymbol{x}, \boldsymbol{\theta}) = P_\theta(\boldsymbol{X} = \boldsymbol{x})$, where $\boldsymbol{x}$ is in the range of $\boldsymbol{X}$. Moreover

$$P_\theta(\boldsymbol{X} = \boldsymbol{x}) = \sum_t P_\theta(\boldsymbol{X} = \boldsymbol{x}, T = t),$$

where the summation over possible values of T. Also since $T = T(\boldsymbol{X})$,

$$\sum_t P_\theta(\boldsymbol{X} = \boldsymbol{x}, T = t) = P_\theta(\boldsymbol{X} = \boldsymbol{x}, T = T(\boldsymbol{x})).$$

Then

$$P_\theta\big(\boldsymbol{X} = \boldsymbol{x}, T = T(\boldsymbol{x})\big) = P_\theta\big(T = T(\boldsymbol{x})\big) P_\theta(\boldsymbol{X} = \boldsymbol{x} | T = T(\boldsymbol{x})),$$

where is used formula $\mathsf{Prob}(A \cap B) = \mathsf{Prob}(B)\mathsf{Prob}(A|B)$ for events $B := \{T = T(\boldsymbol{x})\}$ and $A := \{\boldsymbol{X} = \boldsymbol{x}\}$.

By sufficiency of T we have that the conditional probability $h(\boldsymbol{x}) := P_\theta\big(\boldsymbol{X} = \boldsymbol{x} | T = T(\boldsymbol{x})\big)$ does not depend on θ. Define $g(T(\boldsymbol{x}), \boldsymbol{\theta}) := P_\theta(T = T(\boldsymbol{x}))$. It follows that $f(\boldsymbol{x}, \boldsymbol{\theta}) = g(T(\boldsymbol{x}), \boldsymbol{\theta}) h(\boldsymbol{x})$. This shows that sufficiency implies factorization.

Now suppose that $f(\boldsymbol{x}, \boldsymbol{\theta}) = g(T(\boldsymbol{x}), \boldsymbol{\theta}) h(\boldsymbol{x})$. Then when $T(\boldsymbol{x}) = t$ we have

$$\begin{aligned} P_\theta(\boldsymbol{X} = \boldsymbol{x} | T = t) &= \frac{P_\theta(\boldsymbol{X} = \boldsymbol{x}, T = t)}{P_\theta(T = t)} = \frac{g(T(\boldsymbol{x}), \boldsymbol{\theta}) h(\boldsymbol{x})}{\sum_{T(\boldsymbol{y}) = t} g(T(\boldsymbol{y}), \boldsymbol{\theta}) h(\boldsymbol{y})} \\ &= \frac{g(T(\boldsymbol{x}), \boldsymbol{\theta}) h(\boldsymbol{x})}{\sum_{T(\boldsymbol{y}) = t} f(\boldsymbol{y}, \boldsymbol{\theta})} = \frac{g(t, \boldsymbol{\theta}) h(\boldsymbol{x})}{\sum_{T(\boldsymbol{y}) = t} g(t, \boldsymbol{\theta}) h(\boldsymbol{y})} = \frac{h(\boldsymbol{x})}{\sum_{T(\boldsymbol{y}) = t} h(\boldsymbol{y})}, \end{aligned}$$

which does not depend on $\boldsymbol{\theta}$. If $T(\boldsymbol{x}) \neq t$, then $P_\theta(\boldsymbol{X} = \boldsymbol{x} | T = t) = 0$. It follows that $T(\boldsymbol{X})$ is sufficient. □

Let $T(\boldsymbol{X})$ be a sufficient statistic for $\boldsymbol{\theta}$. Then by the Factorization Theorem, the likelihood function

$$L_N(\boldsymbol{\theta}) = f(\boldsymbol{x}, \boldsymbol{\theta}) = g(T(\boldsymbol{x}), \boldsymbol{\theta}) h(\boldsymbol{x}).$$

It follows that the MLE $\hat{\boldsymbol{\theta}}$ is a function of T, i.e.,

$$\hat{\boldsymbol{\theta}} \in \arg\max_{\boldsymbol{\theta}} g(T(\boldsymbol{x}), \boldsymbol{\theta}).$$

Definition 8.3 *A sufficient statistic $T = T(\boldsymbol{X})$ is said to be* minimal sufficient *if for any other sufficient statistic $S = S(\boldsymbol{X})$, there exists a function $g(\cdot)$ such that $T = g(S)$.*

Theorem 8.4 (Lehmann - Scheffe) *Suppose that there exists $T(\boldsymbol{X})$ such that for any $\boldsymbol{x}$ and $\boldsymbol{y}$, the ratio $\frac{f(\boldsymbol{x},\boldsymbol{\theta})}{f(\boldsymbol{y},\boldsymbol{\theta})}$ is independent of $\boldsymbol{\theta}$ if only if $T(\boldsymbol{x}) = T(\boldsymbol{y})$. Then $T(\boldsymbol{X})$ is a minimal sufficient statistic for $\boldsymbol{\theta}$.*

Proof. Let us show that $T(\boldsymbol{X})$ is sufficient. For t in the image of $T(\boldsymbol{x})$ consider sets $A_t = \{\boldsymbol{x} : T(\boldsymbol{x}) = t\}$. For t in the image of $T(\boldsymbol{x})$, consider a point $\boldsymbol{x}_t \in A_t$. We have that $x_{T(\boldsymbol{x})}$ and $\boldsymbol{x}$ are in the same set A_t, i.e., $T(\boldsymbol{x}) = T(\boldsymbol{x}_{T(\boldsymbol{x})})$, and hence by the assumption of the theorem, the ratio $f(\boldsymbol{x},\boldsymbol{\theta})/f(\boldsymbol{x}_{T(\boldsymbol{x})},\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$. Define $h(\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{\theta})/f(\boldsymbol{x}_{T(\boldsymbol{x})},\boldsymbol{\theta}),\boldsymbol{\theta})$ and $g(t,\boldsymbol{\theta}) = f(\boldsymbol{x}_t,\boldsymbol{\theta})$. Then

$$f(\boldsymbol{x},\boldsymbol{\theta}) = \frac{f(\boldsymbol{x}_{T(\boldsymbol{x})},\boldsymbol{\theta})f(\boldsymbol{x},\boldsymbol{\theta})}{f(\boldsymbol{x}_{T(\boldsymbol{x})},\boldsymbol{\theta})} = g(T(\boldsymbol{x}),\boldsymbol{\theta})h(\boldsymbol{x}).$$

It follows by the Factorization Theorem that $T(\boldsymbol{X})$ is a sufficient statistic.

Let us show that $T(\boldsymbol{X})$ is minimal sufficient. Let $T'(\boldsymbol{X})$ be a sufficient statistic. By the Factorization Theorem, $f(\boldsymbol{x},\theta) = g(T'(\boldsymbol{x}),\boldsymbol{\theta})h(\boldsymbol{x})$. Suppose that $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$. Then

$$\frac{f(\boldsymbol{x},\boldsymbol{\theta})}{f(\boldsymbol{y},\boldsymbol{\theta})} = \frac{g(T'(\boldsymbol{x}),\boldsymbol{\theta})h(\boldsymbol{x})}{g(T'(\boldsymbol{y}),\boldsymbol{\theta})h(\boldsymbol{y})} = \frac{h(\boldsymbol{x})}{h(\boldsymbol{y})}.$$

Since this ratio does not depend on $\boldsymbol{\theta}$, it follows by the assumption of theorem that $T(\boldsymbol{x}) = T(\boldsymbol{y})$. That is, $T'(\boldsymbol{x}) = T'(\boldsymbol{y})$ implies that $T(\boldsymbol{x}) = T(\boldsymbol{y})$. It follows that $T(\boldsymbol{X})$ is a function of $T'(\boldsymbol{X})$. □

Note that the second part of the above proof shows that a sufficient statistic $T(\boldsymbol{X})$ is minimal sufficient if the following implication holds: if the ratio $f(\boldsymbol{x},\boldsymbol{\theta})/f(\boldsymbol{y},\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, then $T(\boldsymbol{x}) = T(\boldsymbol{y})$.

Example 8.2 Consider exponential family of distributions in the canonical form (see eq. (7.1)),

$$f(\boldsymbol{x},\boldsymbol{\theta}) = \exp\Big\{\sum_{i=1}^{k}\theta_i T_i(\boldsymbol{x}) - A(\boldsymbol{\theta})\Big\}h(\boldsymbol{x}), \tag{8.17}$$

with parameter space

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

It follows by the Factorization Theorem that $(T_1(\boldsymbol{X}),...,T_k(\boldsymbol{X}))$ is a sufficient statistic. Note that the set Θ is convex. Also

$$\frac{f(\boldsymbol{x},\boldsymbol{\theta})}{f(y,\boldsymbol{\theta})} = \exp\Big\{\sum_{i=1}^{k}\theta_i\big(T_i(\boldsymbol{x}) - T_i(\boldsymbol{y})\big)\Big\}\frac{h(\boldsymbol{x})}{h(\boldsymbol{y})}.$$

Suppose that the set Θ has a nonempty interior. Then if the ratio $f(\boldsymbol{x},\boldsymbol{\theta})/f(\boldsymbol{y},\boldsymbol{\theta})$ does not depend on $\boldsymbol{\theta}$, then $T_i(\boldsymbol{x}) = T_i(\boldsymbol{y})$, $i = 1,...,k$. Indeed if this ratio does not depend on $\boldsymbol{\theta}$, then

$$\frac{\partial}{\partial\theta_i}\exp\Big\{\sum_{i=1}^{k}\theta_i\big(T_i(\boldsymbol{x}) - T_i(\boldsymbol{y})\big)\Big\}\frac{h(\boldsymbol{x})}{h(\boldsymbol{y})} = \big(T_i(\boldsymbol{x}) - T_i(\boldsymbol{y})\big)\exp\Big\{\sum_{i=1}^{k}\theta_i\big(T_i(\boldsymbol{x}) - T_i(\boldsymbol{y})\big)\Big\}\frac{h(\boldsymbol{x})}{h(\boldsymbol{y})}$$

is zero at every interior point of the set Θ. It follows that $T_i(\boldsymbol{x}) = T_i(\boldsymbol{y})$. This implies that $(T_1(\boldsymbol{X}),...,T_k(\boldsymbol{X}))$ is minimal sufficient. □

We assume in the remainder of this section that $g : \Theta \to \mathbb{R}$ is a real valued (measurable) function.

Definition 8.4 *An estimator $T = T(\boldsymbol{X})$ of $g(\boldsymbol{\theta})$ is a best unbiased estimator if $\mathbb{E}_\theta[T] = g(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, and for any unbiased estimator $S = S(\boldsymbol{X})$ of $g(\boldsymbol{\theta})$ it follows that*

$$\mathrm{Var}_\theta[T] \leq \mathrm{Var}_\theta[S], \ \ \forall \boldsymbol{\theta} \in \Theta.$$

Best unbiased estimator is called Uniform Minimum Variance Unbiased (UMVU) *estimator.*

Finding an UMVU estimator could be not easy. The following result shows that conditioning of any unbiased estimator on a sufficient statistic will result in uniform reduction of the variance. Therefore if an UMVU estimator exists, then it is a function of (minimal) sufficient statistic.

For random variables X and Y, we use below property $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$ of conditional expectation, and the following formula

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathbb{E}(X|Y)], \tag{8.18}$$

for conditional variance

$$\mathrm{Var}(X|Y) = \mathbb{E}\big[\big(X - \mathbb{E}(X|Y)\big)^2|Y\big].$$

Indeed

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}\big[\mathbb{E}[(X - \mathbb{E}(X))^2]|Y\big] \\
&= \mathbb{E}\left[\mathbb{E}\big[\big(X - \mathbb{E}(X|Y) + \mathbb{E}(X|Y) - \mathbb{E}(X)\big)^2\big]|Y\right] \\
&= \mathbb{E}\big[\underbrace{\mathbb{E}[(X - \mathbb{E}(X|Y))^2|Y]}_{\mathrm{Var}(X|Y)}\big] + \underbrace{\mathbb{E}\big[(\mathbb{E}(X|Y) - \mathbb{E}(X))^2\big]}_{\mathrm{Var}[\mathbb{E}(X|Y)]}.
\end{aligned}$$

In the above derivation we used that

$$\mathbb{E}[(\mathbb{E}(X|Y) - \mathbb{E}(X))^2|Y] = \mathbb{E}[(\mathbb{E}(X|Y) - \mathbb{E}(X))^2],$$

since $(\mathbb{E}(X|Y) - \mathbb{E}(X))^2$ is a function of Y, and that

$$\mathbb{E}\big[(\mathbb{E}[(X - \mathbb{E}(X|Y)])(\mathbb{E}(X|Y) - \mathbb{E}(X))|Y\big] = \mathbb{E}\big[(\mathbb{E}(X - \mathbb{E}(X|Y)|Y)(\mathbb{E}(X|Y) - \mathbb{E}(X))\big] = 0.$$

Theorem 8.5 (Rao - Blackwell) *Let W be an unbiased estimator of $g(\boldsymbol{\theta})$, and T be a sufficient statistic for $\boldsymbol{\theta}$. Define $h(t) := \mathbb{E}[W|T = t]$. Then $\mathbb{E}_\theta[h(T)] = g(\boldsymbol{\theta})$ and*

$$\mathrm{Var}_\theta[h(T)] \leq \mathrm{Var}_\theta[W], \ \ \forall \boldsymbol{\theta} \in \Theta. \tag{8.19}$$

Moreover, unless $P_\theta\{W = h(T)\} = 1$, the inequality (8.19) is strict.

Proof. Note that by sufficiency of T, $h(T)$ does not depend on $\boldsymbol{\theta}$ and hence is a statistic. We have that

$$E_\theta[h(T)] = \mathbb{E}_\theta\big[\mathbb{E}[W|T = t]\big] = \mathbb{E}_\theta[W] = g(\boldsymbol{\theta}).$$

That is $h(T)$ is an unbiased estimator of $g(\boldsymbol{\theta})$. Now by using formula (8.18) for conditional variance

$$\begin{aligned}
\mathrm{Var}_\theta[W] &= \mathrm{Var}_\theta[\mathbb{E}(W|T)] + \mathbb{E}_\theta[\mathrm{Var}(W|T)] \\
&= \mathrm{Var}_\theta[h(T)] + \mathbb{E}_\theta[\mathrm{Var}(W|T)] \geq \mathrm{Var}_\theta[h(T)],
\end{aligned}$$

and hence (8.19) follows. Moreover, $\mathbb{E}_\theta[\mathrm{Var}(W|T)] > 0$ and hence the inequality (8.19) is strict unless $P_\theta\{W = h(T)\} = 1$. □

Theorem 8.6 *An UMVU estimator W (if it exists) of $g(\boldsymbol{\theta})$ is unique.*

Proof. Let W' be another UMVU estimator of $g(\boldsymbol{\theta})$. Then $W^* = (W + W')/2$ is unbiased and

$$\mathrm{Var}_\theta(W^*) = \tfrac{1}{4}\mathrm{Var}_\theta(W) + \tfrac{1}{4}\mathrm{Var}_\theta(W') + \tfrac{1}{2}\mathrm{Cov}_\theta(W, W').$$

Now by Cauchy inequality

$$\mathrm{Cov}_\theta(W, W') \leq [\mathrm{Var}_\theta(W) \cdot \mathrm{Var}_\theta(W')]^{1/2},$$

and $\mathrm{Var}_\theta(W) = \mathrm{Var}_\theta(W')$ by the minimum variance assumption. Hence $\mathrm{Var}_\theta(W^*) \leq \mathrm{Var}_\theta(W)$. Since W is UMVU it follows that $\mathrm{Var}_\theta(W^*) = \mathrm{Var}_\theta(W)$ for all $\boldsymbol{\theta} \in \Theta$. The equality in Cauchy inequality holds only if $W' = a(\boldsymbol{\theta})W + b(\boldsymbol{\theta})$. Then

$$\mathrm{Cov}_\theta(W, W') = \mathrm{Cov}_\theta(W, a(\boldsymbol{\theta})W + b(\boldsymbol{\theta})) = a(\boldsymbol{\theta})\mathrm{Var}_\theta(W).$$

Also by the above we have that $\mathrm{Cov}_\theta(W, W') = \mathrm{Var}_\theta(W)$ and hence $a(\boldsymbol{\theta}) \equiv 1$. Moreover $\mathbb{E}_\theta[W'] = g(\boldsymbol{\theta}) = \mathbb{E}_\theta[W]$ and hence $b(\boldsymbol{\theta}) \equiv 0$. It follows that $W = W'$. □

Definition 8.5 *Loss function (for estimating $g(\boldsymbol{\theta})$) is a nonnegative valued function $L(\boldsymbol{\theta}, a)$, $\boldsymbol{\theta} \in \Theta$, $a \in \mathbb{R}$, such that $L(\boldsymbol{\theta}, g(\boldsymbol{\theta})) = 0$ for all $\boldsymbol{\theta} \in \Theta$. Risk function $R(\boldsymbol{\theta}, T) := \mathbb{E}_\theta[L(\boldsymbol{\theta}, T(\boldsymbol{X}))]$, where $T(\boldsymbol{X})$ is an estimator of $g(\boldsymbol{\theta})$.*

For example $L(\boldsymbol{\theta}, a) := |a - g(\boldsymbol{\theta})|^p$, $p > 0$, is a loss function. If $L(\boldsymbol{\theta}, a) = (g(\boldsymbol{\theta}) - a)^2$, then

$$R(\boldsymbol{\theta}, T) = \mathbb{E}_\theta[(g(\boldsymbol{\theta}) - T(\boldsymbol{X}))^2]$$

is the Mean Square Error of estimator T of $g(\boldsymbol{\theta})$.

Theorem 8.7 (Another version of Rao - Blackwell theorem) *Let $L(\boldsymbol{\theta}, a)$ be a loss function, W be a sufficient statistic and $h(t) = \mathbb{E}[W|T = t]$. Suppose that $L(\boldsymbol{\theta}, \cdot)$ is strictly convex. Then*

$$R(\boldsymbol{\theta}, h(T)) \leq R(\boldsymbol{\theta}, W), \tag{8.20}$$

and the above inequality is strict unless $P_\theta\{W = h(T)\} = 1$.

Proof. By using Jensen's inequality

$$R(\boldsymbol{\theta}, h(T)) = \mathbb{E}_\theta\big[L\big(\boldsymbol{\theta}, \mathbb{E}[W|T]\big)\big] \leq \mathbb{E}_\theta\big[\mathbb{E}[L(\boldsymbol{\theta}, W)|T]\big] = \mathbb{E}_\theta[L(\boldsymbol{\theta}, W)] = R(\boldsymbol{\theta}, W).$$

The inequality (8.20) follows and this inequality is strict unless $P_\theta\{W = h(T)\} = 1$. □

9 Hypotheses testing

Let $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$ be a random sample (data). Consider testing $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus $H_1 : \boldsymbol{\theta} \in \Theta_1$, where $\Theta_0, \Theta_1 \subset \mathbb{R}^k$. A procedure for such testing consists of choosing a set $R \subset \mathbb{R}^d$, referred to as the rejection region, and hence defining its complement $R^c = \mathbb{R}^d \setminus R$ referred to as the acceptance region, where d is the dimension of $\boldsymbol{X}$. That is, reject H_0 if $\boldsymbol{X} \in R$. Alternatively this can be formulated as accept H_0 if $\boldsymbol{X} \in R^c$. Rejecting H_0 automatically means acceptance H_1, and acceptance H_0 means rejection of H_1.

There are two types of errors, 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 $\alpha = P_\theta(\text{type I error})$ and $\beta = P_\theta(\text{type II error})$. That is

$$\begin{aligned} \alpha &= P_\theta(\boldsymbol{X} \in R), \ \ \boldsymbol{\theta} \in \Theta_0, \\ \beta &= P_\theta(\boldsymbol{X} \in R^c), \ \ \boldsymbol{\theta} \in \Theta_1. \end{aligned}$$

Power of the test is $1-\beta = P_\theta(\boldsymbol{X} \in R)$, $\boldsymbol{\theta} \in \Theta_1$. Note that $\alpha = \alpha(\boldsymbol{\theta})$ and $\beta = \beta(\boldsymbol{\theta})$ are functions of $\boldsymbol{\theta}$.

