

Alexander Shapiro

# Multivariate Statistical Analysis

November 25, 2019

## Contents

<b>1</b>	<b>Some matrix calculus</b>	<b>3</b>
<b>2</b>	<b>Multivariate normal distribution</b>	<b>3</b>
2.1	Schur complement . . . . .	6
<b>3</b>	<b>Quadratic forms</b>	<b>6</b>
<b>4</b>	<b>Statistical inference of linear models</b>	<b>7</b>
4.1	Distribution theory . . . . .	10
4.2	Estimation with linear restrictions . . . . .	11
<b>5</b>	<b>Shrinkage Methods</b>	<b>12</b>
5.1	Ridge Regression . . . . .	12
5.2	Lasso method . . . . .	14
<b>6</b>	<b>Wishart distribution</b>	<b>15</b>
6.1	Hotelling's $T^2$ statistic . . . . .	18
<b>7</b>	<b>Elements of large samples theory</b>	<b>19</b>
7.1	Maximum likelihood method . . . . .	20
7.1.1	Asymptotic distribution of the ML estimators . . . . .	21
7.1.2	Cramer - Rao lower bound . . . . .	23
<b>8</b>	<b>Hypotheses testing</b>	<b>24</b>
8.1	Likelihood Ratio Test . . . . .	25
8.2	Testing equality constraints . . . . .	26
<b>9</b>	<b>Multinomial distribution</b>	<b>27</b>
<b>10</b>	<b>Logistic regression</b>	<b>29</b>
<b>11</b>	<b>Exponential family of distributions</b>	<b>30</b>
<b>12</b>	<b>Generalized linear models</b>	<b>31</b>

<b>13 Moments structural models</b>	<b>33</b>
13.1 Covariance structural models . . . . .	35
<b>14 Principal components analysis</b>	<b>39</b>
<b>15 Spherical and elliptical distributions</b>	<b>42</b>
15.1 Multivariate cumulants . . . . .	44
<b>16 Correlation analysis</b>	<b>45</b>
16.1 Partial correlation . . . . .	45
16.2 Canonical correlation analysis . . . . .	46
<b>17 Classification problem</b>	<b>47</b>
17.1 Classification with normally distributed populations . . . . .	48
17.2 Fisher discriminant analysis . . . . .	48
17.3 Several populations . . . . .	49
17.3.1 Mahalanobis distance . . . . .	49
17.4 Bayes and Logistic Regression classifies . . . . .	49
<b>18 Support Vector Machines</b>	<b>50</b>
18.1 Kernel PCA . . . . .	53

# 1 Some matrix calculus

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

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

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

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

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

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

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

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

$$\boldsymbol{\Sigma}_Y = \mathbf{A}\boldsymbol{\Sigma}_X\mathbf{A}'. \quad (1.1)$$

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

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

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

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

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

$$\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA}). \quad (1.3)$$

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

$$\mathbb{E}[\mathbf{X}'\mathbf{A}\mathbf{X}] = \mathbb{E}[\text{tr}(\mathbf{A}\mathbf{X}\mathbf{X}')] = \text{tr}(\mathbf{A}\mathbb{E}[\mathbf{X}\mathbf{X}']) = \text{tr}(\mathbf{A}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')) = \text{tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu}. \quad (1.4)$$

## 2 Multivariate normal distribution

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

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

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

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

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

$$f_Y(\mathbf{y}) = f_X(\mathbf{A}^{-1}\mathbf{y}) |\mathbf{A}^{-1}| = \frac{1}{(2\pi)^{m/2} |\mathbf{A}|} \exp(-\mathbf{y}' \mathbf{A}'^{-1} \mathbf{A}^{-1} \mathbf{y}/2) = \frac{1}{(2\pi)^{m/2} |\Sigma_Y|^{1/2}} \exp(-\mathbf{y}' \Sigma_Y^{-1} \mathbf{y}/2).$$

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

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

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

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

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

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

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

Moment generating function of a random variable  $X$  is defined as  $M(t) = \mathbb{E}[e^{tX}]$ . Two random variables  $X$  and  $Y$  have the same distribution if their moment generating functions  $M_X(t)$  and  $M_Y(t)$  are equal to each other for all  $t$  in some neighborhood of zero (provided these moment generating functions are finite valued in this neighborhood). Similarly moment generating function of a random vector  $\mathbf{X} = (X_1, \dots, X_m)'$  is defined as  $M(\mathbf{t}) = \mathbb{E}[e^{t_1 X_1 + \dots + t_m X_m}] = \mathbb{E}[\exp(\mathbf{t}'\mathbf{X})]$ . If  $M_X(\mathbf{t})$  is finite valued in a neighborhood of  $\mathbf{0}$ , then