Theorem 9.1 (Neyman - Pearson Lemma) *Consider simple alternatives $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ with respective pdfs $f(\boldsymbol{x}, \boldsymbol{\theta}_0)$ and $f(\boldsymbol{x}, \boldsymbol{\theta}_1)$. Then the minimal error rejection region is*

$$R = \{\boldsymbol{x} \in \mathbb{R}^d : f(\boldsymbol{x}, \boldsymbol{\theta}_1) \geq \kappa f(\boldsymbol{x}, \boldsymbol{\theta}_0)\}, \tag{9.1}$$

where $\kappa > 0$ is such that $\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \alpha$.

Proof. Note that $\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \alpha$ and $\int_{R^c} f(\boldsymbol{x}, \boldsymbol{\theta}_1) d\boldsymbol{x} = \beta$. We want to choose region R, or equivalently R^c, such that the probability of type I error equals the significance level α, and the probability of type II error is the smallest possible. For a constant $\kappa > 0$ this can be formulated as minimization of

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

with respect to R (or equivalently with respect to R^c), subject to $\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \alpha$. Since $R = \mathbb{R}^d \setminus R^c$ we have that

$$\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \int_{\mathbb{R}^d} f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} - \int_{R^c} f(\boldsymbol{x}, \theta_0) d\boldsymbol{x}.$$

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

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

or equivalently for

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

Note that for $\kappa = 0$ the rejection region $R = \mathbb{R}^d$ and hence $\alpha = 1$. By increasing κ the refection region shrinks and $\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x}$ continuously decreases and tends to zero. Therefore we can choose κ such that $\int_R f(\boldsymbol{x}, \boldsymbol{\theta}_0) d\boldsymbol{x} = \alpha$. □

Suppose that $T(\boldsymbol{X})$ is a sufficient statistic for $\boldsymbol{\theta}$. By Factorization Theorem (Theorem 8.3), $f(\boldsymbol{x}, \boldsymbol{\theta}_0) = g(T(\boldsymbol{x}), \boldsymbol{\theta}_0) h(\boldsymbol{x})$ and $f(\boldsymbol{x}, \boldsymbol{\theta}_1) = g(T(\boldsymbol{x}), \boldsymbol{\theta}_1) h(\boldsymbol{x})$. Therefore the rejection region (9.1) can be written as

$$R = \{\boldsymbol{x} : g(T(\boldsymbol{x}), \boldsymbol{\theta}_1) \geq \kappa\, g(T(\boldsymbol{x}), \boldsymbol{\theta}_0)\}.$$

9.1 Likelihood Ratio Test

Consider

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

where $\Theta = \Theta_0 \cup \Theta_1$ and $L(\theta) = f(\boldsymbol{x}, \boldsymbol{\theta})$ is the corresponding likelihood function. Note that $\lambda(\boldsymbol{x}) \geq 1$ since $\Theta_0 \subset \Theta$. The rejection region of the Likelihood Ratio Test (LRT) is

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

for some $c > 1$. That is, the H_0 is rejected for large value of the LRT statistic.

If $T(\boldsymbol{X})$ is a sufficient statistic for $\boldsymbol{\theta}$, then by the Factorization Theorem

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

That is, the LRT can be formulated in terms of the sufficient statistic $T(\boldsymbol{X})$. For simple alternatives when $\Theta_0 = \{\boldsymbol{\theta}_0\}$ and $\Theta_1 = \{\boldsymbol{\theta}_1\}$ we have that

$$\lambda(\boldsymbol{x}) = \frac{\max\{L(\boldsymbol{\theta}_0), L(\boldsymbol{\theta}_1)\}}{L(\boldsymbol{\theta}_0)} = \max\left\{1, f(\boldsymbol{x}, \boldsymbol{\theta}_1)/f(\boldsymbol{x}, \boldsymbol{\theta}_0\right\},$$

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 hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against the unrestricted alternative $H_1 : \boldsymbol{\theta} \in \mathbb{R}^k$. We have that

$$2\log\lambda(\boldsymbol{X}) = -2\log L(\boldsymbol{\theta}_0) + 2\sup_{\boldsymbol{\theta}\in\mathbb{R}^k} \log L(\boldsymbol{\theta}).$$

Note that

$$\sup_{\boldsymbol{\theta}\in\mathbb{R}^k} \log L(\boldsymbol{\theta}) = \log L(\hat{\boldsymbol{\theta}}),$$

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

$$S(\boldsymbol{\theta}) := \frac{\partial}{\partial\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\partial}{\partial\boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta}), \tag{9.2}$$

called the *score function*. Note that $S(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ (necessary optimality condition), and $\mathbb{E}_\theta[S(\boldsymbol{\theta})] = \mathbf{0}$ (see equation (8.7)). Now using second order Taylor approximation,

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

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

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

Also first order approximation of the score function:

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

Therefore

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

Assuming H_0, we have that $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to $\mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1})$, and by (8.4) and the LLN,

$$-\frac{1}{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log L(\boldsymbol{\theta}_0) = -\frac{1}{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \sum_{i=1}^{N} \log f(\boldsymbol{X}_i, \boldsymbol{\theta}_0)$$

converges in probability to $I(\boldsymbol{\theta}_0)$. It follows that under H_0, the statistic $2\log\lambda(\boldsymbol{X})$ converges in distribution to the quadratic form $Z'[I(\boldsymbol{\theta}_0)]Z$, where $Z \sim \mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1})$. By theorem 3.1 this implies that $2\log\lambda(\boldsymbol{X})$ converges in distribution to χ^2_k. □

In general $2\log\lambda(\boldsymbol{X})$ converges in distribution to χ^2_{k-q} under H_0, where $\Theta_0 \subset \mathbb{R}^k$ is a smooth manifold of dimension $q = \dim\Theta_0$.

Power of the LRT under local alternatives
Suppose the following so-called parameter drift (local alternatives) for testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta}_{0,N} = \boldsymbol{\theta}_0 + N^{-1/2}\boldsymbol{b}$, where $\boldsymbol{b} \in \mathbb{R}^k$ is a fixed vector. Then

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

Hence under local alternatives

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

can be approximated by the noncentral chi-square distribution $\chi^2_k(\delta)$ with the noncentrality parameter $\delta = \boldsymbol{b}' I(\boldsymbol{\theta}_0)\boldsymbol{b}$ (Theorem 3.3).

9.2 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

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^k}{\arg\max}\, L(\boldsymbol{\theta}) \text{ and } \tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{a}(\boldsymbol{\theta})=\mathbf{0}}{\arg\max}\, L(\boldsymbol{\theta})$$

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

Wald test statistic. Consider testing (linear[4]) equality constraints $H_0 : \boldsymbol{A}\boldsymbol{\theta} = \boldsymbol{c}$ against $H_1 : \boldsymbol{A}\boldsymbol{\theta} \neq \boldsymbol{c}$, where $\boldsymbol{A}$ is $q \times k$ matrix of full row rank q. The Wald test statistic is

$$W := N(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{c})'(\boldsymbol{A}\, I(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{A}')^{-1}(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{c}).$$

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

Suppose that the corresponding regularity conditions hold so that the (unrestricted) ML estimator $\hat{\boldsymbol{\theta}}$ is a consistent estimator of the population value $\boldsymbol{\theta}^*$ of the parameter vector, and $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \rightsquigarrow \mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}^*)^{-1})$. Then under H_0 (i.e., $\boldsymbol{A}\theta^* = \boldsymbol{c}$)

$$\sqrt{N}\boldsymbol{A}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \sqrt{N}(\boldsymbol{A}\hat{\boldsymbol{\theta}} - \boldsymbol{c}) \rightsquigarrow \mathcal{N}(\mathbf{0}, \boldsymbol{A}\, I(\boldsymbol{\theta}^*)^{-1}\boldsymbol{A}').$$

It follows that under H_0 the Wald test statistic converges in distribution to $Z'(\boldsymbol{A}\, I(\boldsymbol{\theta}^*)^{-1}\boldsymbol{A}')^{-1}Z$, where $Z \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A}\, I(\boldsymbol{\theta}^*)^{-1}\boldsymbol{A}')$. Therefore by Theorem 3.1, under H_0 the Wald test statistic converges in distribution to χ^2_q.

Note that the LRT

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

Also using formula (8.4) for the information matrix and $L(\boldsymbol{\theta}^*) = f(\boldsymbol{x}, \boldsymbol{\theta}^*)$, by the LLN we have that $-\frac{1}{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log L(\boldsymbol{\theta}^*) \approx I(\boldsymbol{\theta}^*)$ under H_0. Therefore under H_0, the LRT and Wald test statistics are asymptotically equivalent.

Score function test statistic. Consider testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. The score function test statistic is

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

where $S(\boldsymbol{\theta})$ is the score function (see equation (9.2)). Recall that $\mathbb{E}_\theta[S(\boldsymbol{\theta})] = 0$ and $N^{-1/2}S(\boldsymbol{\theta})$ converges in distribution to $\mathcal{N}(\mathbf{0}, I(\boldsymbol{\theta}))$. It follows that under H_0,

$$N^{-1}S(\boldsymbol{\theta}_0)'I(\boldsymbol{\theta}_0)^{-1}S(\boldsymbol{\theta}_0) \rightsquigarrow \chi^2_k.$$

In general, under H_0,

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

when testing q equality constraints.

10 Multinomial distribution

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

$$\mathsf{Prob}(\boldsymbol{Y} = \boldsymbol{y}) = \frac{N!}{y_1! \times \cdots \times y_k!}\prod_{i=1}^{k} p_i^{y_i},$$

where $p_i > 0$, $i = 1, ..., k$, and $p_1 + ... + p_k = 1$. We denote this as $\boldsymbol{Y} \sim Mult(N, \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, ..., p_k)'$. In particular, for $k = 2$ this becomes binomial distribution $Y \sim B(N, p)$ with $\mathsf{Prob}(Y = y) = \binom{N}{y} p^y(1-p)^{N-y}$, $y = 0, 1, ..., N$.

The log-likelihood function, up to a constant independent of $\boldsymbol{p}$, is $L(\boldsymbol{p}) = \sum_{i=1}^k Y_i \log p_i$. Therefore the ML estimator of $\boldsymbol{p}$ is given by the solution of the problem:

$$\max_{p\geq 0} \sum_{i=1}^{k} Y_i \log p_i \;\text{ subject to } p_1 + ... + p_k = 1.$$

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

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

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

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

On the other hand

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

and $\mathrm{Var}(Y_i) = N(p_i - p_i^2)$, $\mathrm{Var}(Y_j) = N(p_j - p_j^2)$. It follows that $\mathrm{Cov}(Y_i, Y_j) = -Np_ip_j$.

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

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

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

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

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

Moreover

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

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

$$2\log \lambda = 2\sum_{i=1}^k Y_i \log \frac{\hat{p}_i}{p_i^*} \approx \sum_{i=1}^k \frac{(Y_i - Np_i^*)^2}{Y_i} \approx \sum_{i=1}^k \frac{(Y_i - Np_i^*)^2}{Np_i^*},$$

where in the last approximation we used $p_i^* \approx \hat{p}_i = Y_i/N$. 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 see that the LRT statistic $2\sum_{i=1}^k Y_i \log \frac{\hat{p}_i}{p_i^*}$ and Pearson's statistic asymptotically are equivalent under H_0. Pearson's statistic can be viewed as quadratic approximation of the LRT statistic.

We can write Pearson's statistic as

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

where $\hat{\boldsymbol{p}} = (Y_1/N, ..., Y_k/N)'$ and $\boldsymbol{Q} := \mathrm{diag}(1/p_1^*, ..., 1/p_k^*)$. By the CLT, under H_0, $\sqrt{N}(\hat{\boldsymbol{p}} - \boldsymbol{p}^*)$ converges in distribution to normal $\mathcal{N}_k(\boldsymbol{0}, \boldsymbol{C})$. Recall that since $\boldsymbol{1}_k'\hat{\boldsymbol{p}} = 1$ and $\boldsymbol{1}_k'\boldsymbol{p}^* = 1$, the

covariance matrix $\boldsymbol{C}$ has rank $k-1$, and hence is singular. Therefore the normal distribution $\mathcal{N}_k(\boldsymbol{0}, \boldsymbol{C})$ is degenerate.

Consider $\boldsymbol{Z} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{C})$, let us show that $\boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}$ has χ^2_{k-1} distribution. For $\boldsymbol{W} := \boldsymbol{Q}^{1/2}\boldsymbol{Z}$ we have that $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{Z}'\boldsymbol{Q}\boldsymbol{Z}$ and $\boldsymbol{W} \sim \mathcal{N}_k\big(\boldsymbol{0}, \boldsymbol{M}\big)$, where $\boldsymbol{M} := \boldsymbol{I}_k - (\boldsymbol{p}^{*1/2})(\boldsymbol{p}^{*1/2})$. Matrix $\boldsymbol{M}$ is a projection matrix of rank

$$\operatorname{rank}(\boldsymbol{M}) = \operatorname{tr}\big(\boldsymbol{I}_k - (\boldsymbol{p}^{*1/2})(\boldsymbol{p}^{*1/2})'\big) = k - (\boldsymbol{p}^{*1/2})'(\boldsymbol{p}^{*1/2}) = k - \textstyle\sum_{i=1}^k p_i^* = k-1.$$

Since $\boldsymbol{M}$ is a projection matrix of rank $k-1$, it has $k-1$ eigenvalues equal 1 and one eigenvalue 0. Therefore it has the spectral decomposition $\boldsymbol{M} = \boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}'$ with $\boldsymbol{\Lambda} = \operatorname{diag}(1, ..., 1, 0)$. Consider $\boldsymbol{Y} := \boldsymbol{T}'\boldsymbol{W}$. Since matrix $\boldsymbol{T}$ is orthogonal we have that $\boldsymbol{W}'\boldsymbol{W} = \boldsymbol{Y}'\boldsymbol{Y}$. Also $\boldsymbol{Y} \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{T}'\boldsymbol{M}\boldsymbol{T})$. Since the last element of matrix $\boldsymbol{T}'\boldsymbol{M}\boldsymbol{T} = \boldsymbol{\Lambda}$ is zero it follows that $\operatorname{Var}(Y_k) = 0$ and hence $Y_k \equiv 0$. Therefore $\boldsymbol{Y}'\boldsymbol{Y} = Y_1^2 + ... + Y_{k-1}^2 \sim \chi^2_{k-1}$. It follows that under H_0, $N(\hat{\boldsymbol{p}} - \boldsymbol{p}^*)'\boldsymbol{Q}(\hat{\boldsymbol{p}} - \boldsymbol{p}^*)$ converges in distribution to χ^2_{k-1}. □

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

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

Suppose that the model is correct, i.e., there is $\boldsymbol{\theta}^* \in \mathbb{R}^q$ such that $\boldsymbol{p}^* = p(\boldsymbol{\theta}^*)$, where $\boldsymbol{p}^*$ is the true (population) value of the parameter vector. Suppose further that the model is identified at $\boldsymbol{\theta}^*$, i.e., if $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}^*)$, then $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Let $\tilde{\boldsymbol{p}}$ be a consistent estimator of $\boldsymbol{p}^*$. Then asimptotically the ML estimation 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 $\tilde{\boldsymbol{Q}} = \operatorname{diag}(1/\tilde{p}_1, ..., 1/\tilde{p}_k)$.

We have here that $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ converges in distribution to normal $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}(\boldsymbol{\theta}^*)^{-1})$ with $\boldsymbol{I}(\boldsymbol{\theta}) = \boldsymbol{P}(\boldsymbol{\theta})'\boldsymbol{C}(\boldsymbol{\theta})\boldsymbol{P}(\boldsymbol{\theta})$, where $\boldsymbol{P}(\boldsymbol{\theta}) = \partial \log \boldsymbol{p}(\boldsymbol{\theta})/\partial \boldsymbol{\theta}'$ is $m \times q$ matrix and $\boldsymbol{C}(\boldsymbol{\theta}) = \boldsymbol{P}(\boldsymbol{\theta}) - \boldsymbol{p}(\boldsymbol{\theta})\boldsymbol{p}(\boldsymbol{\theta})'$. The LRT for testing $H_0 : \boldsymbol{p} = \boldsymbol{p}(\boldsymbol{\theta})$ against the unrestricted alternative is

$$2 \log \lambda = 2 \sum_{i=1}^k Y_i \log \frac{Y_i/N}{p_i(\tilde{\boldsymbol{\theta}})},$$

where $\tilde{\boldsymbol{\theta}}$ is the MLE under H_0. Under H_0, the LRT statistic $2 \log \lambda$ converges in distribution to χ^2_{k-1-q}, and asymptotically is equivalent to Pearson's statistic.

11 Logistic regression

Let $Y_1, ..., Y_N$ be independent random variables such that Y_i has the binomial distribution $B(m_i, \pi_i)$, $i = 1, ..., N$. Consider the logit model:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik})}, \quad i = 1, ..., N, \tag{11.1}$$

where $\beta_0, ..., \beta_k$ are parameters. That is

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik}, \; i = 1, ..., N,$$

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

We can write this model in the matrix form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}, \tag{11.2}$$

where $\eta_i := \log \frac{\pi_i}{1-\pi_i}$, $i = 1, ..., N$, and $\boldsymbol{X} = [\mathbf{1}_N, \boldsymbol{X}_1, ..., \boldsymbol{X}_k]$ is the design matrix. As in the linear regression we assume that matrix $\boldsymbol{X}$ has full column rank $p = k+1$. The multicollinearity problem can also happen here when columns of the design matrix are 'almost' linearly dependent.

We have that

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{m_i - y_i}, \;\; y_i = 0, 1, ..., m_i.$$

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, $0 \times \log 0 = 0$).

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

$$\boldsymbol{W} = \operatorname{diag}\{m_1\pi_1(1-\pi_1), ..., m_n\pi_n(1-\pi_n)\}.$$

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

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

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

$$\mathbb{E}\big[(Y_i - m_i\pi_i)^2\big] = \operatorname{Var}[Y_i] = m_i\pi_i(1-\pi_i), \;\; i = 1, ..., n.$$

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

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

Consider the log-likelihood function $l(\cdot; \boldsymbol{y})$ as a function of $\boldsymbol{x}$ with $\boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{x})$. We have that the terms $y_i \log \pi_i$ are linear functions of $\boldsymbol{x}$, the terms $(m_i - y_i)\log(1-\pi_i)$ consist of linear terms and terms of the form $-(m_i - y_i)\log(1 + \exp(1 + \boldsymbol{\beta}'\boldsymbol{x}))$. Since the function $\phi(x) = \log(1 + e^x)$ is strictly convex, it follows that $l(\cdot; \boldsymbol{y})$ is strictly concave function of $\boldsymbol{x}$, and hence the ML equations for estimating $\boldsymbol{\beta}$ have *unique* solution $\hat{\boldsymbol{\beta}}$ (recall that the design matrix in (11.2) is assumed to have full column rank).

Consider

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... + \hat{\beta}_k X_{ik})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... + \hat{\beta}_k X_{ik})}, \quad i = 1, ..., n,$$

and the following so-called deviance function,

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

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 $l(\boldsymbol{\pi}; \boldsymbol{y}) = \Big(\sum_{i=1}^{N} y_i\Big)\log\tilde{\pi}$.

12 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 $k \times 1$ vector of parameters and $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{ik})'$ 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

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}, \ i = 1, ..., N. \tag{12.1}$$

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

$$\eta_i = g(\mu_i), \ i = 1, ..., N,$$

and

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

We also can write it in the matrix form

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{X}$ is the design matrix with rows $(1, X_{i1}, ..., X_{ik})$, $i = 1, ..., N$. As before it is assumed that the design matrix has full column rank.

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

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{12.2}$$

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 $\mathcal{N}(\mu, \sigma^2)$ 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 (12.2) 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 (12.2) becomes an exponential family in the canonical form with canonical parameter θ.

Consider

$$l(y; \theta, \phi) = \log f_Y(y, \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi). \tag{12.3}$$

By the standard theory of the ML we have that

$$\mathbb{E}[\partial l/\partial\theta] = 0, \tag{12.4}$$

$$\mathbb{E}[\partial^2 l/\partial\theta^2] = -\mathbb{E}[(\partial l/\partial\theta)^2]. \tag{12.5}$$

Also by (12.3)

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

and because of (12.4), $\mathbb{E}[Y - b'(\theta)] = 0$. Thus $\mathbb{E}[Y] = b'(\theta)$, that is (compare with (7.2))

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

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 (7.3))

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

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 (12.1). In order to compute the ML estimate of $\boldsymbol{\beta}$ we need to maximize the corresponding log-likelihood function (given in (12.3)), that is to solve the problem

$$\max_{\boldsymbol{\beta}}\sum_{i=1}^n Y_i\boldsymbol{x}_i'\boldsymbol{\beta}-b(\boldsymbol{x}_i'\boldsymbol{\beta}). \tag{12.6}$$

When $b(\cdot)$ is a convex function, the above problem (12.6) is convex. For the binomial and Poisson distributions the corresponding functions $b(\cdot)$ are convex.

13 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 (pdfs) 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

$$c_1q_1\int_{R_2}p_1(\boldsymbol{x})d\boldsymbol{x}+c_2q_2\int_{R_1}p_2(\boldsymbol{x})d\boldsymbol{x},$$

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 function, 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

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

If $c_1 q_1 p_1(\boldsymbol{x}) = c_2 q_2 p_2(\boldsymbol{x})$, we can take such points either in R_1 or R_2.

It could be noted that the above derivations basically are the same as derivation of the Neyman - Pearson Lemma in section 9. The only difference is that the misclassification errors are treated here symmetrically, unlike in the hypothesis testing.

13.1 Classification with normally distributed populations

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

$$R_1 = \big\{\boldsymbol{x} : \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq \tfrac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \kappa\big\}. \tag{13.1}$$

Note that if $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, 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 $\sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$ is called *Mahalanobis' distance* between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

13.1.1 An optimization problem

Before proceeding further we need the following result. Consider optimization problem

$$\max_{\boldsymbol{d}\neq\boldsymbol{0}} \frac{\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d}}, \tag{13.2}$$

where $\boldsymbol{A}$ is an $m \times m$ symmetric positive semidefinite matrix and $\boldsymbol{B}$ is an $m \times m$ symmetric positive definite matrix. Let $\boldsymbol{B}^{1/2}$ be symmetric positive definite matrix such that $\boldsymbol{B} = \boldsymbol{B}^{1/2}\boldsymbol{B}^{1/2}$ (see section 1 for discussion of such functions of symmetric matrices). By change of variables $\boldsymbol{h} = \boldsymbol{B}^{1/2}\boldsymbol{d}$ we can write problem (13.2) as

$$\max_{\boldsymbol{h}\neq\boldsymbol{0}} \frac{\boldsymbol{h}'(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2})\boldsymbol{h}}{\|\boldsymbol{h}\|^2},$$

where $\|\cdot\|$ denotes the Euclidean norm. This in turn can be written as

$$\max_{\|\boldsymbol{h}\|=1} \boldsymbol{h}'(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2})\boldsymbol{h}. \tag{13.3}$$

Such problems are discussed in section 15.

Matrix $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}$ is symmetric positive semidefinite. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the eigenvalues and $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$ be the corresponding orthonormal eigenvectors of matrix $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}$, i.e., $\|\boldsymbol{e}_i\| = 1$ and $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$ for $i \neq j$. Note that $\boldsymbol{B}^{-1/2}\boldsymbol{e}_i$ are eigenvectors of matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$ corresponding to the same eigenvalues λ_i. This follows from $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i$ by multiplying both sides of this equation by $\boldsymbol{B}^{-1/2}$. The optimal solution $\bar{\boldsymbol{h}}$ of problem (13.3) is the eigenvector $\bar{\boldsymbol{h}} = \boldsymbol{e}_1$. It follows that solution $\boldsymbol{d}_1 = \boldsymbol{B}^{-1/2}\boldsymbol{e}_1$ of problem (13.2) is given by the eigenvector of $\boldsymbol{B}^{-1}\boldsymbol{A}$ corresponding to its largest eigenvalue. Note that the optimal solution of problem (13.2) is defined up to a scale change, i.e., changing $\boldsymbol{d}$ to $t\boldsymbol{d}$ does not change value of the objective function in (13.2) for any nonzero number t.

In particular, suppose that matrix $\boldsymbol{A}$ has rank one, and hence can be written as $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}'$ where $\boldsymbol{a} \neq \boldsymbol{0}$ is an $m \times 1$ vector. Then $\boldsymbol{d}_1 = \boldsymbol{B}^{-1}\boldsymbol{a}$ is an optimal solution of problem (13.2). Indeed, in that case matrix $\boldsymbol{B}^{-1}\boldsymbol{A}$ has only one nonzero eigenvalue, the largest one. Moreover, for $\lambda = \boldsymbol{a}'\boldsymbol{B}^{-1}\boldsymbol{a}$ we have

$$\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{d}_1 = \boldsymbol{B}^{-1}\boldsymbol{a}(\boldsymbol{a}'\boldsymbol{B}^{-1}\boldsymbol{a}) = \lambda\boldsymbol{B}^{-1}\boldsymbol{a} = \lambda\boldsymbol{d}_1. \tag{13.4}$$

It also follows that in that case the optimal value of problem (13.2) is equal to $\boldsymbol{a}'\boldsymbol{B}^{-1}\boldsymbol{a}$.

Next consider the following problem

$$\max_{\boldsymbol{d}\neq\boldsymbol{0}} \frac{\boldsymbol{d}'\boldsymbol{A}\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{B}\boldsymbol{d}} \text{ subject to } \boldsymbol{d}'\boldsymbol{B}\boldsymbol{d}_1 = 0. \tag{13.5}$$

Again by change of variables $\boldsymbol{h} = \boldsymbol{B}^{1/2}\boldsymbol{d}$ we obtain the problem

$$\max_{\|\boldsymbol{h}\|=1} \boldsymbol{h}'(\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2})\boldsymbol{h}, \text{ subject to } \boldsymbol{h}'\boldsymbol{e}_1 = 0.$$

Optimal solution of this problem is the eigenvector $\boldsymbol{e}_2$ of $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2}$, and hence the optimal solution of problem (13.5) is the eigenvector $\boldsymbol{d}_2 = \boldsymbol{B}^{-1/2}\boldsymbol{e}_2$ of $\boldsymbol{B}^{-1}\boldsymbol{A}$ corresponding to its second largest eigenvalue. Note that $\boldsymbol{d}_2'\boldsymbol{B}\boldsymbol{d}_1 = \boldsymbol{e}_2\boldsymbol{e}_1 = 0$.

13.2 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

$$\max_{\boldsymbol{d}\neq\boldsymbol{0}} \left\{ g(\boldsymbol{d}) := \frac{(\boldsymbol{d}'\boldsymbol{\mu}_1 - \boldsymbol{d}'\boldsymbol{\mu}_2)^2}{\boldsymbol{d}'\boldsymbol{\Sigma}_1\boldsymbol{d} + \boldsymbol{d}'\boldsymbol{\Sigma}_2\boldsymbol{d}} \right\}. \tag{13.6}$$

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 (13.6) is (see equation (13.4))

$$\bar{\boldsymbol{d}} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{13.7}$$

In particular if $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, then $\bar{\boldsymbol{d}} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. Recall that the optimal solution of problem (13.6) is defined up to a scale change.

13.3 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 $q_i > 0$, $i = 1, ..., r$). We have that

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

and

$$\mathbb{E}[\boldsymbol{X}\boldsymbol{X}'] = q_1(\boldsymbol{\Sigma}_1 + \boldsymbol{\mu}_1\boldsymbol{\mu}_1')... + q_r(\boldsymbol{\Sigma}_r + \boldsymbol{\mu}_r\boldsymbol{\mu}_r'),$$

and hence

$$\mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}'] - \boldsymbol{\mu}\boldsymbol{\mu}' = \sum_{i=1}^{r} q_i\boldsymbol{\Sigma}_i + \sum_{i=1}^{r} q_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})' = \boldsymbol{\Omega} + \boldsymbol{M},$$

where

$$\boldsymbol{\Omega} := \sum_{i=1}^{r} q_i\boldsymbol{\Sigma}_i \ \text{ and } \ \boldsymbol{M} := \sum_{i=1}^{r} q_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})'. \tag{13.8}$$

Consider the following optimization problem

$$\max_{\boldsymbol{d}\in\mathbb{R}^m} \left\{ g(\boldsymbol{d}) := \frac{\boldsymbol{d}'\boldsymbol{M}\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{\Omega}\boldsymbol{d}} \right\}. \tag{13.9}$$

Note that matrices $\boldsymbol{\Sigma}_i$ are positive definite and hence matrix $\boldsymbol{\Omega}$ is positive definite, and matrix $\boldsymbol{M}$ is positive semidefinite. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the eigenvalues of $\boldsymbol{\Omega}^{-1}\boldsymbol{M}$. Then the optimal solution $\boldsymbol{d}_1$ of problem (13.9) is the eigenvector of $\boldsymbol{\Omega}^{-1}\boldsymbol{M}$ corresponding to its largest eigenvalue λ_1 (see section 13.1.1). Next maximize $g(\boldsymbol{d})$ subject to $\boldsymbol{d}'\boldsymbol{\Omega}\boldsymbol{d}_1 = 0$. The solution of this problem is given by eigenvector $\boldsymbol{d}_2$ of $\boldsymbol{\Omega}^{-1}\boldsymbol{M}$ corresponding to the second largest eigenvalue λ_2. By continuing this process we obtain discriminant functions $\boldsymbol{d}_i'\boldsymbol{X}$, $i = 1, ..., r-1$. Note that $\mathrm{rank}(\boldsymbol{M}) \leq r - 1$ since

$$\textstyle\sum_{i=1}^{r} q_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}) = \sum_{i=1}^{r} q_i\boldsymbol{\mu}_i - \boldsymbol{\mu} = \boldsymbol{0}.$$

Hence $\lambda_r = ... = \lambda_m = 0$. For $r = 2$ we have $\boldsymbol{\mu}_1 - \boldsymbol{\mu} = q_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$, $\boldsymbol{\mu}_2 - \boldsymbol{\mu} = q_1(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and hence $\boldsymbol{M} = q_1q_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'$. In that case the above approach is the same as Fisher's discriminant analysis.

13.3.1 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 $j \neq i$.

Voronoi diagram. The positive definite matrix $\boldsymbol{\Sigma}^{-1}$ defines the corresponding norm $\|\boldsymbol{x}\|_{\Sigma^{-1}} := \sqrt{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$. If $\boldsymbol{\Sigma} = \boldsymbol{I}_m$ this is the Euclidean norm.

Partition of $\mathbb{R}^m$ into regions

$$R_i = \{\boldsymbol{x} : \|\boldsymbol{x}-\boldsymbol{\mu}_i\|_{\Sigma^{-1}} \leq \|\boldsymbol{x}-\boldsymbol{\mu}_j\|_{\Sigma^{-1}},\ j \neq i\},\ \ i = 1, ..., r,$$

is called Voronoi diagram (with respect to the norm $\|\cdot\|_{\Sigma^{-1}}$). Note that each set R_i is polyhedral given by intersection of half spaces

$$\left\{\boldsymbol{x} : \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) \leq \tfrac{1}{2}\left(\boldsymbol{\mu}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i\right)\right\},\ \ j \neq i.$$

Mahalanobis distance classification: classify $\boldsymbol{X}$ in π_i if $\boldsymbol{X} \in R_i$. For $r = 2$ this is the same classification as in (13.1) with $q_1 = q_2$ and $c_1 = c_2$.

13.4 Bayes and Logistic Regression classifiers

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

$$\log \mathrm{odd}(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}. \tag{13.10}$$

We classify $\boldsymbol{X}$ in π_1 if $\mathrm{odd}(\boldsymbol{x}) > 1$. 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 13.1)

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

In that case (assuming $q_1 = q_2$) equation (13.10) holds with $\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

14 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 $\beta_0 + \boldsymbol{\beta}'\boldsymbol{x} < 0$. Then a point $(y_i, \boldsymbol{x}_i)$ is misclassified iff $y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) < 0$.

The data sets are separable iff there exist β_0 and $\boldsymbol{\beta}$ such that $y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) > 0$ for all $i = 1, ..., n$. The largest margin of separation can be obtained by solving the following problem[5]

$$\max_{\beta_0, \boldsymbol{\beta}, \|\boldsymbol{\beta}\|=1} c \tag{14.1}$$

$$\text{subject to} \quad y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq c, \ i = 1, ..., N. \tag{14.2}$$

The data is separable iff the optimal value of the above problem is positive. By making change of variables $c = 1/\|\boldsymbol{\beta}\|$, we can write the above problem as

$$\min_{\beta_0, \boldsymbol{\beta}} \|\boldsymbol{\beta}\|^2 \tag{14.3}$$

$$\text{subject to} \quad y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq 1, \ i = 1, ..., N. \tag{14.4}$$

Constraints (14.4) define a nonempty feasible set iff the data is separable. Problem (14.3) - (14.4) is a convex quadratic programming problem, and can be solved efficiently.

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

$$y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq c - \xi_i, \ i = 1, ..., N, \tag{14.5}$$

or

$$y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq c(1 - \xi_i), \ i = 1, ..., N, \tag{14.6}$$

where $\xi_i \geq 0$, $i = 1, ..., N$, and $\sum_{i=1}^N \xi_i \leq const.$ Similar to (14.3)–(14.4), formulation (14.6) leads to the following optimization problem

$$\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi}} \|\boldsymbol{\beta}\|^2 \tag{14.7}$$

$$\text{subject to} \quad y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq 1 - \xi_i, \ i = 1, ..., N, \tag{14.8}$$

$$\xi_i \geq 0, \ i = 1, ..., N, \tag{14.9}$$

$$\sum_{i=1}^N \xi_i \leq C, \tag{14.10}$$

where $C > 0$ is a chosen constant. The above problem (14.7) - (14.10) is a convex quadratic programming problem.

Recall that a point $(y_i, \boldsymbol{x}_i)$ is misclassified iff $y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) < 0$. Therefore if $(\beta_0, \boldsymbol{\beta}, \boldsymbol{\xi})$ is a feasible point of the problem (14.7) - (14.10) and a point $\boldsymbol{x}_i$ is misclassified, then $0 > y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) \geq 1 - \xi_i$, and hence $\xi_i > 1$. It follows that if C is smaller than the minimal number of possible misclassifications, then problem (14.7) - (14.10) does not have a feasible solution. On the other hand, for given β_0 and $\boldsymbol{\beta}$ consider the corresponding set of misclassified points. If C is equal to the number of misclassifications, then we can take $\xi_i = 1$ for every misclassified point and $\xi_i = 0$ for every classified point. This will give a feasible point of problem (14.7) - (14.10).

[5]The norm $\|\cdot\|$ here is the Euclidean norm.

We can look at the classification problem from the following point of view. Suppose that we want to find the hyperplane such that the number of misclassified points is minimal. That is, we would like to solve the following problem

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{N} \delta\big(-y_i(\beta_0+\boldsymbol{\beta}'\boldsymbol{x}_i)\big), \tag{14.11}$$

where $\delta(t)=1$ if $t>0$, and $\delta(t)=0$ if $t\le 0$. That is, for given β_0 and $\boldsymbol{\beta}$ the sum $\sum_{i=1}^{N}\delta\big(-y_i(\beta_0+\boldsymbol{\beta}'\boldsymbol{x}_i)\big)$ is equal to the number of misclassified points.

Problem (14.11) is a difficult combinatorial problem. Note that $\delta(t)\le[1+t]_+$, where $[a]_+=\max\{0,a\}$. Therefore we can approximate problem (14.11) by the following *convex* problem

$$\min_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{N}[1-y_i(\beta_0+\boldsymbol{\beta}'\boldsymbol{x}_i)]_+ + c\|\boldsymbol{\beta}\|^2. \tag{14.12}$$

Equivalently we can formulate problem (14.12) as

$$\min_{\beta_0,\boldsymbol{\beta},\boldsymbol{\xi}} \quad \tfrac{1}{2}\|\boldsymbol{\beta}\|^2+\gamma\textstyle\sum_{i=1}^{N}\xi_i \tag{14.13}$$

$$\text{s.t.} \quad y_i(\beta_0+\boldsymbol{\beta}'\boldsymbol{x}_i)\ge 1-\xi_i,\ i=1,...,N, \tag{14.14}$$

$$\xi_i\ge 0,\ i=1,...,N, \tag{14.15}$$

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 (14.13)– (14.15) is the problem

$$\max_{\boldsymbol{\lambda}\ge 0,\boldsymbol{\mu}\ge 0}\ \min_{\beta_0,\boldsymbol{\beta},\boldsymbol{\xi}\ge 0} L(\beta_0,\boldsymbol{\beta},\boldsymbol{\xi},\boldsymbol{\lambda},\boldsymbol{\mu}). \tag{14.16}$$

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 (14.16). That is, by setting derivatives of the Lagrangian to zero, with respect to $\boldsymbol{\beta},\beta_0,\boldsymbol{\xi}$, we have

$$\boldsymbol{\beta}=\sum_{i=1}^{N}\lambda_i y_i\boldsymbol{x}_i \tag{14.17}$$

$$0=\sum_{i=1}^{N}\lambda_i y_i \tag{14.18}$$

$$\lambda_i=\gamma-\mu_i,\ i=1,...,N, \tag{14.19}$$

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

$$\max_{\boldsymbol{\lambda}} \quad \sum_{i=1}^{N}\lambda_i-\tfrac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j\boldsymbol{x}_i'\boldsymbol{x}_j \tag{14.20}$$

$$\text{s.t.} \quad 0\le\lambda_i\le\gamma,\ i=1,...,N, \tag{14.21}$$

$$\sum_{i=1}^{N}\lambda_i y_i=0. \tag{14.22}$$

We also have the following complementarity conditions for problem (14.13)– (14.15):

$$\lambda_i[y_i(\beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_i) - (1 - \xi_i)] = 0,\ i = 1, ..., N, \tag{14.23}$$

$$\mu_i\xi_i = 0,\ i = 1, ..., N. \tag{14.24}$$

Given solution $\bar{\boldsymbol{\lambda}}$ of problem (14.20)–(14.22) the optimal $\boldsymbol{\beta}$ can be computed using equation (14.17), that is

$$\bar{\boldsymbol{\beta}} = \sum_{i=1}^{N} \bar{\lambda}_i y_i \boldsymbol{x}_i. \tag{14.25}$$

The complementarity conditions (14.24) mean that $\xi_i = 0$ if $\mu_i > 0$, and similarly for the complementarity conditions (14.23) . By (14.19) we have that $\mu_i > 0$ if $\lambda_i < \gamma$. Therefore by using equation (14.23), 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 (14.20) is replaced by

$$\sum_{i=1}^{N} \lambda_i - \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \boldsymbol{h}(\boldsymbol{x}_i)'\boldsymbol{h}(\boldsymbol{x}_j). \tag{14.26}$$

Consequently, by using $\boldsymbol{\beta} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{h}(\boldsymbol{x}_i)$ (see (14.25)), the classification is performed according to the sign of

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}'\boldsymbol{h}(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{h}(\boldsymbol{x})'\boldsymbol{h}(\boldsymbol{x}). \tag{14.27}$$

Both expressions (14.26) and (14.27) are defined by the so-called kernel function

$$K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{x})'\boldsymbol{h}(\boldsymbol{z}) = \sum_{s=1}^{q} h_s(\boldsymbol{x}) h_s(\boldsymbol{z}). \tag{14.28}$$

In terms of the kernel function the objective function (14.26) can be written as

$$\sum_{i=1}^{N} \lambda_i - \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{14.29}$$

and the classifier (14.27) as

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{N} \lambda_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i). \tag{14.30}$$

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 $\tanh x = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, $\sinh x = -i\sin(ix) = \frac{e^x - e^{-x}}{2}$, $\cosh(x) = \cos(ix) = \frac{e^x + e^{-x}}{2}$.

15 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[6] that the eigenvectors are orthonormal, i.e., $\|\boldsymbol{e}_i\| = 1$, $i = 1, ..., m$, and $\boldsymbol{e}_i'\boldsymbol{e}_j = 0$ for $i \neq j$. Recall that then (spectral decomposition)

$$\boldsymbol{\Sigma} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}' = \lambda_1\boldsymbol{e}_1\boldsymbol{e}_1' + ... + \lambda_m\boldsymbol{e}_m\boldsymbol{e}_m', \tag{15.1}$$

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

Suppose that we want to find a linear combinations $\boldsymbol{w}'\boldsymbol{X} = w_1X_1 + ... + w_mX_m$ with largest variance. That is we want to solve the problem

$$\max_{\|\boldsymbol{w}\|=1} \mathrm{Var}(\boldsymbol{w}'\boldsymbol{X}). \tag{15.2}$$

Note that $\boldsymbol{w}'\boldsymbol{x} = \|\boldsymbol{w}\|\|\boldsymbol{x}\|\cos\theta$, where θ is the angle between vectors $\boldsymbol{w}$ and $\boldsymbol{x}$. If $\|\boldsymbol{w}\| = 1$, then $\boldsymbol{w}'\boldsymbol{x} = \|\boldsymbol{x}\|\cos\theta$ is the orthogonal projection of vector $\boldsymbol{x}$ onto the straight line in the direction of vector $\boldsymbol{w}$. Therefore problem (15.2) can be viewed as finding a direction such that projection of $\boldsymbol{X}$ onto that direction has the largest variance.