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

Similarly the Hessian matrix of second order partial derivatives

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

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

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

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

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

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

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

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

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

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

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

### Conditional normal distribution

Suppose that  $\mathbf{X} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  is partitioned  $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$  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  $\mathbf{X}_1$  given  $\mathbf{X}_2 = \mathbf{x}_2$ . Consider

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

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

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

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

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

and covariance matrix

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

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

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

## 2.1 Schur complement

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

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

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

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

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

Note that

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

and

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

Hence it follows from (2.3) that

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

This implies that

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

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

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

Using (2.8) it is possible to compute

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

## 3 Quadratic forms

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

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

$$\mathbf{A}\mathbf{e}_i = \lambda_i\mathbf{e}_i, \quad i = 1, \dots, m. \quad (3.1)$$

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

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

$$\mathbf{A} = \mathbf{T}\mathbf{\Lambda}\mathbf{T}' = \sum_{i=1}^m \lambda_i \mathbf{e}_i \mathbf{e}_i'. \quad (3.2)$$

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

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

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

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

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

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

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

### Noncentral chi square distribution

Let  $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{I}_m)$  and consider  $Q = \mathbf{X}'\mathbf{X} = X_1^2 + \dots + X_m^2$ . Note that if  $\mathbf{Y} = \mathbf{T}\mathbf{X}$ , where  $\mathbf{T}$  is an orthogonal matrix, then  $\mathbf{Y}'\mathbf{Y} = \mathbf{X}'\mathbf{X}$  and  $\mathbb{E}[\mathbf{Y}] = \mathbf{T}\boldsymbol{\mu}$  and the covariance matrix of  $\mathbf{Y}$  is  $\mathbf{I}_m$ . It follows that the distribution of  $Q$  depends on  $\delta = \mu_1^2 + \dots + \mu_m^2$  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  $\delta = \mu_1^2 + \dots + \mu_m^2$  and  $m$  degrees of freedom, denoted  $Q \sim \chi_m^2(\delta)$ . Similar to Theorem 3.1 it is possible to show that if  $\mathbf{X} \sim N_m(\boldsymbol{\mu}, \mathbf{\Sigma})$ . Then  $\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X} \sim \chi_m^2(\delta)$  with noncentrality parameter  $\delta = \boldsymbol{\mu}'\mathbf{\Sigma}^{-1}\boldsymbol{\mu}$ . If  $\mathbf{X} \sim N_m(\boldsymbol{\mu}, \mathbf{I}_m)$  and  $\mathbf{P}$  is symmetric projection matrix of rank  $r$ , then  $\mathbf{X}'\mathbf{P}\mathbf{X} \sim \chi_r^2(\delta)$ , where  $\delta = \boldsymbol{\mu}'\mathbf{P}\boldsymbol{\mu}$ .

## 4 Statistical inference of linear models

Consider linear regression model

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

In matrix form we can write this model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4.2)$$

where  $\mathbf{Y} = (Y_1, \dots, Y_n)'$ ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$  and  $\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ & \dots & \dots & \dots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$  is  $n \times p$ ,  $p = k + 1$ ,

design matrix. Note that the first column of  $\mathbf{X}$  is formed by ones. We assume that  $\mathbf{X}$  has full column rank  $p$ .

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

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

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

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

- (i)  $\mathbf{H}$  is symmetric.
- (ii)  $\mathbf{H}$  and  $\mathbf{I}_n - \mathbf{H}$  are idempotent (projection) matrices, i.e.  $\mathbf{H}^2 = \mathbf{H}$  and  $(\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n - \mathbf{H}$ .
- (iii)  $\text{tr}(\mathbf{H}) = \text{tr}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] = \text{tr}(\mathbf{I}_p) = p$  and  $\text{tr}(\mathbf{I}_n - \mathbf{H}) = n - p$ .
- (iv)  $\mathbf{H}\mathbf{X} = \mathbf{X}$  and  $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$ .