We have that $\mathrm{Var}(\boldsymbol{w}'\boldsymbol{X}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}$ and by (15.1)

$$\boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} = \boldsymbol{w}'\boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}'\boldsymbol{w} = \boldsymbol{v}'\boldsymbol{\Lambda}\boldsymbol{v} = \lambda_1v_1^2 + ... + \lambda_mv_m^2, \tag{15.3}$$

where $\boldsymbol{v} = \boldsymbol{E}'\boldsymbol{w}$. Moreover, since matrix $\boldsymbol{E}$ is orthogonal,

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

That is, $\boldsymbol{v}'\boldsymbol{\Lambda}\boldsymbol{v}$ is a convex combination of eigenvalues λ_i. Thus $\boldsymbol{v}'\boldsymbol{\Lambda}\boldsymbol{v}$ is maximized when $\boldsymbol{v} = (1, 0, ..., 0)'$. Since $\boldsymbol{w} = \boldsymbol{E}\boldsymbol{v}$, it follows that solution of problem (15.2) is given by the eigenvector $\boldsymbol{e}_1$ corresponding to the largest eigenvalue of matrix $\boldsymbol{\Sigma}$. Note that

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

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 $\mathrm{Cov}(Y_1, Y_2) = 0$ and Y_2 has the largest possible variance. Since

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

this means that we want to solve the problem

$$\max_{\|\boldsymbol{w}\|=1} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} \text{ subject to } \boldsymbol{w}'\boldsymbol{e}_1 = 0. \tag{15.4}$$

[6] The norm in this section is the Euclidean norm.

Again we need to find $\boldsymbol{v}$ which maximizes the right hand side of (15.3) and such that the sum of its squared components is one, and is orthogonal to vector $(1, 0, ..., 0)'$, i.e., the first component of $\boldsymbol{v}$ is zero. This is vector $(0, 1, 0, ..., 0)'$, and hence solution of problem (15.4) 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 $\mathrm{Var}(Y_i) = \lambda_i$, $i = 1, ..., m$, $\mathrm{Cov}(Y_i, Y_j) = 0$ for $i \neq j$ and

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

Note also that vector $\boldsymbol{Y}$ of principal components can be written as $\boldsymbol{Y} = \boldsymbol{E}'\boldsymbol{X}$. Multiplying both sides of this equation by $\boldsymbol{E}$, and since matrix $\boldsymbol{E}$ is orthogonal, we obtain

$$\boldsymbol{X} = \boldsymbol{E}\boldsymbol{Y} = Y_1\boldsymbol{e}_1 + ... + Y_m\boldsymbol{e}_m. \tag{15.5}$$

That is, $\boldsymbol{X}$ can be recovered from $\boldsymbol{Y}$ if vectors $\boldsymbol{e}_1, ..., \boldsymbol{e}_m$ are known. This can be used for approximation of $\boldsymbol{X}$ by removing from the right hand side of (15.5) terms corresponding to small eigenvalues λ_i.

Note that principal components analysis is not scale invariant. That is, suppose we rescale components of $\boldsymbol{X}$ say by changing units of measurements. So we change $\boldsymbol{X}$ to $\boldsymbol{D}\boldsymbol{X}$, where $\boldsymbol{D}$ is a diagonal matrix with positive diagonal elements representing change of scale. Then the covariance matrix $\boldsymbol{\Sigma}$ is changed to $\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}$. The eigenvalues and eigenvectors of matrix $\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}$ do not have a simple relation to the respective eigenvalues and eigenvectors of matrix $\boldsymbol{\Sigma}$.

The true (population) covariance matrix $\boldsymbol{\Sigma}$ is unknown. It is estimated by the sample covariance matrix

$$\boldsymbol{S} = (N-1)^{-1} \sum_{i=1}^{N} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'. \tag{15.6}$$

Therefore the PCA usually performed on the sample covariance matrix $\boldsymbol{S}$, or because of the lack of scale invariance, on the sample correlation[7] matrix. 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? In order to apply Delta Theorem we need to compute derivatives of eigenvalues and eigenvectors considered as functions of symmetric matrices. We are going to discus this next.

15.1 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 has *multiplicity one*, i.e., λ_i is different from the previous eigenvalue λ_{i-1} and the next eigenvalue λ_{i+1}. Then $\lambda_i(\cdot)$, considered as a function $\lambda_i : \mathbb{S}^{m\times m} \to \mathbb{R}$, is continuous at $\boldsymbol{A}$. Let us make small perturbations of elements of matrix $\boldsymbol{A}$ by adding the differential $d\boldsymbol{A} \in \mathbb{S}^{m\times m}$. Then the eigenvalue equations for perturbed matrix are

$$(\boldsymbol{A} + d\boldsymbol{A})(\boldsymbol{e}_i + d\boldsymbol{e}_i) = (\lambda_i + d\lambda_i)(\boldsymbol{e}_i + d\boldsymbol{e}_i), \tag{15.7}$$

where $d\lambda_i$ and $d\boldsymbol{e}_i$ are the corresponding small changes in the eigenvalue and eigenvectors. Moreover we have that

$$(\boldsymbol{A} + d\boldsymbol{A})(\boldsymbol{e}_i + d\boldsymbol{e}_i) = \boldsymbol{A}\boldsymbol{e}_i + (d\boldsymbol{A})\boldsymbol{e}_i + \boldsymbol{A}(d\boldsymbol{e}_i) + (d\boldsymbol{A})d\boldsymbol{e}_i. \tag{15.8}$$

[7] The sample correlation matrix is obtained by scaling matrix $\boldsymbol{D}\boldsymbol{S}\boldsymbol{D}$ such that its diagonal elements are equal one.

By disregarding the high order terms $(d\lambda_i)(d\boldsymbol{e}_i)$ and $(d\boldsymbol{A})(d\boldsymbol{e}_i)$ in (15.7) and (15.8), and since $\boldsymbol{A}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i$, we can write

$$(d\boldsymbol{A})\boldsymbol{e}_i + \boldsymbol{A}(d\boldsymbol{e}_i) = (d\lambda_i)\boldsymbol{e}_i + \lambda_i(d\boldsymbol{e}_i). \tag{15.9}$$

Furthermore up to high order terms

$$(\boldsymbol{e}_i + d\boldsymbol{e}_i)'(\boldsymbol{e}_j + d\boldsymbol{e}_j) = (d\boldsymbol{e}_i)'\boldsymbol{e}_j + \boldsymbol{e}_i'(d\boldsymbol{e}_j) + \boldsymbol{e}_i'\boldsymbol{e}_j. \tag{15.10}$$

It follows that for $i = j$, since $(\boldsymbol{e}_i + d\boldsymbol{e}_i)'(\boldsymbol{e}_i + d\boldsymbol{e}_i) = \boldsymbol{e}_i'\boldsymbol{e}_i = 1$,

$$\boldsymbol{e}_i'(d\boldsymbol{e}_i) = 0, \tag{15.11}$$

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

$$(d\boldsymbol{e}_i)'\boldsymbol{e}_j + \boldsymbol{e}_i'(d\boldsymbol{e}_j) = 0. \tag{15.12}$$

Consequently by multiplying both sides of (15.9) 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

$$d\lambda_i = \boldsymbol{e}_i'(d\boldsymbol{A})\boldsymbol{e}_i. \tag{15.13}$$

It is also possible to write (15.13) as

$$d\lambda_i = \mathrm{tr}(\boldsymbol{e}_i\boldsymbol{e}_i'(d\boldsymbol{A})). \tag{15.14}$$

Equation (15.13) (equation (15.14)) gives an expression for the linear approximation of the eigenvalue λ_i for small perturbations $d\boldsymbol{A}$ of matrix $\boldsymbol{A}$, i.e.,

$$\lambda_i(\boldsymbol{A} + d\boldsymbol{A}) = \lambda_i(\boldsymbol{A}) + \boldsymbol{e}_i'(d\boldsymbol{A})\boldsymbol{e}_i + o(\|d\boldsymbol{A}\|). \tag{15.15}$$

The assumption that the eigenvalue is simple is essential in the above derivations.

Now let us compute $d\boldsymbol{e}_i$. Note that since it is assumed that the eigenvalue λ_i is simple and $\|\boldsymbol{e}_i\| = 1$, the eigenvector $\boldsymbol{e}_i$ of $\boldsymbol{A}$ is defined uniquely up to sign change from $\boldsymbol{e}_i$ to $-\boldsymbol{e}_i$. 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 $c_j = \boldsymbol{e}_j'd\boldsymbol{e}_i$, $j = 1, ..., m$. For $i \neq j$ we have by (15.9) and since $\boldsymbol{e}_j'\boldsymbol{e}_i = 0$ that

$$\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i + \boldsymbol{e}_j'\boldsymbol{A}(d\boldsymbol{e}_i) = \lambda_i \boldsymbol{e}_j'(d\boldsymbol{e}_i), \tag{15.16}$$

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

$$\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i = (\lambda_i - \lambda_j)\boldsymbol{e}_j'(d\boldsymbol{e}_i). \tag{15.17}$$

This implies that

$$c_j = (\lambda_i - \lambda_j)^{-1}\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i,\ j \neq i. \tag{15.18}$$

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

$$d\boldsymbol{e}_i = \sum_{\substack{j=1 \\ j\neq i}}^{m} \left[\frac{\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i}{\lambda_i - \lambda_j}\right] \boldsymbol{e}_j. \tag{15.19}$$

That is, for small perturbations $d\boldsymbol{A}$ of matrix $\boldsymbol{A}$,

$$\boldsymbol{e}_i(\boldsymbol{A} + d\boldsymbol{A}) = \sum_{\substack{j=1 \\ j\neq i}}^{m} \left[\frac{\boldsymbol{e}_j'(d\boldsymbol{A})\boldsymbol{e}_i}{\lambda_i - \lambda_j}\right] \boldsymbol{e}_j + o(\|d\boldsymbol{A}\|). \tag{15.20}$$

15.2 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

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

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

Note the following matrix identities

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = (\boldsymbol{AC}) \otimes (\boldsymbol{BD}) \tag{15.21}$$

and

$$\mathsf{vec}(\boldsymbol{BXC}) = (\boldsymbol{C}' \otimes \boldsymbol{B})\mathsf{vec}(\boldsymbol{X}), \tag{15.22}$$

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 $\boldsymbol{b} = \mathsf{vec}(\boldsymbol{B})$,

$$\mathrm{tr}(\boldsymbol{A}'\boldsymbol{B}) = \sum_{i,j} a_{ij}b_{ij} = \boldsymbol{a}'\boldsymbol{b}. \tag{15.23}$$

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid sample of realizations of random vector $\boldsymbol{X} = (X_1, ..., X_m)'$. Assume that the distribution of $\boldsymbol{X}$ has finite fourth order moments. Let $\boldsymbol{s} = \mathsf{vec}(\boldsymbol{S})$ and $\boldsymbol{\sigma}_0 = \mathsf{vec}(\boldsymbol{\Sigma}_0)$, 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 since matrices $\boldsymbol{S}$ and $\boldsymbol{\Sigma}_0$ are symmetric, vectors $\boldsymbol{s}$ and $\boldsymbol{\sigma}_0$ have not more than $m(m+1)/2$ different elements, therefore $\mathrm{rank}(\boldsymbol{\Gamma}) \leq m(m+1)/2$. The typical element of matrix $\boldsymbol{\Gamma}$ is

$$\begin{array}{rcl} [\boldsymbol{\Gamma}]_{ij,k\ell} & = & \mathbb{E}\big\{[(X_i - \mu_i)(X_j - \mu_j) - \sigma_{ij}][(X_k - \mu_k)(X_\ell - \mu_\ell) - \sigma_{k\ell}]\big\} \\ & = & \mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)(X_k - \mu_k)(X_\ell - \mu_\ell)\big] - \sigma_{ij}\sigma_{k\ell}. \end{array}$$

In particular if $\boldsymbol{X}$ has 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

$$[\boldsymbol{\Gamma}]_{ij,k\ell} = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}. \tag{15.24}$$

In a matrix form equations (15.24) can be written as

$$\boldsymbol{\Gamma} = 2\boldsymbol{M}_m(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0), \tag{15.25}$$

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

$$\boldsymbol{\Gamma} = 2\boldsymbol{M}_m(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0)\boldsymbol{M}_m. \tag{15.26}$$

15.3 Asymptotics of PCA

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid sample from $\mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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

$$\sum_{j=1,\, j\neq i}^{m} \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \boldsymbol{e}_j \boldsymbol{e}'_j. \tag{15.27}$$

By the Delta Theorem and (15.15) 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 $\mathcal{N}(0, \sigma^2)$, where σ^2 can be calculated as follows. We have

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

and hence

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

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

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

Similarly, by (15.20)

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

where

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

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

$$\frac{[\mathsf{vec}(\boldsymbol{e}_i\boldsymbol{e}_j')]'\boldsymbol{\Gamma}[\mathsf{vec}(\boldsymbol{e}_i\boldsymbol{e}_k')]}{(\lambda_i-\lambda_j)^2} = \frac{2\mathrm{tr}[(\boldsymbol{e}_i\boldsymbol{e}_j')\boldsymbol{M}_m\boldsymbol{\Sigma}(\boldsymbol{e}_i\boldsymbol{e}_k')\boldsymbol{M}_m\boldsymbol{\Sigma}]}{(\lambda_i-\lambda_j)^2}. \tag{15.28}$$

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 (15.28) is equal to

$$\frac{\mathrm{tr}[((\boldsymbol{e}_i\boldsymbol{e}_j') + (\boldsymbol{e}_j\boldsymbol{e}_i'))\boldsymbol{\Sigma}((\boldsymbol{e}_i\boldsymbol{e}_k') + (\boldsymbol{e}_k\boldsymbol{e}_i'))\boldsymbol{\Sigma}]}{2(\lambda_i-\lambda_j)^2}. \tag{15.29}$$

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 (15.29) 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 (15.27) follows.

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

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

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

15.4 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 $\boldsymbol{D} = \mathrm{diag}(\sigma_1, ..., \sigma_r)$ 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 decomposition of the (symmetric positive semidefinite) $m \times m$ matrix $\boldsymbol{X}\boldsymbol{X}'$. It follows that σ_i^2, $i = 1, ..., r$, are the nonzero eigenvalues of $\boldsymbol{X}\boldsymbol{X}'$. Similarly $\boldsymbol{W}\boldsymbol{D}^2\boldsymbol{W}'$ is the spectral decomposition of the $n \times n$ matrix $\boldsymbol{X}'\boldsymbol{X}$ with the same nonzero eigenvalues σ_i^2, $i = 1, ..., r$.

For $s < r$ consider the (truncated) matrix $\boldsymbol{X}_s = \boldsymbol{V}_s\boldsymbol{D}_s\boldsymbol{W}_s'$, where $\boldsymbol{V}_s = [\boldsymbol{v}_1, ..., \boldsymbol{v}_s]$, $\boldsymbol{W}_s = [\boldsymbol{w}_1, ..., \boldsymbol{w}_s]$ and $\boldsymbol{D}_s = \mathrm{diag}(\sigma_1, ..., \sigma_s)$, i.e.,

$$\boldsymbol{X}_s = \sigma_1\boldsymbol{v}_1\boldsymbol{w}_1' + ... + \sigma_s\boldsymbol{v}_s\boldsymbol{w}_s'.$$

The matrix $\boldsymbol{X}_s$ is the nearest matrix of rank s 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

$$\min_{Z \in \mathbb{R}^{m \times n}} \|\boldsymbol{X} - \boldsymbol{Z}\|_F \text{ subject to } \mathrm{rank}(\boldsymbol{Z}) \leq s$$

is $\bar{\boldsymbol{Z}} = \boldsymbol{X}_s$. Frobenius norm of a matrix $\boldsymbol{A}$ is

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

16 Factor analysis model

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

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\varepsilon}, \tag{16.1}$$

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

It follows then that

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\varepsilon})(\boldsymbol{\Lambda}\boldsymbol{f} + \boldsymbol{\varepsilon})'] = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \tag{16.2}$$

where $\boldsymbol{\Phi} = \mathrm{Cov}(\boldsymbol{f})$ and $\boldsymbol{\Psi} = \mathrm{Cov}(\boldsymbol{\varepsilon})$. Since it is assumed that the errors are uncorrelated, matrix $\boldsymbol{\Psi}$ is diagonal. Matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ are covariance matrices and hence are positive semidefinite. Since matrix $\boldsymbol{\Psi}$ is diagonal, it is positive semidefinite iff all its diagonal elements are nonnegative. Often it is assumed that $\boldsymbol{\Phi} = \boldsymbol{I}_k$, i.e., the factors are standardized. Then the model becomes

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \tag{16.3}$$

Note that $\mathrm{rank}(\boldsymbol{\Sigma} - \boldsymbol{\Psi}) = \mathrm{rank}(\boldsymbol{\Lambda}) \leq k$. Note also that if $\boldsymbol{T}$ is a $k \times k$ orthogonal matrix, then $(\boldsymbol{\Lambda}\boldsymbol{T})(\boldsymbol{\Lambda}\boldsymbol{T})' = \boldsymbol{\Lambda}\boldsymbol{T}\boldsymbol{T}'\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$. Therefore the right hand side of (16.3) is defined up to change of $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda}\boldsymbol{T}$. This can be viewed as rotation of the row vectors of matrix $\boldsymbol{\Lambda}$ by orthogonal matrix $\boldsymbol{T}$.

There is a certain similarity between Factor Analysis (FA) and PCA. Both try to explain covariances between components of the response vector $\boldsymbol{X}$ by a smaller number of factors. But there are also essential differences, FA is a model and, unlike PCA, is scale invariant. That is, if $\boldsymbol{D}$ is a diagonal matrix with positive diagonal elements, then rescaling $\boldsymbol{X}$ to $\boldsymbol{D}\boldsymbol{X}$ results in rescaling $\boldsymbol{\Sigma}$ to $\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}$, $\boldsymbol{\Lambda}$ to $\boldsymbol{D}\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ to $\boldsymbol{D}^2\boldsymbol{\Psi}$. It is possible to develop a statistical inference of the FA model (below).

Given data (sample) $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ of observations (realizations) of $\boldsymbol{X}$, FA is performed on the sample covariance matrix $\boldsymbol{S}$. That is, $\boldsymbol{S}$ is approximated by matrix of the form $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}' + \hat{\boldsymbol{\Psi}}$, where $\hat{\boldsymbol{\Lambda}}$ is an $m \times k$ matrix and $\hat{\boldsymbol{\Psi}}$ is a diagonal matrix with nonnegative diagonal elements. If it is assumed that the sample has normal distribution, i.e., $\boldsymbol{X}_i \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it is possible to proceed to the corresponding statistical inference. Let us show that the ML estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{X}}$ and $\hat{\boldsymbol{\Sigma}} = N^{-1}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$. Note that $\hat{\boldsymbol{\Sigma}} = \frac{N-1}{N}\boldsymbol{S}$.