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

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

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

$$\begin{aligned} \mathbb{E}[\sum_{i=1}^n (e_i^2 + \dots + e_n^2)] &= \text{Var}(\mathbf{e}'\mathbf{e}) = \mathbb{E}[\text{tr}(\mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y})] \\ &= \text{tr}\{(\mathbf{I}_n - \mathbf{H})\mathbb{E}[\mathbf{Y}\mathbf{Y}']\} = \text{tr}\{(\mathbf{I}_n - \mathbf{H})(\sigma^2\mathbf{I}_n + (\mathbf{X}\boldsymbol{\beta})(\mathbf{X}\boldsymbol{\beta})')\} \\ &= \text{tr}\{(\mathbf{I}_n - \mathbf{H})(\sigma^2\mathbf{I}_n)\} = \sigma^2\text{tr}(\mathbf{I}_n - \mathbf{H}) = \sigma^2(n - p). \end{aligned}$$

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

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

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

$$\text{Var}(\mathbf{a}'\tilde{\boldsymbol{\beta}}) \geq \text{Var}(\mathbf{a}'\hat{\boldsymbol{\beta}}) \quad (4.3)$$

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



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

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

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

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

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

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

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

Hence

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

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

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

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

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

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

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

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

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

Now

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

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

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

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

#### 4.1 Distribution theory

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

$$\sigma^{-2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi_p^2, \quad (4.5)$$

where  $p = k + 1$ .

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

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

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

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

By Theorem 3.2 it follows that

$$\frac{(n-p)S^2}{\sigma^2} \sim \chi_{n-p}^2. \quad (4.6)$$

Moreover

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

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

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

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

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

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

$$S_{YY} = SS_R + SS_E. \quad (4.8)$$

Indeed, we have that

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

and

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

Also

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

and

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

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

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

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

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

and hence  $SS_E$  and  $SS_R$  are independent.

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

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

Recall that  $SS_E/\sigma^2 \sim \chi_{n-p}^2$ . Also under  $H_0$  we have that  $\mathbb{E}[\mathbf{Y}] = \beta_0\mathbf{1}_n$  and hence

$$(\mathbf{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbb{E}[\mathbf{Y}] = \beta_0(\mathbf{H} - n^{-1}\mathbf{1}_n\mathbf{1}_n')\mathbf{1}_n = \beta_0(\mathbf{H}\mathbf{1}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'\mathbf{1}_n) = \mathbf{1}_n - \mathbf{1}_n = \mathbf{0}.$$

Consequently

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

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

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

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

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

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

## 4.2 Estimation with linear restrictions

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

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

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

Consider the Lagrangian of the above problem (4.12):

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

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

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

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

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

The corresponding estimators are given by

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

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

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

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

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

and

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

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

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

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

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

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

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

## 5 Shrinkage Methods

For a vector  $\mathbf{x} \in \mathbb{R}^m$  the  $\ell_q$  norm is  $\|\mathbf{x}\|_q = (|x_1|^q + \dots + |x_m|^q)^{1/q}$ ,  $q \geq 1$ . In particular, the  $\ell_2$  norm is the Euclidean norm, and  $\ell_1$  norm is  $\|\mathbf{x}\|_1 = |x_1| + \dots + |x_m|$ . Note that function  $\|\cdot\|_q$  is positively homogeneous (i.e.,  $\|\lambda\mathbf{x}\|_q = \lambda\|\mathbf{x}\|_q$  for  $\lambda \geq 0$ ) for any  $q > 0$ . However for  $q \in (0, 1)$ ,  $\|\cdot\|_q$  is not convex.

### 5.1 Ridge Regression

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

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \varepsilon\|\boldsymbol{\beta}\|_2^2, \quad (5.18)$$

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

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

That is  $\tilde{\beta}_\varepsilon = (\mathbf{X}'\mathbf{X} + \varepsilon\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y}$  (recall that  $p = k + 1$  is the number of estimated parameters). Of course for  $\varepsilon = 0$  the estimator  $\tilde{\beta}_\varepsilon$  coincides with Least Squares (LS) estimator  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ . It is also possible to formulate problem (5.18) in the following form

$$\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \quad \text{s.t. } \|\beta\|_2 \leq c, \quad (5.19)$$

for a certain value of  $c > 0$  (take  $c = \|\tilde{\beta}_\varepsilon\|_2$ ). Conversely solution of problem (5.19), for some  $c > 0$ , is also the solution of problem (5.18) when  $\varepsilon$  is the corresponding Lagrange multiplier. (If  $\|\hat{\beta}\|_2 \leq c$ , then the corresponding  $\varepsilon = 0$ .) Therefore in a sense problems (5.18) and (5.19) are equivalent to each other for a proper choice of the respective positive constants  $\varepsilon$  and  $c$ .

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

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

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

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

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

and hence

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

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

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

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

We also have that

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

Therefore

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

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

## 5.2 Lasso method

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

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \varepsilon\|\boldsymbol{\beta}\|_1. \quad (5.20)$$

Equivalently this can be formulated as

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq c, \quad (5.21)$$

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

Note that

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

where  $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$  is the usual least squares estimator. When  $c < \|\hat{\boldsymbol{\beta}}\|_1$ , an optimal solution of problem (5.21) is on the boundary of the feasible set  $C = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq c\}$  and the corresponding optimality conditions are

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

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

Optimality conditions for problem (5.20) are

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

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  $\mathbf{g}$  (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  $\varepsilon$  is bigger than the absolute value of every component  $[\mathbf{X}'\mathbf{Y}]_i$  of vector  $\mathbf{X}'\mathbf{Y}$ , then  $\tilde{\boldsymbol{\beta}}_\varepsilon = \mathbf{0}$ . If  $\mathbf{X}'\mathbf{X}$  is diagonal, then  $\tilde{\beta}_{\varepsilon,i} = 0$ , when  $\varepsilon > [\mathbf{X}'\mathbf{Y}]_i$ .

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

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_0 \leq c, \quad (5.22)$$

This is a difficult combinatorial problem. Problem (5.21) can be viewed as a convex approximation of problem (5.22). Problem (5.21) can be written as the following quadratic programming problem

$$\begin{aligned} \min_{\boldsymbol{\beta}, \boldsymbol{\xi}} \quad & \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{s.t.} \quad & \beta_i \leq \xi_i, \quad -\beta_i \leq \xi_i, \quad i = 0, \dots, k, \\ & \sum_{i=0}^k \xi_i \leq c. \end{aligned} \quad (5.23)$$

Problem (5.20) can be formulated as the following problem

$$\begin{aligned} \min_{\boldsymbol{\beta} \in \mathbb{R}^p, z \in \mathbb{R}} \quad & \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + z \\ \text{s.t.} \quad & \varepsilon\|\boldsymbol{\beta}\|_1 \leq z, \end{aligned} \quad (5.24)$$

which in turn can be written as a quadratic programming problem similar to (5.23).

## 6 Wishart distribution

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

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

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

$$\text{vec}(\mathbf{A}) = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_q \end{bmatrix},$$

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

For basic properties of these notions see pages 6-9 of *Multivariate analysis notes*. In particular we use the following matrix identities

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}) \quad (6.1)$$

and

$$\text{vec}(\mathbf{BXC}) = (\mathbf{C}' \otimes \mathbf{B})\text{vec}(\mathbf{X}), \quad (6.2)$$

where  $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}, \mathbf{X}$  are matrices of appropriate order.

Let  $\mathbf{Z}_1, \dots, \mathbf{Z}_n$  be an iid sequence of random vectors having normal distribution  $N_m(\mathbf{0}, \mathbf{\Sigma})$ . Consider random matrix

$$\mathbf{A} = \mathbf{Z}_1\mathbf{Z}_1' + \dots + \mathbf{Z}_n\mathbf{Z}_n'. \quad (6.3)$$

By definition  $\mathbf{A}$  has Wishart distribution, denoted  $\mathbf{A} \sim W_m(n, \mathbf{\Sigma})$ . In particular, for  $m = 1$  and  $Z_i \sim N(0, \sigma^2)$ , the corresponding  $A/\sigma^2$  has chi-square distribution with  $n$  degrees of freedom.

Wishart distribution has the following properties.

- (i) If  $\mathbf{A} \sim W_m(n, \mathbf{\Sigma})$  and  $\alpha > 0$ , then  $\alpha\mathbf{A} \sim W_m(n, \alpha\mathbf{\Sigma})$ . Indeed,

$$\alpha\mathbf{A} = (\alpha^{1/2}\mathbf{Z}_1)(\alpha^{1/2}\mathbf{Z}_1)' + \dots + (\alpha^{1/2}\mathbf{Z}_n)(\alpha^{1/2}\mathbf{Z}_n)',$$

and  $\alpha^{1/2}\mathbf{Z}_i \sim N(\mathbf{0}, \alpha\mathbf{\Sigma})$ .