The likelihood function is

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\big\{-(\boldsymbol{X}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu})/2\big\}.$$

Up to a constant independent of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ we can write logarithm of the likelihood function as

$$
\begin{aligned}
\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\tfrac{1}{2} N \log |\boldsymbol{\Sigma}| - \tfrac{1}{2} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu}) \\
&= -\tfrac{1}{2} N \log |\boldsymbol{\Sigma}| - \tfrac{1}{2} \sum_{i=1}^{N} \mathrm{tr} \left[\boldsymbol{\Sigma}^{-1} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})' \right] \\
&= -\tfrac{1}{2} N \log |\boldsymbol{\Sigma}| - \tfrac{1}{2} \mathrm{tr} \boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})' \right] \\
&= -\tfrac{1}{2} N \log |\boldsymbol{\Sigma}| - \tfrac{1}{2} \mathrm{tr} \left\{ \boldsymbol{\Sigma}^{-1} \left[\sum_{i=1}^{N} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' + N(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \right] \right\} \\
&= -\tfrac{1}{2} N \log |\boldsymbol{\Sigma}| - \tfrac{1}{2} \mathrm{tr} \left[\boldsymbol{\Sigma}^{-1} \boldsymbol{A} \right] - \tfrac{1}{2} N (\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}).
\end{aligned}
$$

where

$$
\boldsymbol{A} := \sum_{i=1}^{N} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})' = (N-1)\boldsymbol{S}.
$$

That is

$$
\log L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\tfrac{1}{2} N \log |\boldsymbol{\Sigma}| - \tfrac{1}{2} (N-1) \mathrm{tr} \left[\boldsymbol{\Sigma}^{-1} \boldsymbol{S} \right] - \tfrac{1}{2} N (\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}). \tag{16.4}
$$

Since $\boldsymbol{\Sigma}^{-1}$ is positive definite, we have that $(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) \geq 0$ and its minimum of zero is attained for $\boldsymbol{\mu} = \bar{\boldsymbol{X}}$. It follows that $\bar{\boldsymbol{X}}$ is the ML estimator of $\boldsymbol{\mu}$. Now in order to find the ML estimator of $\boldsymbol{\Sigma}$ we need to minimize $N \log |\boldsymbol{\Sigma}| + \mathrm{tr} \left[\boldsymbol{\Sigma}^{-1} \boldsymbol{A} \right]$ over positive definite matrices $\boldsymbol{\Sigma}$. Let $\lambda_1, ..., \lambda_m$ be eigenvalues of $\boldsymbol{\Sigma}^{-1} \boldsymbol{A}$ (note that since matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{A}$ are positive definite, matrix $\boldsymbol{\Sigma}^{-1} \boldsymbol{A}$ has positive real valued eigenvalues, see section 1). Then

$$
N \log |\boldsymbol{\Sigma}| + \mathrm{tr} \left[\boldsymbol{\Sigma}^{-1} \boldsymbol{A} \right] = N \log |\boldsymbol{\Sigma} \boldsymbol{A}^{-1}| + \mathrm{tr} \left[\boldsymbol{\Sigma}^{-1} \boldsymbol{A} \right] + N \log |\boldsymbol{A}| = \sum_{i=1}^{m} (\lambda_i - N \log \lambda_i) + N \log |\boldsymbol{A}|.
$$

Note that function $f(\lambda) = \lambda - N \log \lambda$ is convex and has unique minimizer $\lambda = N$. It follows that the minimum is attained when all eigenvalues $\lambda_i = N$, that is $\hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{A} = N \boldsymbol{I}_m$. It follows that $\hat{\boldsymbol{\Sigma}} = N^{-1} \boldsymbol{A} = \frac{N-1}{N} \boldsymbol{S}$. □

Assuming that the sample is from normally distributed population, by (16.4) and since the MLE of $\boldsymbol{\mu}$ is $\bar{\boldsymbol{X}}$, the MLE of parameters $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$ of the FA model are obtained by solving the problem

$$
\min_{\boldsymbol{\Lambda}, \boldsymbol{\Psi} \geq 0} \log \left| \boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi} \right| + \frac{N-1}{N} \mathrm{tr} \left[(\boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} \boldsymbol{S} \right],
$$

where matrix $\boldsymbol{\Psi}$ is diagonal (by writing $\boldsymbol{\Psi} \geq 0$ we mean that diagonal elements of $\boldsymbol{\Psi}$ are nonnegative).

An important question in FA is how many factors should be in the model. The LRT statistic for testing FA model (16.3), with k factors, is

$$
2 \log \lambda = N \min_{\boldsymbol{\Lambda}, \boldsymbol{\Psi} \geq 0} \left\{ \log \left| \boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi} \right| - \log |\hat{\boldsymbol{\Sigma}}| + \mathrm{tr} \left[(\boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi})^{-1} \hat{\boldsymbol{\Sigma}} \right] - m \right\}
$$

where $\hat{\boldsymbol{\Sigma}} = \frac{N-1}{N} \boldsymbol{S}$ is the unrestricted MLE of $\boldsymbol{\Sigma}$. Under H_0 of the FA model with k factors, the statistic $2 \log \lambda$ asymptotically has χ^2_ν distribution with $\nu = m(m+1)/2 - m(k+1) + k(k-1)/2$

degrees of freedom. In calculation of the degrees of freedom, $m(m+1)/2$ is the number of nonduplicated elements of the covariance matrix, $mk+m$ is the number of estimated parameters and the last term $k(k-1)/2$ is the correction because of the possible rotation of the factor loadings matrix by $k \times k$ orthogonal matrix. Consequently H_0 hypothesis of k factors is rejected if the statistic $2\log\lambda$ is larger than critical value of the χ^2_ν distribution.

The above statistical inference is based on the assumption that the population has a normal distribution. In various applications this assumption can be questionable. Also if the sample size n is large, this procedure tends to reject H_0 even if the FA model gives a reasonable approximation of the sample covariance matrix. Various indexes of fit, with questionable justifications, were suggested in the literature trying to resolve the question of 'correct' number of factors.

17 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 $\lambda_s \neq 0$)

$$\boldsymbol{e}_s = \frac{1}{\lambda_s N}\sum_{i=1}^N \alpha_{is}\boldsymbol{Z}_i, \tag{17.1}$$

where $\alpha_{is} = \boldsymbol{Z}_i'\boldsymbol{e}_s$, $s = 1, ..., q$, $i = 1, ..., N$. It follows by (17.1) that

$$\alpha_{is} = \frac{1}{\lambda_s N}\boldsymbol{Z}_i'\Big(\sum_{j=1}^N \alpha_{js}\boldsymbol{Z}_j\Big) = \frac{1}{\lambda_s N}\sum_{j=1}^N \alpha_{js}\boldsymbol{Z}_i'\boldsymbol{Z}_j = \frac{1}{\lambda_s N}\sum_{j=1}^N \alpha_{js}\boldsymbol{h}(\boldsymbol{X}_i)'\boldsymbol{h}(\boldsymbol{X}_j). \tag{17.2}$$

Consider kernel function (compare with (14.28)) $K(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{h}(\boldsymbol{x})'\boldsymbol{h}(\boldsymbol{z})$. In terms of the kernel function equation (17.2) can be written as

$$\sum_{j=1}^N \alpha_{js}K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \lambda_s N\alpha_{is}. \tag{17.3}$$

Consider $N \times N$ matrix $\boldsymbol{K}$ with components $\boldsymbol{K}_{ij} = K(\boldsymbol{X}_i, \boldsymbol{X}_j)$, $i, j = 1, ..., N$. Equation (17.3) can be written as

$$\boldsymbol{K}\boldsymbol{\alpha}_s = \lambda_s N\boldsymbol{\alpha}_s, \ s = 1, ..., q, \tag{17.4}$$

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 (17.4) 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 (17.1) 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 $N^{-1}\sum_{i=1}^N \boldsymbol{h}(X_i) \neq \boldsymbol{0}$ we can make the following correction to the matrix $\boldsymbol{K}$:

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

18 Correlation analysis

18.1 Partial correlation

Let X, Y and Z be random variables. Partial correlation between X and Y given Z, denoted $\mathrm{Corr}(X, Y|Z)$ or $\rho_{XY.Z}$, is defined as the correlation between residuals of X and Y regressed on Z. That is, let us consider regression X on Z. Without loss of generality we can assume that $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z] = 0$. The regression is obtained by solving

$$\min_\beta \mathbb{E}[(X - \beta Z)^2].$$

Solution of this problem is $\beta = \mathrm{Cov}(X, Z)/\mathrm{Var}(Z) = \mathrm{Corr}(X, Z)$. Hence the partial correlation is

$$\mathrm{Corr}(X, Y|Z) = \mathrm{Corr}\left(X - \rho_{XZ}Z, Y - \rho_{YZ}Z\right) = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}},$$

where $\rho_{XZ} = \mathrm{Corr}(X, Z)$ and $\rho_{YZ} = \mathrm{Corr}(Y, Z)$.

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 $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z_1] = ... = \mathbb{E}[Z_n] = 0$. Consider the problem

$$\min_{\boldsymbol{\beta}} \mathbb{E}[(X - \boldsymbol{\beta}'\boldsymbol{Z})^2].$$

Solution of this problem is $\boldsymbol{\beta} = \boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\Sigma}_{ZX}$, where $\boldsymbol{\Sigma}_Z$ is the covariance matrix of random vector $\boldsymbol{Z} = (Z_1, ..., Z_n)$ and $\boldsymbol{\Sigma}_{XZ} = \mathrm{Cov}(X, \boldsymbol{Z})$. Hence

$$\mathrm{Corr}(X, Y|\boldsymbol{Z}) = \mathrm{Corr}(X - \boldsymbol{\Sigma}_{XZ}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{Z}, Y - \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_Z^{-1}\boldsymbol{Z}).$$

18.2 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 $(\boldsymbol{X}', \boldsymbol{Y}')'$, i.e., $\boldsymbol{\Sigma}_{11} = \mathrm{Cov}(\boldsymbol{X})$, $\boldsymbol{\Sigma}_{22} = \mathrm{Cov}(\boldsymbol{Y})$ and $\boldsymbol{\Sigma}_{12} = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})$. 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

$$\max_{\boldsymbol{a},\boldsymbol{b}} \mathrm{Corr}(U, V). \tag{18.1}$$

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 (18.1) becomes

$$\max_{\boldsymbol{a},\boldsymbol{b}} \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}'\boldsymbol{a}}\sqrt{\boldsymbol{b}'\boldsymbol{b}}}. \tag{18.2}$$

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 (18.2) is attained at $\boldsymbol{b} = \boldsymbol{\Sigma}_{21}\boldsymbol{a}$. Hence with respect to $\boldsymbol{a}$ problem (18.2) becomes

$$\max_{\boldsymbol{a}} \left\{ \frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}{\sqrt{\boldsymbol{a}'\boldsymbol{a}}\sqrt{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}} = \sqrt{\frac{\boldsymbol{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\boldsymbol{a}}{\boldsymbol{a}'\boldsymbol{a}}} \right\}. \tag{18.3}$$

Optimal solution $\bar{\boldsymbol{a}}$ of problem (18.3) is given by the eigenvector of matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}$ corresponding to its largest eigenvalue λ_1, and the maximum in (18.1) 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

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\bar{\boldsymbol{a}} = \lambda_1\bar{\boldsymbol{a}}, \tag{18.4}$$

and similarly

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\bar{\boldsymbol{b}} = \lambda_1\bar{\boldsymbol{b}}. \tag{18.5}$$

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

$$\max_{\boldsymbol{c}} \frac{\boldsymbol{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{d}}{\sqrt{\boldsymbol{c}'\boldsymbol{c}}\sqrt{\boldsymbol{d}'\boldsymbol{d}}} \text{ subject to } \boldsymbol{c}'\boldsymbol{c}_1 = 0.$$

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.

19 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: $p(y_i = k) = \pi_k$, $k = 1, ..., K$, and the conditional distributions $f(\boldsymbol{x}_i|y_i = k) \sim \mathcal{N}_m(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are normal. The corresponding log-likelihood function is

$$\ell_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k \phi(\boldsymbol{X}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right),$$

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 $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K)$ is vector of parameters.

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

20 Von Mises statistical functionals

Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid sample of random vectors with probability distribution (probability measure) $\boldsymbol{X}_i \sim F(\cdot)$. 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 random numbers, the empirical cdf $\hat{F}_N(x) = \frac{\#(X_i \le x)}{N}$. That is, if the sample is arranged in the increasing order $X_{(1)} \le \cdots \le X_{(N)}$, then $\hat{F}_N(x) = 0$ for $x < X_{(1)}$, $\hat{F}_N(x) = 1/N$ for $X_{(1)} \le x < X_{(2)}$, $\hat{F}_N(x) = 2/n$ for $X_{(2)} \le x < X_{(3)}$, 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:

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

Its sample estimate

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

- 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[8] it is defined

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

Its sample estimate

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

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

It is known (Glivenko-Cantelli Theorem) that the empirical cdf $\hat{F}_N(x)$ converges w.p.1 to $F(x)$ uniformly in $x \in \mathbb{R}$, that is $\sup_{x\in\mathbb{R}} |\hat{F}_N(x) - F(x)|$ 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 Consider probability distributions F and G. Their convex combination is

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

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 $t = 0$. Let $G = \hat{F}_N$. Then

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

Suppose further that $T'(F, \cdot)$ is linear (as a function of the direction), then it follows by the above that

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

Now we use the following approximation

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

[8] Quantile $F^{-1}(\alpha)$, $\alpha \in (0, 1)$, is defined by the equation $F(x) = \alpha$. Solution of this equation could be not unique or does not exist if the cdf $F(\cdot)$ is discontinuous. Therefore the left side quantile is defined as $\inf\{x : F(x) \geq \alpha\}$, and the right side quantile is defined as $\sup\{x : F(x) \leq \alpha\}$. If the left side and right side quantiles are different from each other, sometimes their average is taken as the corresponding quantile.

where

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

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

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

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

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 $\mathbb{E}_F\left[IC_{T,F}(\boldsymbol{X})\right] = 0$.

By the above

$$N^{1/2}\Big[T(\hat{F}_N) - T(F)\Big] \approx N^{-1/2} \sum_{i=1}^N IC_{T,F}(\boldsymbol{X}_i). \tag{20.1}$$

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

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

This suggests that $N^{1/2}\big[T(\hat{F}_N) - T(F)\big]$ converges in distribution to normal $\mathcal{N}(0, \sigma^2)$ with $\sigma^2 = \mathrm{Var}_F\left[IC_{T,F}(\boldsymbol{X})\right]$.

These derivations of asymptotic normality of $T(\hat{F}_N)$ are somewhat heuristic since the approximation (20.1) is not rigorously justified. Nevertheless it usually gives correct results, which could be proved by ad hoc methods, and is routinely used in applications.

For example, consider the median functional $T(F) = F^{-1}(1/2)$ (here F is the cumulative distribution function). Suppose that the (population) median $\mathsf{m} = F^{-1}(1/2)$ is uniquely defined and the distribution has density $dF(\mathsf{m})/dx = f(\mathsf{m})$ at $x = \mathsf{m}$.

Let us compute the 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

$$-F(T(F_t)) + (1-t)\frac{dF(T(F_t))}{dt} + G(T(F_t)) + t\frac{dG(T(F_t))}{dt} = 0. \tag{20.2}$$

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

$$\frac{dF(T(F_t))}{dt}\Big|_{t=0} = f(\mathsf{m})\frac{dT(F_t))}{dt}\Big|_{t=0}. \tag{20.3}$$

Equation (20.2) (for $t = 0$) together with (20.2) imply that

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

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

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

We obtain that

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

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

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

We obtain that $N^{1/2}\big[T(\hat{F}_N) - T(F)\big]$ converges in distribution to normal $\mathcal{N}\big(0, \frac{1}{4f(\mathsf{m})^2}\big)$. That is the sample median has approximately normal distribution with variance $\frac{1}{4Nf(\mathsf{m})^2}$, provided that the population median m is defined uniquely and the distribution has density $f(\mathsf{m}) = dF(\mathsf{m})/dx$ at $x = \mathsf{m}$.

For example suppose that variables X_i have normal distribution $\mathcal{N}(\mu, \sigma^2)$. In that case the median $\mathsf{m} = \mu$. Asymptotic variance of the sample median is $N^{-1}\sigma^2(\pi/2)$, while variance of $\bar{X}$ is $N^{-1}\sigma^2$. In that case $\bar{X}$ is a better estimator of $\mathsf{m} = \mu$.

However, suppose now that X_i have Laplace distribution with $f(x, \theta) = \frac{1}{2}e^{-|x-\theta|}$, $\theta \in \mathbb{R}$. Then θ is the mean and median of the distribution, and $N^{1/2}(\hat{\theta} - \theta)$ converges in distribution to normal $\mathcal{N}(0, 1)$. We have here that $\mathrm{Var}(X_i) = 2$ and hence variance of $\bar{X}$ is $2N^{-1}$, while the asymptotic variance of the sample median is N^{-1}. It is also interesting to note that the MLE $\hat{\theta}$ is the sample median. Now $\partial \log f(x, \theta)/\partial\theta$ is equal 1 if $\theta < x$ and -1 if $\theta > x$. Thus $\partial^2 \log f(x, \theta)/\partial\theta^2 = 0$ for $\theta \neq x$, and $\partial^2 \log f(x, \theta)/\partial\theta^2$ is not defined for $\theta = x$. Hence formula (8.4) for the information number cannot be applied, i.e., the situation here is not standard.

As another example suppose that Y has has Cauchy distribution, i.e., $Y = V/W$ with independent $V \sim \mathcal{N}(0, 1)$ and $W \sim \mathcal{N}(0, 1)$. Cauchy distribution has pdf $f_Y(y) = \frac{1}{\pi(1+y^2)}$. Therefore in that case asymptotic variance of the sample median is $N^{-1}\pi^2/4$. On the other hand, $\mathbb{E}|Y| = +\infty$ and the average $\bar{X}$ has the same distribution as $\mathsf{m} + Y$ for any sample size N, and will not converge to m as $N \to \infty$.

Finite sample interpretation of the influence curve. By adding one more observation X_{N+1} to sample $X_1, ..., X_N$, 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

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

This shows sensitivity of the estimator to one observation. If $\mathrm{Var}_F[IC_{T,F}(X)]$ is large, the estimator $T(\hat{F}_N)$ can be sensitive just to one observation.

21 Bootstrap

21.1 Jackknife bias estimation.

Consider an estimator $\hat{\theta} = \hat{\theta}(\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$. Denote

$$\hat{\theta}_{-i} = \hat{\theta}(\boldsymbol{X}_1, ..., \boldsymbol{X}_{i-1}, \boldsymbol{X}_{i+1}, ..., \boldsymbol{X}_N), \;\; i = 1, ..., N,$$

i.e., $\hat{\theta}_{-i}$ is obtained by removing data point $\boldsymbol{X}_i$ from calculation of $\hat{\boldsymbol{\theta}}$. Let $\bar{\theta} = N^{-1} \sum_{i=1}^{N} \hat{\theta}_{-i}$.

The Jackknife estimator of the bias is $(N-1)(\bar{\theta} - \hat{\theta})$. Bias corrected version Jackknife estimator

$$\hat{\theta}_{jack} = \hat{\theta} - (N-1)(\bar{\theta} - \hat{\theta}) = N\hat{\theta} - (N-1)(\bar{\theta}). \tag{21.1}$$

Theoretical justification. Suppose that

$$\mathbb{E}_\theta[\hat{\theta}] = \theta + N^{-1} a(\theta),$$

i.e., bias $b_\theta(\hat{\theta}) = N^{-1}a(\theta)$, of $\hat{\theta}$, is of order $O(1/N)$. Then $\mathbb{E}[\hat{\theta}_{-i}] = \theta + (N-1)^{-1}a(\theta)$ and hence

$$\mathbb{E}[\bar{\theta}] = N^{-1} \sum_{i=1}^{n} \mathbb{E}[\hat{\theta}_{-i}] = \theta + (N-1)^{-1} a(\theta),$$

and thus

$$\mathbb{E}[\hat{\theta} - \bar{\theta}] = N^{-1}a(\theta) - (N-1)^{-1}a(\theta) = [N(N-1)]^{-1}a(\theta).$$

It follows that

$$\mathbb{E}[(N-1)(\hat{\theta} - \bar{\theta})] = -N^{-1}a(\theta),$$

and hence $\mathbb{E}[\hat{\theta}_{jack}] = \theta$, i.e. $\hat{\theta}_{jack}$ is an unbiased estimator of θ.

21.2 Bootstrap method

The idea of resampling used in the Jackknife estimation is further extended in the Bootstrap method. Let $\hat{\theta} = \hat{\theta}(\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$ be an estimator which is a function of sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$. 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 $\tilde{\boldsymbol{X}}_1, ..., \tilde{\boldsymbol{X}}_N$ from F and compute $\tilde{\theta} = \hat{\theta}(\tilde{\boldsymbol{X}}_1, ..., \tilde{\boldsymbol{X}}_N)$. We repeat this procedure independently M times. In that way we generate M independent replications $\tilde{\theta}_1, ..., \tilde{\theta}_M$ of the random variable $\hat{\theta}$. Consequently 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 estimate variance of $\hat{\theta}$ as

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

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

Of course the true distribution F is not known. So we replace it by the empirical distribution $\hat{F}_N = N^{-1} \sum_{i=1}^{N} \delta_{X_i}$. Then we proceed by generating a random sample $\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*$ from $\hat{F}_N$ and compute $\hat{\theta}^* = \hat{\theta}(\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*)$. We repeat this procedure M times to obtain values $\hat{\theta}_1^*, ..., \hat{\theta}_M^*$,

which can be used to estimate quantity of interest. Generating a sample $\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*$ from $\hat{F}_N$ means resampling from the data set (sample) $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$. That is, an element $\boldsymbol{X}_i^*$ is chosen at random from the set $\{\boldsymbol{X}_1, ..., \boldsymbol{X}_N\}$. This is repeated N times with replacement, to generate one realization $\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*$. So each element of the generated sample $\boldsymbol{X}_1^*, ..., \boldsymbol{X}_N^*$ coincides with some element of the original sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$.