- (ii) If  $\mathbf{A} \sim W_m(n, \mathbf{\Sigma})$  and  $\mathbf{B}$  is  $m \times k$  deterministic matrix, then  $\mathbf{B}'\mathbf{A}\mathbf{B} \sim W_k(n, \mathbf{B}'\mathbf{\Sigma}\mathbf{B})$ . Indeed,

$$\mathbf{B}'\mathbf{A}\mathbf{B} = (\mathbf{B}'\mathbf{Z}_1)(\mathbf{B}'\mathbf{Z}_1)' + \dots + (\mathbf{B}'\mathbf{Z}_n)(\mathbf{B}'\mathbf{Z}_n)',$$

and  $\mathbf{B}'\mathbf{Z}_i \sim N(\mathbf{0}, \mathbf{B}'\mathbf{\Sigma}\mathbf{B})$ .

- (iii) Equation (6.3) can be written in the following form  $\mathbf{A} = \mathbf{Z}'\mathbf{Z}$ , where  $\mathbf{Z}$  is  $n \times m$  matrix

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_n' \end{bmatrix} \quad \text{with } \mathbf{Z}' = [\mathbf{Z}_1, \dots, \mathbf{Z}_n]. \quad \text{Note that } \mathbb{E}[\mathbf{Z}] = \mathbf{0} \text{ and the covariance matrix of}$$

the corresponding  $mn \times 1$  vector  $\text{vec}(\mathbf{Z}')$  is

$$\text{Cov}(\text{vec}(\mathbf{Z}')) = \mathbf{I}_n \otimes \mathbf{\Sigma}. \quad (6.4)$$

**Proposition 6.1** Let  $\mathbf{X}_1, \dots, \mathbf{X}_N \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and  $\mathbf{S} = (N-1)^{-1} \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$  be the corresponding sample covariance matrix. Then  $\mathbf{S} \sim W_m(n, n^{-1}\boldsymbol{\Sigma})$ , where  $n = N-1$ .

**Proof.** Consider  $N \times m$  matrix  $\mathbf{W}$  with rows  $\mathbf{X}_i - \bar{\mathbf{X}}$ , i.e.,  $\mathbf{W}' = [\mathbf{X}_1 - \bar{\mathbf{X}}, \dots, \mathbf{X}_N - \bar{\mathbf{X}}]$ . Note that  $\mathbf{S} = n^{-1}\mathbf{W}'\mathbf{W}$  and  $\mathbf{W} = (\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N')\mathbf{X}$ , where  $\mathbf{X}' = [\mathbf{X}_1, \dots, \mathbf{X}_N]$ . Matrix  $\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N'$  is a symmetric projection matrix of rank  $N-1$ . Hence  $\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N' = \mathbf{H}\mathbf{H}'$ , where  $\mathbf{H}$  is  $N \times n$  matrix with  $\mathbf{H}'\mathbf{H} = \mathbf{I}_n$  and  $\mathbf{H}'\mathbf{1}_N = \mathbf{0}$  (spectral decomposition). Consider the following  $n \times m$  matrix  $\mathbf{Z} = \mathbf{H}'\mathbf{X}$ . Then

$$\mathbf{Z}'\mathbf{Z} = \mathbf{X}'\mathbf{H}\mathbf{H}'\mathbf{X} = \mathbf{X}'(\mathbf{I}_N - N^{-1}\mathbf{1}_N\mathbf{1}_N')\mathbf{X} = \mathbf{W}'\mathbf{W},$$

and hence  $\mathbf{S} = n^{-1}\mathbf{Z}'\mathbf{Z}$ . Note that  $\mathbb{E}[\mathbf{X}] = \mathbf{1}_N\boldsymbol{\mu}'$  and hence  $\mathbb{E}[\mathbf{Z}] = \mathbf{H}'\mathbb{E}[\mathbf{X}] = \mathbf{H}'\mathbf{1}_N\boldsymbol{\mu}' = \mathbf{0}$ . Now  $\text{Cov}(\text{vec}(\mathbf{X}')) = \mathbf{I}_N \otimes \boldsymbol{\Sigma}$  and (see (6.2))

$$\text{vec}(\mathbf{X}'\mathbf{H}) = (\mathbf{H}' \otimes \mathbf{I}_m)\text{vec}(\mathbf{X}').$$

Thus using (6.1),

$$\text{Cov}(\text{vec}(\mathbf{Z}')) = \text{Cov}(\text{vec}(\mathbf{X}'\mathbf{H})) = (\mathbf{H}' \otimes \mathbf{I}_m)(\mathbf{I}_N \otimes \boldsymbol{\Sigma})(\mathbf{H} \otimes \mathbf{I}_m) = (\mathbf{H}'\mathbf{H}) \otimes \boldsymbol{\Sigma} = \mathbf{I}_n \otimes \boldsymbol{\Sigma}.$$

Hence  $\mathbf{Z}'\mathbf{Z} \sim W_m(n, \boldsymbol{\Sigma})$ , and  $\mathbf{S} \sim W_m(n, n^{-1}\boldsymbol{\Sigma})$ . ■

**Theorem 6.1** If  $\mathbf{A} \sim W_m(n, \boldsymbol{\Sigma})$  and  $\mathbf{Y}$  is an  $m \times 1$  random vector independent of  $\mathbf{A}$  and such that  $\text{Prob}(\mathbf{Y} = \mathbf{0}) = 0$ , then random variable  $\frac{\mathbf{Y}'\mathbf{A}\mathbf{Y}}{\mathbf{Y}'\boldsymbol{\Sigma}\mathbf{Y}} \sim \chi_n^2$  and is independent of  $\mathbf{Y}$ .

**Proof.** Conditional on  $\mathbf{Y}$ , we have that  $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim W_1(n, \mathbf{Y}'\boldsymbol{\Sigma}\mathbf{Y})$  and hence

$$\frac{\mathbf{Y}'\mathbf{A}\mathbf{Y}}{\mathbf{Y}'\boldsymbol{\Sigma}\mathbf{Y}} \sim W_1(n, 1) = \chi_n^2.$$

That is, the conditional distribution of  $\frac{\mathbf{Y}'\mathbf{A}\mathbf{Y}}{\mathbf{Y}'\boldsymbol{\Sigma}\mathbf{Y}}$  does not depend on  $\mathbf{Y}$ . It follows that  $\frac{\mathbf{Y}'\mathbf{A}\mathbf{Y}}{\mathbf{Y}'\boldsymbol{\Sigma}\mathbf{Y}}$  is independent of  $\mathbf{Y}$  and its (unconditional) distribution is  $\chi_n^2$ . ■

Together with Proposition 6.1 this implies the following.

**Proposition 6.2** Let  $\mathbf{X}_1, \dots, \mathbf{X}_N \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  and  $\mathbf{S}$  be the corresponding sample covariance matrix. Then  $\frac{n\bar{\mathbf{X}}'\mathbf{S}\bar{\mathbf{X}}}{\bar{\mathbf{X}}'\boldsymbol{\Sigma}\bar{\mathbf{X}}} \sim \chi_n^2$  and is independent of  $\bar{\mathbf{X}}$  (recall that  $n = N-1$ ).

**Proof.** Since  $\mathbf{S}$  and  $\bar{\mathbf{X}}$  are independent and  $n\mathbf{S} \sim W_m(n, \boldsymbol{\Sigma})$ , the result follows from Theorem 6.1. ■

**Theorem 6.2** Let  $\mathbf{A} \sim W_m(n, \boldsymbol{\Sigma})$  be partitioned  $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ , where  $\mathbf{A}_{11}$  is of order  $k \times k$  and  $\mathbf{A}_{22}$  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  $\mathbf{A}_{11.2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$  and  $\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ . Then

$$\mathbf{A}_{11.2} \sim W_k(n - m + k, \boldsymbol{\Sigma}_{11.2}), \quad (6.5)$$

and  $\mathbf{A}_{11.2}$  is independent of  $\mathbf{A}_{22}$ .



**Proof.** Since  $\mathbf{A} \sim W_m(n, \Sigma)$  it can be written in the form (6.3), or equivalently as  $\mathbf{A} = \mathbf{Z}'\mathbf{Z}$ ,

where  $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}'_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{Z}'_n \end{bmatrix}$  is the respective  $n \times m$  matrix. Let us partition  $\mathbf{Z} = [\tilde{\mathbf{Z}}_1, \tilde{\mathbf{Z}}_2]$ , where  $\tilde{\mathbf{Z}}_1$  is

of order  $n \times k$  and  $\tilde{\mathbf{Z}}_2$  is of order  $n \times (m - k)$ . Note that  $\tilde{\mathbf{Z}}_1 = \begin{bmatrix} \mathbf{Z}'_{11} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{Z}'_{1n} \end{bmatrix}$  and  $\tilde{\mathbf{Z}}_2 = \begin{bmatrix} \mathbf{Z}'_{21} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{Z}'_{2n} \end{bmatrix}$ ,