This procedure is easy to implement and does not require any modelling assumptions. On the other hand, it is solely based on the sample (the data) $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ and can be very sensitive to outliers. Its theoretical analysis is quite sophisticated and is based of theory of statistical estimators of functionals $\theta = T(F)$.

22 Robust statistics

Let $\rho : \mathbb{R} \to \mathbb{R}_+$ be a convex function such that $\rho(0) = 0$. Consider the problem

$$\min_{\theta} \mathbb{E}\big[\rho(X - \theta)\big].$$

For example if $\rho(t) = t^2$ this becomes the least squares problem, its solution is $\theta^* = \mathbb{E}[X]$. Another example $\rho(t) = |t|$. In that case solution θ^* is the median of the distribution of X. The sample median is much less sensitive to outliers than the average. Another example which tries to combine local efficiency of least squares with robustness of absolute value deviations is $\rho(t) = t^2$ for $|t| \leq 2$, and $\rho(t) = |t| + 2$ for $|t| \geq 2$. So when observation in the interval $[-2, 2]$ it works like least squares, outside that interval it could deal with outliers as the absolute deviation method.

As another example for $\alpha \in (0, 1)$ let

$$\rho_\alpha(t) = \begin{cases} -(1-\alpha)t & \text{if} \quad t \leq 0, \\ \alpha t & \text{if} \quad t \geq 0. \end{cases} \tag{22.1}$$

Consider the problem

$$\min_{\theta \in \mathbb{R}} \mathbb{E}\big[\rho_\alpha(X - \theta)\big]. \tag{22.2}$$

We have that

$$\partial \mathbb{E}[\rho_\alpha(X - \theta)]/\partial\theta = \mathbb{E}[\partial \rho_\alpha(X - \theta)/\partial\theta],$$

with $\partial \rho_\alpha(X - \theta)/\partial\theta$ is equal to $(1 - \alpha)$ for $X - \theta < 0$, and $-\alpha$ for $X - \theta > 0$. Suppose that $F(x)$ is continuous at $x = \theta$. It follows that

$$\partial \mathbb{E}[\rho_\alpha(X - \theta)]/\partial\theta = F(\theta) - \alpha,$$

where $F(x) = \mathsf{Prob}(X \leq x)$ is the cdf of X. Thus the quantile $\theta = F^{-1}(\alpha)$ is the optimal solution of problem (22.2). In particular for $\alpha = 1/2$, $\rho_\alpha(t) = \frac{1}{2}|t|$ and solution of problem (22.2) is the median of the distribution. As it was discussed in section 20, the left side $\inf\{x : F(x) \geq \alpha\}$, and the right side $\sup\{x : F(x) \leq \alpha\}$ quantiles can be different from each other. In that case optimal solution of problem (22.2) can be any point between the left side and right side quantiles.

22.1 Quantile regression

Quantile regression use function $\rho_\alpha(\cdot)$, defined in (22.1), to fit linear model to the data. That is, consider the problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \mathbb{E}[\rho_\alpha(Y - \boldsymbol{\beta}'\boldsymbol{X})], \tag{22.3}$$

where Y and $\boldsymbol{X} = (1, X_1, ..., X_k)'$ are random variables. Given data Y_i and $\boldsymbol{X}_i = (1, X_{i1}, ..., X_{ik})'$, $i = 1, ..., N$, the sample counterpart of problem (22.3) is

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \sum_{i=1}^{N} \rho_\alpha(Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i), \tag{22.4}$$

The solution $\hat{\boldsymbol{\beta}}$ of problem (22.4) can be viewed as an estimator of the solution of problem (22.3). For $\alpha = 1/2$ this becomes the least absolute deviations method for solving linear regression. Both problems (22.3) and (22.4) could have more than one optimal solution.

Problem (22.4) can be written as the linear program

$$\begin{array}{ll} \min\limits_{\boldsymbol{\beta}, \boldsymbol{v}^+, \boldsymbol{v}^-} & \sum\limits_{i=1}^{N} (1-\alpha) v_i^- + \alpha v_i^+ \\ \text{s.t.} & Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i = v_i^+ - v_i^-, \ i = 1, ..., N, \\ & v_i^- \geq 0, \ v_i^+ \geq 0, \ i = 1, ..., N. \end{array}$$

It is possible to show that under some regularity conditions, in particular if Y has pdf $f_Y(\cdot)$, $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to normal with zero mean vector and covariance matrix $\alpha(1-\alpha)\boldsymbol{\Psi}^{-1}\boldsymbol{\Omega}\boldsymbol{\Psi}^{-1}$ with $\boldsymbol{\Psi} = \mathbb{E}[f_Y(\boldsymbol{\beta}'\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}']$ and $\boldsymbol{\Omega} = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}']$.

23 Bayes estimators

Recall Bayes' formula: if $\{A_i\}$ is a partition of the sample space and B is an event such that $P(B) \neq 0$ then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}.$$

Let $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_N)$ be a sample with $\boldsymbol{X} \sim f(\boldsymbol{x}, \boldsymbol{\theta})$. Suppose that $\boldsymbol{\theta}$ is random with pdf $\pi(\boldsymbol{\theta})$, referred to as the *prior distribution*. Denote by $f(\boldsymbol{x}|\boldsymbol{\theta})$ the sampling distribution conditional on $\boldsymbol{\theta}$. Then the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{\theta}$ is $f(\boldsymbol{x}, \boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. By Bayes' formula, the distribution of $\boldsymbol{\theta}$, conditional on $\boldsymbol{X} = \boldsymbol{x}$, is

$$\pi(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}, \tag{23.1}$$

that is $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is proportional to $f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, written $\pi(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. The distribution (pdf) $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ is called the *posterior distribution*.

Example 23.1 Suppose that $X_i \sim \mathcal{N}(\theta, \sigma^2)$ and $\theta \sim \mathcal{N}(\mu, \tau^2)$, where σ^2, μ and τ^2 are supposed to be known. We have that

$$f(\boldsymbol{x}|\theta)\pi(\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^N} e^{-\sum_{i=1}^N (x_i - \theta)^2 / 2\sigma^2} \frac{1}{\sqrt{2\pi}\tau} e^{-(\theta-\mu)^2/2\tau^2},$$

and hence

$$f(\boldsymbol{x}|\theta)\pi(\theta) \propto \exp\left\{ \frac{-\left(\theta - \frac{\tau^2 \sum_{i=1}^N x_i + \mu\sigma^2}{N\tau^2 + \sigma^2}\right)^2}{2\left(\frac{\sigma^2\tau^2}{N\tau^2+\sigma^2}\right)} \right\}.$$

It follows that the posterior distribution $\pi(\theta|\boldsymbol{x})$ is normal with conditional mean

$$\mathbb{E}[\theta|\boldsymbol{x}] = \frac{\tau^2}{\tau^2 + \sigma^2/N}\bar{x} + \frac{\sigma^2/N}{\tau^2 + \sigma^2/N}\mu$$

and conditional variance

$$\mathrm{Var}(\theta|\boldsymbol{x}) = \frac{(\sigma^2/N)\tau^2}{\sigma^2/N + \tau^2}.$$

Note that $\mathbb{E}[\theta|\boldsymbol{x}] - \bar{x}$ tends to 0 and $\mathrm{Var}(\theta|\boldsymbol{x})$ tends to 0 as $N \to \infty$. That is, if we view the 'true' distribution of the sample as normal with mean θ^* and variance σ^2, then the average $\bar{X}$ converges in probability to θ^*, and the Bayes estimator converges in probability to θ^*, i.e. for any $\varepsilon > 0$ the probability $\mathsf{Prob}(|\pi(\theta|\boldsymbol{x}) - \theta^*| > \varepsilon)$ converges to 0 w.p.1 as $N \to \infty$. The probability is with respect to the true distribution of X_i and the convergence w.p.1 is with respect to the true distribution.

> This is a general property of Bayes estimators. If we assume that the true distribution of the sample X_i, $i = 1, ...,$ is $f(x, \theta^*)$ for some $\theta^* \in \Theta$, then (under some regularity conditions) Bayes estimator converges in probability to θ^* for almost every (with respect to the true distribution) sequence $X_1, ...$.

In general it may be not easy to compute the posterior distribution. The problem is in calculation of the integral in the right hand side of (23.1). In the above example it was possible to compute the posterior distribution in a closed form, and the posterior distribution was in the same family of normal distributions. Such families of distributions are called *conjugate families*.

Example 23.2 One parameter exponential family

$$f(x|\theta) = \exp\big[\eta(\theta)T(x) - A(\theta)\big]h(x),$$

with prior

$$\pi(\theta) \propto \exp[\alpha\eta(\theta) - \beta A(\theta)].$$

Then posterior distribution

$$f(\theta|x)\pi(\theta) \propto \exp\big[\eta(\theta)(T(x) + \alpha) - (\beta + 1)A(\theta)\big]$$

is in the same family of one parameter exponential distributions. □

23.1 Bayesian decisions

Consider a loss function $L(\boldsymbol{\theta}, a)$ (see definition 8.5) and let $\delta(\boldsymbol{X})$ be a decision rule, e.g., $\delta(\boldsymbol{X})$ is an estimator of parameter $\boldsymbol{\theta}$. The corresponding risk function is

$$R(\boldsymbol{\theta}, \delta) = \mathbb{E}_\theta[L(\boldsymbol{\theta}, \delta(\boldsymbol{X}))] = \int L(\boldsymbol{\theta}, \delta(\boldsymbol{x}))\, f(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{x}.$$

For example, let $L(\theta, a) = (\theta - a)^2$. Then

$$R(\theta, \delta) = \mathbb{E}_\theta[(\theta - \delta(\boldsymbol{X}))^2] = \mathrm{Var}_\theta(\delta(\boldsymbol{X})) + \big(\underbrace{\mathbb{E}_\theta[\delta(\boldsymbol{X})] - \theta}_{\mathrm{bias}(\delta(X))}\big)^2$$

is the Mean Square Error of the estimator $\delta(\boldsymbol{X})$ of θ.

Bayes risk, with prior $\pi(\boldsymbol{\theta})$:

$$B(\pi, \delta) = \mathbb{E}_\pi[R(\boldsymbol{\theta}, \delta)] = \int R(\boldsymbol{\theta}, \delta)\pi(\boldsymbol{\theta})d\theta = \int \Big(\int L(\boldsymbol{\theta}, \delta(\boldsymbol{x}))f(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x}\Big)\pi(\boldsymbol{\theta})d\theta.$$

The Bayes rule with respect to the prior $\pi(\boldsymbol{\theta})$ is

$$\delta^\pi \in \arg\min_{\delta \in \mathcal{D}} B(\pi, \delta), \tag{23.2}$$

where $\mathcal{D}$ is a family of decision rules.

Theorem 23.1 *Define*

$$r(\boldsymbol{x}, a) = \int L(\boldsymbol{\theta}, a)\pi(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta},$$

and let $\delta^\pi(\boldsymbol{x})$ be a minimizer of $r(\boldsymbol{x}, a)$, i.e., $\delta^\pi(\boldsymbol{x}) \in \arg\min_a r(\boldsymbol{x}, a)$. Suppose that $\delta^\pi \in \mathcal{D}$. Then δ^π is the Bayes rule with respect to π.

Proof. Denote $\mathsf{m}(\boldsymbol{x}) := \int f(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{x})$. We can write

$$\begin{aligned} B(\pi, \delta) &= \int \Big(\int L(\boldsymbol{\theta}, \delta(\boldsymbol{x})) f(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} \Big) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \int L(\boldsymbol{\theta}, \delta(\boldsymbol{x})) f(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{x} d\boldsymbol{\theta} \\ &= \int \int L(\boldsymbol{\theta}, \delta(\boldsymbol{x})) \pi(\boldsymbol{\theta}|\boldsymbol{x}) \mathsf{m}(\boldsymbol{x}) d\theta d\boldsymbol{x} \\ &= \int \int r(\boldsymbol{x}, \delta(\boldsymbol{x})) \mathsf{m}(\boldsymbol{x}) d\boldsymbol{x}. \end{aligned}$$

Since $\delta^\pi(\boldsymbol{x}) \in \arg\min_a r(\boldsymbol{x}, a)$, we have that for any $\delta \in \mathcal{D}$,

$$r(\boldsymbol{x}, \delta^\pi(\boldsymbol{x}))\mathsf{m}(\boldsymbol{x}) \leq r(\boldsymbol{x}, \delta(\boldsymbol{x}))\mathsf{m}(\boldsymbol{x}).$$

It follows that δ^π is a minimizer of $B(\pi, \delta)$ over $\delta \in \mathcal{D}$, provided $\delta^\pi \in \mathcal{D}$. □

For squared error loss $L(\theta, a) = (\theta - a)^2$, we have

$$r(\boldsymbol{x}, a) = \int (\theta - a)^2 \pi(\theta|\boldsymbol{x}) d\theta,$$

and hence

$$\delta^\pi(\boldsymbol{x}) = \int \theta\, \pi(\theta|\boldsymbol{x}) d\theta = \mathbb{E}_\pi[\theta|\boldsymbol{X} = \boldsymbol{x}]$$

is the posterior mean. For the absolute error loss $L(\theta, a) = |\theta - a|$, the Bayes rule $\delta^\pi(\boldsymbol{x})$ is the posterior median.

Definition 23.1 *It is said that decision rule δ' is as good as δ if $R(\boldsymbol{\theta}, \delta') \leq R(\boldsymbol{\theta}, \delta)$ for all $\boldsymbol{\theta} \in \Theta$. Moreover, if $R(\boldsymbol{\theta}, \delta') < R(\boldsymbol{\theta}, \delta)$ for some $\boldsymbol{\theta} \in \Theta$, it is said that δ' is better than δ.*

A decision rule $\delta \in \mathcal{D}$ is admissible *if there is no $\delta' \in \mathcal{D}$ that is better than δ.*

The following example shows that an admissible decision rule can be quite awkward.

Example 23.3 Suppose that X has Binomial distribution $Bin(n, \theta)$, i.e., for $x = 0, 1, ..., n$,

$$P(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x},\ \theta \in (0, 1).$$

Let $\delta(x) = c$ for some constant $c \in (0, 1)$ and all $x = 0, ..., n$, and $L(\theta, a) = (\theta - a)^2$ be the loss function. Then

$$R(c, \delta) = \sum_{x=0}^{n} (\delta(x) - c)^2 P(X = x|\theta = c) = 0.$$

Let δ' be as good as δ. Then

$$0 \leq \sum_{x=0}^{n} (\delta'(x) - c)^2 P(X = x|\theta = c) = R(c, \delta') \leq R(c, \delta) = 0.$$

Therefore $\delta'(x) = c$ for all $x = 0, ..., n$, and hence $\delta' = \delta$. It follows that δ is admissible. □

Theorem 23.2 *Suppose that $R(\boldsymbol{\theta}, \delta)$ is continuous in $\boldsymbol{\theta}$ and for every $\boldsymbol{\theta} \in \Theta$ there is $\varepsilon > 0$ such that $\int_{V_{\varepsilon,\theta}} \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0$, where $V_{\varepsilon,\theta} = \{\boldsymbol{\theta}' \in \Theta : \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| \le \varepsilon\}$ is the ε - neighborhood of θ. Then δ^π is an admissible decision rule.*

Proof. We argue by a contradiction. Suppose that $\delta \in \mathcal{D}$ is a decision rule which is better than δ^π. Since

$$\delta^\pi \in \arg\min_{\delta \in \mathcal{D}} B(\pi, \delta) = \arg\min_{\delta \in \mathcal{D}} \int R(\boldsymbol{\theta}, \delta) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

we have that

$$\begin{aligned} 0 &\ge \int R(\boldsymbol{\theta}, \delta^\pi) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} - \int R(\boldsymbol{\theta}, \delta) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \big[R(\boldsymbol{\theta}, \delta^\pi) - R(\boldsymbol{\theta}, \delta)\big] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}. \end{aligned}$$

On the other hand, since δ is better than δ^π,

$$R(\boldsymbol{\theta}, \delta^\pi) - R(\boldsymbol{\theta}, \delta) \ge 0, \ \ \forall \boldsymbol{\theta} \in \Theta,$$

and there is a point $\boldsymbol{\theta}^* \in \Theta$ such that $R(\boldsymbol{\theta}^*, \delta^\pi) - R(\boldsymbol{\theta}^*, \delta) > 0$. Since $R(\boldsymbol{\theta}, \delta)$ is continuous in $\boldsymbol{\theta}$, there is a neighborhood Ξ of $\boldsymbol{\theta}^*$ and $\gamma > 0$ such that $R(\boldsymbol{\theta}, \delta^\pi) - R(\boldsymbol{\theta}, \delta) \ge \gamma$ for all $\boldsymbol{\theta} \in \Xi$. By the assumption of the theorem there is ε - neighborhood V of $\boldsymbol{\theta}^*$ such that $V \subset \Xi$ and $\int_V \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0$. It follows that

$$\int [R(\boldsymbol{\theta}, \delta^\pi) - R(\boldsymbol{\theta}, \delta)] \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \ge \gamma \int_V \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} > 0.$$

This gives the required contradiction. □

Let $T(\boldsymbol{X})$ be a sufficient statistic for $\boldsymbol{\theta}$, and let $L(\boldsymbol{\theta}, a)$ be a loss function. Suppose that $L(\boldsymbol{\theta}, a)$ is convex in a for all $\boldsymbol{\theta}$. Consider $\delta^*(t) = \mathbb{E}[\delta(\boldsymbol{X}) | T = t]$. Note that by sufficiency of T, $\delta^*(t)$ does not depend on $\boldsymbol{\theta}$. Since $L(\boldsymbol{\theta}, a)$ is convex in a, we have by Jensen inequality

$$\mathbb{E}[L(\boldsymbol{\theta}, \delta(\boldsymbol{X}) | T] \ge L(\boldsymbol{\theta}, \mathbb{E}[\delta(\boldsymbol{X}) | T]) = L(\boldsymbol{\theta}, \delta^*(T)).$$

It follows

$$R(\boldsymbol{\theta}, \delta) = \mathbb{E}_\theta\big[\mathbb{E}[L(\boldsymbol{\theta}, \delta(\boldsymbol{X}))|T]\big] \ge \mathbb{E}_\theta[L(\boldsymbol{\theta}, \mathbb{E}[\delta(\boldsymbol{X})|T])] = \mathbb{E}_\theta[L(\boldsymbol{\theta}, \delta^*(T))] = R(\boldsymbol{\theta}, \delta^*).$$

That is, δ^* is as good as δ. Therefore if δ is admissible, then δ^* is also admissible.

Minimax decision rules. Consider

$$\delta' \in \arg\min_{\delta \in \mathcal{D}} \Big\{ \sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \delta) \Big\}.$$

That is, decision rule δ' is minimax if

$$\sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \delta') = \inf_{\delta \in \mathcal{D}} \Big\{ \sup_{\boldsymbol{\theta} \in \Theta} R(\boldsymbol{\theta}, \delta) \Big\}.$$

Theorem 23.3 *Suppose that δ is a* unique *minimax decision rule. Then δ is admissible.*

Proof. Consider $\delta' \in \mathcal{D}$. Then since δ is minimax

$$\sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta') \geq \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta).$$

Moreover since δ is unique we have, that if $\delta' \neq \delta$, then

$$\sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta') > \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta),$$

i.e., δ' is not better than δ. It follows that δ is admissible. □

How minimax decision rules are related to Bayes rules.

Proposition 23.1 *If*

$$\sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta^\pi) \leq B(\pi, \delta^\pi),$$

then δ^π is a minimax decision rule.

Proof. If δ^π is not minimax, then for some δ'

$$\sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta') < \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta^\pi).$$

For any prior $\pi(\boldsymbol{\theta})$ we have

$$B(\pi, \delta) = \int R(\boldsymbol{\theta}, \delta)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \leq \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta) \int \pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta).$$

Hence then

$$B(\pi, \delta') \leq \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta') < \sup_{\boldsymbol{\theta}\in\Theta} R(\boldsymbol{\theta}, \delta^\pi) \leq B(\pi, \delta^\pi),$$

which contradicts minimality of δ^π for $B(\pi, \cdot)$. □

Saddle point

Consider problems

$$\max_{y\in Y} \min_{x\in X} g(x, y), \tag{23.3}$$

$$\min_{x\in X} \max_{y\in Y} g(x, y), \tag{23.4}$$

where X and Y are nonempty sets and $g : X \times Y \to \mathbb{R}$ is a real valued function. We have that for any $(x', y') \in X \times Y$,

$$\psi(y') = \min_{x\in X} g(x, y') \leq g(x', y') \leq \max_{y\in Y} g(x', y) = \varphi(x').$$

It follows that

$$\max_{y\in Y} \psi(y) \leq \min_{x\in X} \varphi(x).$$

Therefore we have that

$$\max_{y\in Y} \min_{x\in X} g(x, y) \leq \min_{x\in X} \max_{y\in Y} g(x, y), \tag{23.5}$$

i.e., optimal value of problem (23.3) is less then or equal to the optimal value of problem (23.4).

Now suppose that $\psi(\bar{y}) = \varphi(\bar{x})$ for some $(\bar{x}, \bar{y}) \in X \times Y$. By (23.5) this implies that optimal values of problems (23.3) and (23.4)are equal to each other and

$$\bar{y} \in \arg\max_{y \in Y} \psi(y) \text{ and } \bar{x} \in \arg\min_{x \in X} \varphi(x).$$

That is

$$\max_{y \in Y} g(\bar{x}, y) = g(\bar{x}, \bar{y}) = \min_{x \in X} g(x, \bar{y}). \tag{23.6}$$

A point $(\bar{x}, \bar{y}) \in X \times Y$ satisfying the above condition (23.6) is called *saddle point*.

Let $(\bar{x}, \bar{y}) \in X \times Y$ be a saddle point. Then

$$\varphi(\bar{x}) = \max_{y \in Y} g(\bar{x}, y) = g(\bar{x}, \bar{y}) = \min_{x \in X} g(x, \bar{y}) = \psi(\bar{y}).$$

- It follows that if a saddle point $(\bar{x}, \bar{y})$ exists, then the optimal values of problems (23.3) and (23.4) are equal to each other, $\bar{y}$ is an optimal solution of problem (23.3) and $\bar{x}$ is an optimal solution of problem (23.4). Conversely if the optimal values of problems (23.3) and (23.4) are equal to each other, and $\bar{y}$ is an optimal solution of problem (23.3) and $\bar{x}$ is an optimal solution of problem (23.4), then $(\bar{x}, \bar{y})$ is a saddle point. □

24 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 $\boldsymbol{X} \sim \mathcal{N}_m(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m)$. 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 $\mathcal{N}_m(\boldsymbol{0}, \sigma_i^2 \boldsymbol{I}_m)$, $i = 1, 2$.

(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 $Z \sim \chi_n^2$ and $\boldsymbol{Y} \sim N_m(\boldsymbol{0}, \boldsymbol{I}_m)$, and Z and $\boldsymbol{Y}$ are independent. This is the multivariate counterpart of t-distribution with n degrees of freedom.

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

$$f(\boldsymbol{x}) = \int_0^\infty (2\pi z)^{-m/2} \exp\left(-\tfrac{1}{2}z^{-1}\boldsymbol{x}'\boldsymbol{x}\right) dG(z).$$

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 $Z \sim n/\chi_n^2$, then $\boldsymbol{X}$ has m-variate t-distribution with n degrees of freedom.

Recall that the characteristic function of a random vector $\boldsymbol{X}$ is $\phi_X(\boldsymbol{t}) := \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{X})]$, where $i^2 = -1$ and $e^{i\theta} = \cos\theta + i\sin\theta$. If $\boldsymbol{X}$ has spherical distribution, then $\boldsymbol{X}$ and $\boldsymbol{T}\boldsymbol{X}$ have the same distribution for any orthogonal matrix $\boldsymbol{T}$ and hence

$$\phi_X(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{X})] = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{T}\boldsymbol{X})] = \mathbb{E}[\exp(i(\boldsymbol{T}'\boldsymbol{t})'\boldsymbol{X})] = \phi_X(\boldsymbol{T}'\boldsymbol{t}).$$

It follows that $\phi_X(\boldsymbol{t})$ is a function of $\boldsymbol{t}'\boldsymbol{t}$, i.e.,

$$\phi_X(\boldsymbol{t}) = \psi(\boldsymbol{t}'\boldsymbol{t}) \tag{24.1}$$

for some function $\psi(\cdot)$ of nonnegative real valued variable. Conversely suppose that the characteristic function of a random vector $\boldsymbol{X}$ can be represented in the form (24.1). Then for any orthogonal matrix $\boldsymbol{T}$ the characteristic function of $\boldsymbol{X}$ is the same as the characteristic function of $\boldsymbol{T}\boldsymbol{X}$ and hence they have the same distribution. It follows that distribution of $\boldsymbol{X}$ has spherical distribution iff the characteristic function of $\boldsymbol{X}$ can be represented in the form (24.1).

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 $\boldsymbol{V} = [v_{ij}]_{i,j=1,\ldots,m}$ 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 $h : \mathbb{R} \to \mathbb{R}_+$. The constant $c_m > 0$ is adjusted in such a way that $\int f(\boldsymbol{x})d\boldsymbol{x} = 1$. We use notation $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ for elliptical distributions. Note that $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ iff $\boldsymbol{Y} = \boldsymbol{V}^{-1/2}(\boldsymbol{X}-\boldsymbol{\mu})$ has spherical distribution.

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, and hence its characteristic function can be written as

$$\phi_X(\boldsymbol{t}) = \mathbb{E}\left[\exp\left(i\boldsymbol{t}'(\boldsymbol{\mu} + \boldsymbol{V}^{1/2}\boldsymbol{Y})\right)\right] = \exp(i\boldsymbol{t}'\boldsymbol{\mu})\mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{V}^{1/2}\boldsymbol{Y})].$$

Since $\boldsymbol{Y}$ has spherical distribution we have by (24.1) 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

$$\phi_X(\boldsymbol{t}) = \exp(i\boldsymbol{t}'\boldsymbol{\mu})\psi(\boldsymbol{t}'\boldsymbol{V}\boldsymbol{t}) \tag{24.2}$$

for some function $\psi(\cdot)$. If $\boldsymbol{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then by equation (2.2),

$$\phi_X(\boldsymbol{t}) = \exp(i\boldsymbol{t}'\boldsymbol{\mu} - \boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}/2). \tag{24.3}$$

In that case $\psi(u) = e^{-u/2}$ for $u \geq 0$ with $\boldsymbol{V} = \boldsymbol{\Sigma}$.

Let $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{A}$ be an $k \times m$ matrix of full row rank k. Then random vector $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$ has characteristic function

$$\phi_Y(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{A}\boldsymbol{X})] = \mathbb{E}[\exp(i(\boldsymbol{A}'\boldsymbol{t})'\boldsymbol{X})] = \phi_X(\boldsymbol{A}'\boldsymbol{t}) = \exp(i\boldsymbol{t}'\boldsymbol{A}\boldsymbol{\mu})\psi(\boldsymbol{t}'\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}'\boldsymbol{t}).$$

It follows that $\boldsymbol{Y} \sim E_k(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{V}\boldsymbol{A}')$. In particular let $\boldsymbol{X}$ 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 $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix}$, where $\boldsymbol{X}_1$ is $m_1 \times 1$ and $\boldsymbol{X}_2$ is $m_2 \times 1$ subvectors of $\boldsymbol{X}$. Then $\boldsymbol{X}_1 \sim E_{m_1}(\boldsymbol{\mu}_1, \boldsymbol{V}_{11})$ with the characteristic function

$$\phi_{X_1}(\boldsymbol{t}_1) = \exp(i\boldsymbol{t}_1'\boldsymbol{\mu}_1)\psi(\boldsymbol{t}_1'\boldsymbol{V}_{11}\boldsymbol{t}_1), \tag{24.4}$$

and similarly for $\boldsymbol{X}_2$.

Now suppose that components of random vector $\boldsymbol{X} = (X_1, ..., X_m)$ have finite second order moments. Then

$$\partial \phi_X(\mathbf{0})/\partial \boldsymbol{t} = i\,\mathbb{E}[\boldsymbol{X}] \tag{24.5}$$

and

$$\partial^2 \phi_X(\mathbf{0})/\partial \boldsymbol{t}\partial \boldsymbol{t}' = -\mathbb{E}[\boldsymbol{X}\boldsymbol{X}'] = -\boldsymbol{\mu}\boldsymbol{\mu}' - \mathrm{Cov}(\boldsymbol{X}). \tag{24.6}$$

It follows from (24.2) together with (24.5) and (24.6), 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

$$\mathrm{Corr}(X_i, X_j) = \frac{v_{ij}}{\sqrt{v_{ii} v_{jj}}},\ i, j = 1, ..., m. \tag{24.7}$$

By (24.7) we have that if $\boldsymbol{X} \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$ and $\boldsymbol{V} = \mathrm{diag}(v_{11}, ..., v_{mm})$ is diagonal, then $X_1, ..., X_m$ are uncorrelated.

Theorem 24.1 *Let $\boldsymbol{X} = (X_1, ..., X_m)' \sim E_m(\boldsymbol{\mu}, \boldsymbol{V})$, $m \geq 2$, 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. By replacing $\boldsymbol{X}$ with $\boldsymbol{X} - \boldsymbol{\mu}$, we can assume without loss of generality that $\boldsymbol{\mu} = \mathbf{0}$. Since matrix $\boldsymbol{V} = \mathrm{diag}(v_{11}, ..., v_{mm})$ is diagonal we have by (24.2) that the characteristic function of $\boldsymbol{X}$ is

$$\phi_X(\boldsymbol{t}) = \psi(\boldsymbol{t}'\boldsymbol{V}\boldsymbol{t}) = \psi\left(\textstyle\sum_{i=1}^m t_i^2 v_{ii}\right).$$

Since $X_1, ..., X_m$ are independent, we have that

$$\phi_X(\boldsymbol{t}) = \mathbb{E}[\exp(i\boldsymbol{t}'\boldsymbol{X})] = \mathbb{E}\left[\textstyle\prod_{i=1}^m e^{it_i X_i}\right] = \textstyle\prod_{i=1}^m \mathbb{E}\left[e^{it_i X_i}\right] = \prod_{i=1}^m \phi_i(t_i),$$

where $\phi_i(t_i)$ is the characteristic function of X_i, $i = 1, ..., m$. By (24.4), $\phi_i(t_i) = \psi(t_i^2 v_{ii})$, and thus it follows that

$$\psi\left(\textstyle\sum_{i=1}^m u_i^2\right) = \prod_{i=1}^m \psi(u_i^2), \tag{24.8}$$

where $u_i := t_i v_{ii}^{1/2}$.

In turn equation (24.8) implies that $\psi(u) = e^{-\kappa u/2}$ for some κ and $u > 0$. Indeed suppose that equation (24.8) holds. Then for any natural number p, $\psi(1) = \psi(1/p + ... + 1/p) = \psi(1/p)^p$ and hence $\psi(1/p) = \psi(1)^{1/p}$. Furthermore for a rational positive number q/p we have $\psi(q/p) = \psi(1/p + ... + 1/p) = \psi(1/p)^q$. It follows that $\psi(q/p) = \psi(1)^{q/p}$ for any positive rational number p/q. Moreover since function $\psi(\cdot)$ is continuous, it follows that for $u > 0$, $\psi(u) = \psi(1)^u = e^{-\kappa u/2}$ for $\kappa := -2\log\psi(1)$. Since $\mathrm{Cov}(\boldsymbol{X}) = \alpha \boldsymbol{V}$, where $\alpha = -2\psi'(0)$, we have then that $\mathrm{Cov}(\boldsymbol{X}) = \kappa \boldsymbol{V}$, and hence $\kappa > 0$. It follows by (24.2) that the characteristic function of $\boldsymbol{X}$ is $\phi_X(\boldsymbol{t}) = \exp(-k\boldsymbol{t}'\boldsymbol{V}\boldsymbol{t}/2)$. That is, the characteristic function of $\boldsymbol{X}$ coincides with the characteristic function of normal distribution with mean vector $\boldsymbol{\mu} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = k\boldsymbol{V}$. □

24.1 Multivariate cumulants

Consider a random variable X. Let

$$\log \mathbb{E}[e^{tX}] = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} \tag{24.9}$$

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 for $t \neq 0$, it is preferable to define cumulants in terms of the characteristic function as

$$\log \mathbb{E}[e^{itX}] = \sum_{n=1}^{\infty} \kappa_n \frac{(it)^n}{n!}, \tag{24.10}$$

where $\kappa_n = \frac{\partial^n \log \mathbb{E}[e^{itX}]}{\partial t^n}\big|_{t=0}$.

Denote $\mu_k := \mathbb{E}[X^k]$ the k-th moment of X. Then

$$\begin{aligned}
\kappa_1 &= \mu_1 = \mathbb{E}[X], \\
\kappa_2 &= \mu_2 - \mu_1^2 = \mathrm{Var}(X), \\
\kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3, \\
\kappa_4 &= \mu_4 - 4\mu_1\mu_3 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4,
\end{aligned}$$

provided these moments are finite. 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. In particular, if $Y = a$ where a is (deterministic) number, then the first cumulant of $X + a$ is $\kappa_1 + a$, and the cumulants of the higher order are the same as the cumulants of X.

Skewness of X is defined as

$$\gamma_1 := \frac{\kappa_3}{\kappa_2^{3/2}}, \tag{24.11}$$

kurtosis of X is defined as

$$\gamma_2 := \frac{\kappa_4}{\kappa_2^2}. \tag{24.12}$$

As it was pointed above, the skewness and kurtosis of X are the same as the respective skewness and kurtosis of $X + a$ for any number a. If distribution of X is symmetrical around its mean, then $\gamma_1 = 0$. If $X \sim N(0, \sigma^2)$, then $\mu_4 = 3\mu_2^2$ (see equation (2.4)). It follows that if $X \sim N(\mu, \sigma^2)$ then its kurtosis $\gamma_2 = 0$.

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

$$\kappa_{1111}^{ijk\ell} = \kappa(\sigma_{ij}\sigma_{k\ell} + \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}).$$

Let $\boldsymbol{S}$ be the sample covariance matrix of sample of size N. 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 $\boldsymbol{\Gamma}_{\mathcal{N}}$ the corresponding $m^2 \times m^2$ covariance matrix (see section 15.2), where the subscript $\mathcal{N}$ emphasizes that this is under the assumption of normal distribution. 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

$$\boldsymbol{\Gamma} = (1 + \kappa)\boldsymbol{\Gamma}_{\mathcal{N}} + \kappa\boldsymbol{\sigma}\boldsymbol{\sigma}'.$$

25 Wishart distribution

Recall that

$$\boldsymbol{S} = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'$$

is the sample covariance matrix of random sample $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$. Note that if $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ is an iid sample from normal distribution $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent. Indeed

$$\mathrm{Cov}(\bar{\boldsymbol{X}}, \boldsymbol{X}_i - \bar{\boldsymbol{X}}) = \mathrm{Cov}(\bar{\boldsymbol{X}}, \boldsymbol{X}_i) - \mathrm{Cov}(\bar{\boldsymbol{X}}).$$

Now $\mathrm{Cov}(\bar{\boldsymbol{X}}, \boldsymbol{X}_i) = N^{-1}\boldsymbol{\Sigma}$ and $\mathrm{Cov}(\bar{\boldsymbol{X}}) = N^{-2}\sum_{i=1}^{N}\mathrm{Cov}(\boldsymbol{X}_i) = N^{-1}\boldsymbol{\Sigma}$. It follows that $\mathrm{Cov}(\bar{\boldsymbol{X}}, \boldsymbol{X}_i - \bar{\boldsymbol{X}}) = \boldsymbol{0}$. That is, $\bar{\boldsymbol{X}}$ and $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$ are uncorrelated, and because their joint distribution is normal, are independent. Since $\boldsymbol{S}$ is a function of $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$, $i = 1, ..., N$, it follows that $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent.

Let $\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n$ be an iid sequence of random vectors having normal distribution $\mathcal{N}_m(\boldsymbol{0}, \boldsymbol{\Sigma})$. Consider random matrix

$$\boldsymbol{A} = \boldsymbol{Z}_1\boldsymbol{Z}_1' + ... + \boldsymbol{Z}_n\boldsymbol{Z}_n'. \tag{25.1}$$

By definition $\boldsymbol{A}$ has Wishart distribution, denoted $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$. In particular, for $m = 1$ and $Z_i \sim \mathcal{N}(0, \sigma^2)$, the corresponding A/σ^2 has chi-square distribution with n degrees of freedom.

Wishart distribution has the following properties.

(i) If $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$ and $\alpha > 0$, then $\alpha\boldsymbol{A} \sim W_m(n, \alpha\boldsymbol{\Sigma})$. Indeed,

$$\alpha\boldsymbol{A} = (\alpha^{1/2}\boldsymbol{Z}_1)(\alpha^{1/2}\boldsymbol{Z}_1)' + ... + (\alpha^{1/2}\boldsymbol{Z}_n)(\alpha^{1/2}\boldsymbol{Z}_n)',$$

and $\alpha^{1/2}\boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{0}, \alpha\boldsymbol{\Sigma})$.

(ii) If $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$ and $\boldsymbol{B}$ is $m \times k$ deterministic matrix, then $\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B} \sim W_k(n, \boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})$. Indeed,

$$\boldsymbol{B}'\boldsymbol{A}\boldsymbol{B} = (\boldsymbol{B}'\boldsymbol{Z}_1)(\boldsymbol{B}'\boldsymbol{Z}_1)' + ... + (\boldsymbol{B}'\boldsymbol{Z}_n)(\boldsymbol{B}'\boldsymbol{Z}_n)',$$

and $\boldsymbol{B}'\boldsymbol{Z}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}'\boldsymbol{\Sigma}\boldsymbol{B})$.