where  $\mathbf{Z}_i = \begin{bmatrix} \mathbf{Z}_{1i} \\ \mathbf{Z}_{2i} \end{bmatrix}$  are respective partitions of vectors  $\mathbf{Z}_i$ ,  $i = 1, \dots, n$ . Recall that conditional on  $\mathbf{Z}_{2i} = \mathbf{z}_2$ ,

$$\mathbf{Z}_{1i} \sim N(\Sigma_{12}\Sigma_{22}^{-1}\mathbf{z}_2, \Sigma_{11.2}) \quad (6.6)$$

(see equation (2.2)). Note that matrix  $\mathbf{I}_n - \tilde{\mathbf{Z}}_2(\tilde{\mathbf{Z}}_2'\tilde{\mathbf{Z}}_2)^{-1}\tilde{\mathbf{Z}}_2'$  is idempotent (projection) of rank  $n - (m - k) = n - m + k$ , and

$$[\mathbf{I}_n - \tilde{\mathbf{Z}}_2(\tilde{\mathbf{Z}}_2'\tilde{\mathbf{Z}}_2)^{-1}\tilde{\mathbf{Z}}_2']\tilde{\mathbf{Z}}_2 = \mathbf{0}. \quad (6.7)$$

Because of (6.6) and (6.7) we have that conditional on  $\tilde{\mathbf{Z}}_2$ ,

$$\tilde{\mathbf{Z}}_1'[\mathbf{I}_n - \tilde{\mathbf{Z}}_2(\tilde{\mathbf{Z}}_2'\tilde{\mathbf{Z}}_2)^{-1}\tilde{\mathbf{Z}}_2']\tilde{\mathbf{Z}}_1 \sim W_k(n - m + k, \Sigma_{11.2}).$$

Moreover

$$\tilde{\mathbf{Z}}_1'[\mathbf{I}_n - \tilde{\mathbf{Z}}_2(\tilde{\mathbf{Z}}_2'\tilde{\mathbf{Z}}_2)^{-1}\tilde{\mathbf{Z}}_2']\tilde{\mathbf{Z}}_1 = \underbrace{\tilde{\mathbf{Z}}_1'\tilde{\mathbf{Z}}_1}_{\mathbf{A}_{11}} - \underbrace{\tilde{\mathbf{Z}}_1'\tilde{\mathbf{Z}}_2}_{\mathbf{A}_{12}} \underbrace{(\tilde{\mathbf{Z}}_2'\tilde{\mathbf{Z}}_2)^{-1}}_{\mathbf{A}_{22}^{-1}} \underbrace{\tilde{\mathbf{Z}}_2'\tilde{\mathbf{Z}}_1}_{\mathbf{A}_{21}} = \mathbf{A}_{11.2}.$$

It follows that the (unconditional) distribution of  $\mathbf{A}_{11.2}$  is  $W_k(n - m + k, \Sigma_{11.2})$ , and that  $\mathbf{A}_{11.2}$  is independent of  $\tilde{\mathbf{Z}}_2$  and hence of  $\mathbf{A}_{22}$ . ■

**Theorem 6.3** *Let  $\mathbf{A} \sim W_m(n, \Sigma)$  and  $\mathbf{B}$  be (deterministic)  $m \times k$  matrix of rank  $k$ . Then  $(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} \sim W_k(n - m + k, (\mathbf{B}'\Sigma^{-1}\mathbf{B})^{-1})$ .*

**Proof.** Note that the assertion is invariant under linear transformations. That is, if  $\mathbf{C}$  is an  $m \times m$  nonsingular matrix, then by replacing  $\mathbf{B}$  with  $\tilde{\mathbf{B}} = \mathbf{C}\mathbf{B}$  and  $\mathbf{A}$  with  $\tilde{\mathbf{A}} = \mathbf{C}\mathbf{A}\mathbf{C}'$  we have  $\tilde{\mathbf{B}}'\tilde{\mathbf{A}}^{-1}\tilde{\mathbf{B}} = \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}$ . Moreover  $\tilde{\mathbf{A}} \sim W_m(n, \tilde{\Sigma})$ , where  $\tilde{\Sigma} = \mathbf{C}\Sigma\mathbf{C}'$ , and  $\tilde{\mathbf{B}}'\tilde{\Sigma}^{-1}\tilde{\mathbf{B}} = \mathbf{B}'\Sigma^{-1}\mathbf{B}$ .