(iii) Equation (25.1) can be written in the following form $\boldsymbol{A} = \boldsymbol{Z}'\boldsymbol{Z}$, where $\boldsymbol{Z}$ is $n \times m$ matrix $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1' \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{Z}_n' \end{bmatrix}$ with $\boldsymbol{Z}' = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n]$. Note that $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{0}$ and the covariance matrix of the corresponding $mn \times 1$ vector $\mathsf{vec}(\boldsymbol{Z}')$ is [9]

$$\mathrm{Cov}(\mathsf{vec}(\boldsymbol{Z}')) = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}. \tag{25.2}$$

Proposition 25.1 *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N \overset{iid}{\sim} \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{S}$ be the sample covariance matrix. Then $\boldsymbol{S} \sim W_m(n, n^{-1}\boldsymbol{\Sigma})$, where $n = N - 1$.*

Proof. Consider $N \times m$ matrix $\boldsymbol{W}$ with rows $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$, i.e., $\boldsymbol{W}' = [\boldsymbol{X}_1 - \bar{\boldsymbol{X}}, ..., \boldsymbol{X}_N - \bar{\boldsymbol{X}}]$. Note that $\boldsymbol{S} = n^{-1}\boldsymbol{W}'\boldsymbol{W}$ and $\boldsymbol{W} = (\boldsymbol{I}_N - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N')\boldsymbol{X}$, where $\boldsymbol{X}' = [\boldsymbol{X}_1, ..., \boldsymbol{X}_N]$. Matrix $\boldsymbol{I}_N - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N'$ is a symmetric projection matrix of rank $N-1$. Hence $\boldsymbol{I}_N - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N' = \boldsymbol{H}\boldsymbol{H}'$, where $\boldsymbol{H}$ is $N \times n$ matrix with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$ and $\boldsymbol{H}'\boldsymbol{1}_N = \boldsymbol{0}$ (spectral decomposition). Consider the following $n \times m$ matrix $\boldsymbol{Z} = \boldsymbol{H}'\boldsymbol{X}$. Then

$$\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{X}'\boldsymbol{H}\boldsymbol{H}'\boldsymbol{X} = \boldsymbol{X}'(\boldsymbol{I}_N - N^{-1}\boldsymbol{1}_N\boldsymbol{1}_N')\boldsymbol{X} = \boldsymbol{W}'\boldsymbol{W},$$

and hence $\boldsymbol{S} = n^{-1}\boldsymbol{Z}'\boldsymbol{Z}$. Note that $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{1}_N\boldsymbol{\mu}'$ and hence $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{H}'\mathbb{E}[\boldsymbol{X}] = \boldsymbol{H}'\boldsymbol{1}_N\boldsymbol{\mu}' = \boldsymbol{0}$. Now $\mathrm{Cov}(\mathsf{vec}(\boldsymbol{X}')) = \boldsymbol{I}_N \otimes \boldsymbol{\Sigma}$ and (see (15.22))

$$\mathsf{vec}(\boldsymbol{X}'\boldsymbol{H}) = (\boldsymbol{H}' \otimes \boldsymbol{I}_m)\mathsf{vec}(\boldsymbol{X}').$$

Thus using (15.21),

$$\mathrm{Cov}(\mathsf{vec}(\boldsymbol{Z}')) = \mathrm{Cov}(\mathsf{vec}(\boldsymbol{X}'\boldsymbol{H})) = (\boldsymbol{H}' \otimes \boldsymbol{I}_m)(\boldsymbol{I}_N \otimes \boldsymbol{\Sigma})(\boldsymbol{H} \otimes \boldsymbol{I}_m) = (\boldsymbol{H}'\boldsymbol{H}) \otimes \boldsymbol{\Sigma} = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}.$$

Hence $\boldsymbol{Z}'\boldsymbol{Z} \sim W_m(n, \boldsymbol{\Sigma})$, and $\boldsymbol{S} \sim W_m(n, n^{-1}\boldsymbol{\Sigma})$. □

Theorem 25.1 *If $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$ and $\boldsymbol{Y}$ is an $m \times 1$ random vector independent of $\boldsymbol{A}$ and such that $\mathsf{Prob}(\boldsymbol{Y} = \boldsymbol{0}) = 0$, then random variable $\frac{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{\Sigma}\boldsymbol{Y}} \sim \chi^2_n$ and is independent of $\boldsymbol{Y}$.*

Proof. Conditional on $\boldsymbol{Y}$, we have that $\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y} \sim W_1(n, \boldsymbol{Y}'\boldsymbol{\Sigma}\boldsymbol{Y})$ and hence

$$\frac{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{\Sigma}\boldsymbol{Y}} \sim W_1(n, 1) = \chi^2_n.$$

That is, the conditional distribution of $\frac{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{\Sigma}\boldsymbol{Y}}$ does not depend on $\boldsymbol{Y}$. It follows that $\frac{\boldsymbol{Y}'\boldsymbol{A}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{\Sigma}\boldsymbol{Y}}$ is independent of $\boldsymbol{Y}$ and its (unconditional) distribution is χ^2_n. □

Together with Proposition 25.1 this implies the following.

Proposition 25.2 *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N \overset{iid}{\sim} \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{S}$ be the corresponding sample covariance matrix. Then $\frac{n\bar{\boldsymbol{X}}'\boldsymbol{S}\bar{\boldsymbol{X}}}{\bar{\boldsymbol{X}}'\boldsymbol{\Sigma}\bar{\boldsymbol{X}}} \sim \chi^2_n$ and is independent of $\bar{\boldsymbol{X}}$ (recall that $n = N - 1$).*

Proof. Since $\boldsymbol{S}$ and $\bar{\boldsymbol{X}}$ are independent and $n\boldsymbol{S} \sim W_m(n, \boldsymbol{\Sigma})$, the result follows from Theorem 25.1. □

[9] Recall definitions of Kronecker product of matrices and vec operator discussed in section 15.2.

Theorem 25.2 *Let* $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$ *be partitioned* $\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}$, *where* $\boldsymbol{A}_{11}$ *is of order* $k \times k$ *and* $\boldsymbol{A}_{22}$ *is of order* $(m-k) \times (m-k)$, *and matrix* $\boldsymbol{\Sigma}$ *is partitioned accordingly* $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$. *Consider* $\boldsymbol{A}_{11.2} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ *and* $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. *Then*

$$\boldsymbol{A}_{11.2} \sim W_k(n-m+k, \boldsymbol{\Sigma}_{11.2}), \tag{25.3}$$

and $\boldsymbol{A}_{11.2}$ *is independent of* $\boldsymbol{A}_{22}$.

Proof. Since $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$ it can be written in the form (25.1), or equivalently as $\boldsymbol{A} = \boldsymbol{Z}'\boldsymbol{Z}$, where $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1' \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{z}_n' \end{bmatrix}$ is the respective $n \times m$ matrix. Let us partition $\boldsymbol{Z} = [\tilde{\boldsymbol{Z}}_1, \tilde{\boldsymbol{Z}}_2]$, where $\tilde{\boldsymbol{Z}}_1$ is of order $n \times k$ and $\tilde{\boldsymbol{Z}}_2$ is of order $n \times (m-k)$. Note that $\tilde{\boldsymbol{Z}}_1 = \begin{bmatrix} \boldsymbol{z}_{11}' \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{z}_{1n}' \end{bmatrix}$ and $\tilde{\boldsymbol{Z}}_2 = \begin{bmatrix} \boldsymbol{z}_{21}' \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{z}_{2n}' \end{bmatrix}$, where $\boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{z}_{1i} \\ \boldsymbol{z}_{2i} \end{bmatrix}$ are respective partitions of vectors $\boldsymbol{Z}_i$, $i = 1, ..., n$. Recall that conditional on $\boldsymbol{Z}_{2i} = \boldsymbol{z}_2$,

$$\boldsymbol{Z}_{1i} \sim \mathcal{N}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{z}_2, \boldsymbol{\Sigma}_{11.2}) \tag{25.4}$$

(see equation (2.3)). Note that matrix $\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_2(\tilde{\boldsymbol{Z}}_2'\tilde{\boldsymbol{Z}}_2)^{-1}\tilde{\boldsymbol{Z}}_2'$ is idempotent (projection) of rank $n - (m-k) = n - m + k$, and

$$[\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_2(\tilde{\boldsymbol{Z}}_2'\tilde{\boldsymbol{Z}}_2)^{-1}\tilde{\boldsymbol{Z}}_2']\tilde{\boldsymbol{Z}}_2 = \boldsymbol{0}. \tag{25.5}$$

Because of (25.4) and (25.5) we have that conditional on $\tilde{\boldsymbol{Z}}_2$,

$$\tilde{\boldsymbol{Z}}_1'[\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_2(\tilde{\boldsymbol{Z}}_2'\tilde{\boldsymbol{Z}}_2)^{-1}\tilde{\boldsymbol{Z}}_2']\tilde{\boldsymbol{Z}}_1 \sim W_k(n-m+k, \boldsymbol{\Sigma}_{11.2}).$$

Moreover

$$\tilde{\boldsymbol{Z}}_1'[\boldsymbol{I}_n - \tilde{\boldsymbol{Z}}_2(\tilde{\boldsymbol{Z}}_2'\tilde{\boldsymbol{Z}}_2)^{-1}\tilde{\boldsymbol{Z}}_2']\tilde{\boldsymbol{Z}}_1 = \underbrace{\tilde{\boldsymbol{Z}}_1'\tilde{\boldsymbol{Z}}_1}_{\boldsymbol{A}_{11}} - \underbrace{\tilde{\boldsymbol{Z}}_1'\tilde{\boldsymbol{Z}}_2}_{\boldsymbol{A}_{12}}\underbrace{(\tilde{\boldsymbol{Z}}_2'\tilde{\boldsymbol{Z}}_2)^{-1}}_{\boldsymbol{A}_{22}^{-1}}\underbrace{\tilde{\boldsymbol{Z}}_2'\tilde{\boldsymbol{Z}}_1}_{\boldsymbol{A}_{21}} = \boldsymbol{A}_{11.2}.$$

It follows that the (unconditional) distribution of $\boldsymbol{A}_{11.2}$ is $W_k(n-m+k, \boldsymbol{\Sigma}_{11.2})$, and that $\boldsymbol{A}_{11.2}$ is independent of $\tilde{\boldsymbol{Z}}_2$ and hence of $\boldsymbol{A}_{22}$. □

Theorem 25.3 *Let* $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$ *and* $\boldsymbol{B}$ *be (deterministic)* $m \times k$ *matrix of rank* k. *Then* $(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \sim W_k(n-m+k, (\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B})^{-1})$.

Proof. Note that the assertion is invariant under linear transformations. That is, if $\boldsymbol{C}$ is an $m \times m$ nonsingular matrix, then by replacing $\boldsymbol{B}$ with $\tilde{\boldsymbol{B}} = \boldsymbol{C}\boldsymbol{B}$ and $\boldsymbol{A}$ with $\tilde{\boldsymbol{A}} = \boldsymbol{C}\boldsymbol{A}\boldsymbol{C}'$ we have $\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}} = \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B}$. Moreover $\tilde{\boldsymbol{A}} \sim W_m(n, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}'$, and $\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{B}} = \boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$.

Therefore by applying an appropriate linear transformation, we can assume that $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{I}_k \\ \boldsymbol{0} \end{bmatrix}$. Then $\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B} = \boldsymbol{A}^{11}$, where $\boldsymbol{A}^{-1} = \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{bmatrix}$. Now (see (2.7))

$$\boldsymbol{A}^{11} = (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1},$$

and hence $(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} = \boldsymbol{A}_{11.2}$. By Theorem 25.2 we have that $\boldsymbol{A}_{11.2} \sim W_k(n-m+k, \boldsymbol{\Sigma}_{11.2})$. It remains to note that here $\boldsymbol{\Sigma}_{11.2} = (\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B})^{-1}$. □

Proposition 25.3 *If $\boldsymbol{A} \sim W_m(n, \boldsymbol{\Sigma})$ and $\boldsymbol{Y}$ is an $m \times 1$ random vector independent of $\boldsymbol{A}$ and such that* $\mathsf{Prob}(\boldsymbol{Y} = \boldsymbol{0}) = 0$*, then*

$$\frac{\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y}}{\boldsymbol{Y}'\boldsymbol{A}^{-1}\boldsymbol{Y}} \sim \chi^2_{n-m+1}. \tag{25.6}$$

Proof. By Theorem 25.3 we have that conditional on $\boldsymbol{Y}$, $(\boldsymbol{Y}'\boldsymbol{A}^{-1}\boldsymbol{Y})^{-1} \sim W_1(n-m+1, (\boldsymbol{Y}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Y})^{-1})$. This implies (25.6). □

25.1 Hotelling's T^2 statistic

Hotelling's T^2 statistic is an extension of t distribution to a multivariate setting. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N$ be an iid sample from normal distribution $\mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{S}$ be the sample covariance matrix. Recall that $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$ are independent.

Suppose that we want to test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ against $H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0$ is a given $m \times 1$ vector. Hotelling's T^2 statistic for testing H_0 is

$$T^2 = N(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0). \tag{25.7}$$

For $m = 1$ this statistic can be written as $\frac{(\bar{X}-\mu_0)^2}{S^2/N}$, where $S^2 = (N-1)^{-1}\sum_{i=1}^N (X_i - \bar{X})^2$ is the sample variance. So in that case $T^2 = t^2$, where $t = \frac{\bar{X}-\mu_0}{S/\sqrt{N}}$ is the usual t statistic.

We proceed now to statistical inference of Hotelling's statistic. For $n = N-1$ we can write

$$\frac{T^2}{n} = \frac{N(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)}{\left(\frac{n(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)}{(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}-\boldsymbol{\mu}_0)}\right)}$$

Under H_0 we have that $N^{1/2}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, and hence $N(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0) \sim \chi^2_m$. Also by Proposition 25.1 we have that $n\boldsymbol{S} \sim W_m(n, \boldsymbol{\Sigma})$ and hence by Proposition 25.3,

$$\frac{n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)}{(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)} \sim \chi^2_{n-m+1}.$$

We obtain the following result.

Theorem 25.4 *Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_N \overset{iid}{\sim} N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then under $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$,*

$$\frac{(N-m)T^2}{m(N-1)} \sim F_{m,N-m}. \tag{25.8}$$

Note that as $N \to \infty$, the coefficient $\frac{N-m}{m(N-1)}$ in (25.8) tends to $1/m$. Therefore for large N the distribution of T^2 becomes like χ^2_m. This should be not surprising since by the LLN, $\boldsymbol{S}$ converges w.p.1 to $\boldsymbol{\Sigma}$, and $N(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$ has χ^2_m distribution when $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ (Theorem 3.1).

Suppose now that we want to test linear model $H_0 : \boldsymbol{A\mu} = \boldsymbol{c}$, where $\boldsymbol{A}$ is a $k \times m$ matrix of rank k and $\boldsymbol{c}$ is $k \times 1$ vector. The corresponding Hotelling's T^2 statistic is

$$T^2 = N \min_{\boldsymbol{A\mu}=\boldsymbol{c}} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}). \tag{25.9}$$

It is possible to write this in the form

$$T^2 = N(\boldsymbol{A}\bar{\boldsymbol{X}} - \boldsymbol{c})' (\boldsymbol{ASA}')^{-1} (\boldsymbol{A}\bar{\boldsymbol{X}} - \boldsymbol{c}). \tag{25.10}$$

Indeed, suppose for the sake of simplicity that $\boldsymbol{c} = \boldsymbol{0}$. Consider $\tilde{\boldsymbol{X}} = \boldsymbol{S}^{-1/2}\bar{\boldsymbol{X}}$ and $\tilde{\boldsymbol{A}} = \boldsymbol{AS}^{1/2}$. Then making change of variables $\boldsymbol{\tau} = \boldsymbol{S}^{-1/2}\boldsymbol{\mu}$ we have

$$\min_{\boldsymbol{A\mu}=\boldsymbol{0}} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})' \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu}) = \min_{\tilde{\boldsymbol{A}}\boldsymbol{\tau}=\boldsymbol{0}} (\tilde{\boldsymbol{X}} - \boldsymbol{\tau})' (\tilde{\boldsymbol{X}} - \boldsymbol{\tau}). \tag{25.11}$$

The right hand side of (25.11) is the squared distance from $\tilde{\boldsymbol{X}}$ to the space orthogonal to the one generated by matrix $\tilde{\boldsymbol{A}}$. Hence

$$\min_{\tilde{\boldsymbol{A}}\boldsymbol{\tau}=\boldsymbol{0}} (\tilde{\boldsymbol{X}} - \boldsymbol{\tau})' (\tilde{\boldsymbol{X}} - \boldsymbol{\tau}) = \tilde{\boldsymbol{X}}' \tilde{\boldsymbol{A}}' (\tilde{\boldsymbol{A}}\tilde{\boldsymbol{A}}')^{-1} \tilde{\boldsymbol{A}}\tilde{\boldsymbol{X}} = (\boldsymbol{A}\bar{\boldsymbol{X}})' (\boldsymbol{ASA}')^{-1} (\boldsymbol{A}\bar{\boldsymbol{X}}).$$

Under H_0, $N^{1/2}(\boldsymbol{A}\bar{\boldsymbol{X}} - \boldsymbol{c}) \sim \mathcal{N}_k(\boldsymbol{0}, \boldsymbol{A\Sigma A}')$ and

$$\frac{(N-k)T^2}{k(N-1)} \sim F_{k,N-k}. \tag{25.12}$$

Indeed, consider $\boldsymbol{Y}_i = \boldsymbol{AX}_i$, $i = 1, ..., N$. We have that $\boldsymbol{Y}_i \sim N_k(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}')$. Also $\bar{\boldsymbol{Y}} = \boldsymbol{A}\bar{\boldsymbol{X}}$ and the corresponding sample covariance matrix is $\boldsymbol{ASA}'$. Hotelling's T^2 statistic for testing $H_0 : \boldsymbol{A\mu} = \boldsymbol{c}$ is given by the left hand side of (25.12).

26 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 $c(\mathbf{0}) = \mathrm{Var}(Z(\boldsymbol{h})) = \mathbb{E}\big[Z(\boldsymbol{h})^2\big]$, and thus

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

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 $c_0 = c(\mathbf{0})$ and $c_{ij} = c(\boldsymbol{x}_i - \boldsymbol{x}_j)$. In matrix form this can be written as $\boldsymbol{\Gamma} = c_0 \mathbf{1}_m \mathbf{1}_m' - \boldsymbol{C}$, 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

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

Moreover

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

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 $c_0 + c_\ell$.

Positive-definite functions. Recall that for complex number $c = a + bi$ its conjugate $\bar{c} = a - bi$, where $i^2 = -1$. A function $\phi : \mathbb{R}^n \to \mathbb{C}$ is positive-definite if $\phi(-\boldsymbol{x}) = \overline{\phi(\boldsymbol{x})}$ and for any $\boldsymbol{x}_1, ..., \boldsymbol{x}_m \in \mathbb{R}^n$ and $c_1, ..., c_m \in \mathbb{C}$ it follows that

$$\sum_{k,\ell=1}^{m} c_k\bar{c}_\ell\phi(\boldsymbol{x}_k - \boldsymbol{x}_\ell) \geq 0.$$

This means that the corresponding $m \times m$ matrix $\boldsymbol{\Gamma}$ with components $\gamma_{k\ell} = \phi(\boldsymbol{x}_k - \boldsymbol{x}_\ell)$ is Hermitian[10]. If $\phi(\boldsymbol{x})$ is real valued, then the corresponding matrix $\boldsymbol{\Gamma}$ is positive semidefinite. For $m = 1$ and $c_1 = 1$ it follows that $\phi(\boldsymbol{x}_1 - \boldsymbol{x}_1) \geq 0$, i.e., $\phi(\boldsymbol{0}) \geq 0$. Also $|\phi(\boldsymbol{x})| \leq \phi(\boldsymbol{0})$.

Recall that $e^{i\theta} = \cos\theta + i\sin\theta$. Consider Fourier transform of finite positive Borel measure μ on $\mathbb{R}^n$

$$\hat{\mu}(\boldsymbol{z}) = \int_{\mathbb{R}^n} e^{-i\boldsymbol{z}'\boldsymbol{x}}d\mu(\boldsymbol{x}),\ \boldsymbol{z} \in \mathbb{R}^n.$$

If $d\mu(\boldsymbol{x}) = f(\boldsymbol{x})d\boldsymbol{x}$, then

$$\hat{\mu}(\boldsymbol{z}) = \int_{\mathbb{R}^n} e^{-i\boldsymbol{z}'\boldsymbol{x}}f(\boldsymbol{x})d\boldsymbol{x},$$

is the Fourier transform of function f. Note that measure μ is positive if $f(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

For any $\boldsymbol{z}_1, ..., \boldsymbol{z}_m \in \mathbb{R}^n$ and $c_1, ..., c_m \in \mathbb{C}$ we have

$$\sum_{k,\ell=1}^{m} c_k\bar{c}_\ell\hat{\mu}(\boldsymbol{z}_k - \boldsymbol{z}_\ell) = \sum_{k,\ell=1}^{m} c_k\bar{c}_\ell\int_{\mathbb{R}^n} e^{-i(\boldsymbol{z}_k-\boldsymbol{z}_\ell)'\boldsymbol{x}}d\mu(\boldsymbol{x}) =$$

$$\int_{\mathbb{R}^n}\sum_{k,\ell=1}^{m} c_k\bar{c}_\ell e^{-i(\boldsymbol{z}_k-\boldsymbol{z}_\ell)'\boldsymbol{x}}d\mu(\boldsymbol{x}) = \int_{\mathbb{R}^n}\left(\sum_{k=1}^{m} c_k e^{-i\boldsymbol{z}_k'\boldsymbol{x}}\right)\overline{\left(\sum_{\ell=1}^{m} c_\ell e^{-i\boldsymbol{z}_j'\boldsymbol{x}}\right)}d\mu(\boldsymbol{x}) =$$

$$\int_{\mathbb{R}^n}\left|\sum_{k=1}^{m} c_k e^{-i\boldsymbol{z}_k'\boldsymbol{x}}\right|^2 d\mu(\boldsymbol{x}) \geq 0.$$

That is, Fourier transform of a finite positive Borel measure is a positive definite function. The converse of that is also true (its proof is not trivial).

Theorem 26.1 (Bochner) *If $\phi : \mathbb{R}^n \to \mathbb{C}$ is positive definite, continuous, and satisfies $\phi(\boldsymbol{0}) = 1$, then there is Borel probability measure μ on $\mathbb{R}^n$ such that ϕ is Fourier transform of μ.*

[10] A matrix $\boldsymbol{A} = [a_{k\ell}]$ is said to be Hermitian if $a_{k\ell} = \bar{a}_{k\ell}$ and $\sum_{k,\ell=1}^{m} a_{k\ell}x_k\bar{x}_\ell \geq 0$ for any $x_1, ..., x_m \in \mathbb{C}$.