Therefore by applying an appropriate linear transformation, we can assume that  $\mathbf{B} = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix}$ .

Then  $\mathbf{B}'\mathbf{A}^{-1}\mathbf{B} = \mathbf{A}^{11}$ , where  $\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}$ . Now (see (2.9))

$$\mathbf{A}^{11} = (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1},$$

and hence  $(\mathbf{B}'\mathbf{A}^{-1}\mathbf{B})^{-1} = \mathbf{A}_{11.2}$ . By Theorem 6.2 we have that  $\mathbf{A}_{11.2} \sim W_k(n - m + k, \Sigma_{11.2})$ . It remains to note that here  $\Sigma_{11.2} = (\mathbf{B}'\Sigma^{-1}\mathbf{B})^{-1}$ . ■

**Proposition 6.3** *If  $\mathbf{A} \sim W_m(n, \Sigma)$  and  $\mathbf{Y}$  is an  $m \times 1$  random vector independent of  $\mathbf{A}$  and such that  $\text{Prob}(\mathbf{Y} = \mathbf{0}) = 0$ , then*

$$\frac{\mathbf{Y}'\Sigma^{-1}\mathbf{Y}}{\mathbf{Y}'\mathbf{A}^{-1}\mathbf{Y}} \sim \chi_{n-m+1}^2. \quad (6.8)$$

**Proof.** By Theorem 6.3 we have that conditional on  $\mathbf{Y}$ ,  $(\mathbf{Y}'\mathbf{A}^{-1}\mathbf{Y})^{-1} \sim W_1(n - m + 1, (\mathbf{Y}'\mathbf{\Sigma}^{-1}\mathbf{Y})^{-1})$ . This implies (6.8). ■

### 6.1 Hotelling's $T^2$ statistic

Let  $\mathbf{X}_1, \dots, \mathbf{X}_N \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  be an iid sample. 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

$$T^2 = N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0). \quad (6.9)$$

Recall that  $\bar{\mathbf{X}}$  and  $\mathbf{S}$  are independent. For  $n = N - 1$  we can write

$$\frac{T^2}{n} = \frac{N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)}{\left( \frac{n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)}{(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)} \right)}$$

Assuming  $H_0$  we have that  $N^{1/2}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ , and hence  $N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0) \sim \chi_m^2$ . Also by Proposition 6.1 we have that  $n\mathbf{S} \sim W_m(n, \boldsymbol{\Sigma})$  and hence by Proposition 6.3,

$$\frac{n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)}{(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)} \sim \chi_{n-m+1}^2.$$

We obtain the following result.

**Theorem 6.4** *Let  $\mathbf{X}_1, \dots, \mathbf{X}_N \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  be an iid sample. Then under  $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ ,*

$$\frac{(N - m)T^2}{m(N - 1)} \sim F_{m, N-m}. \quad (6.10)$$

Suppose now that we want to test linear model  $H_0 : \mathbf{A}\boldsymbol{\mu} = \mathbf{c}$ , where  $\mathbf{A}$  is a  $k \times m$  matrix of rank  $k$  and  $\mathbf{c}$  is  $k \times 1$  vector. The corresponding Hotelling's  $T^2$  statistic is

$$T^2 = N \min_{\mathbf{A}\boldsymbol{\mu}=\mathbf{c}} (\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}). \quad (6.11)$$

It is possible to write this in the form

$$T^2 = N(\mathbf{A}\bar{\mathbf{X}} - \mathbf{c})'(\mathbf{A}\mathbf{S}\mathbf{A}')^{-1}(\mathbf{A}\bar{\mathbf{X}} - \mathbf{c}). \quad (6.12)$$

Indeed, suppose for the sake of simplicity that  $\mathbf{c} = \mathbf{0}$ . Consider  $\tilde{\mathbf{X}} = \mathbf{S}^{-1/2}\bar{\mathbf{X}}$  and  $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{S}^{1/2}$ . Then making change of variables  $\boldsymbol{\tau} = \mathbf{S}^{-1/2}\boldsymbol{\mu}$  we have

$$\min_{\mathbf{A}\boldsymbol{\mu}=\mathbf{0}} (\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) = \min_{\tilde{\mathbf{A}}\boldsymbol{\tau}=\mathbf{0}} (\tilde{\mathbf{X}} - \boldsymbol{\tau})'(\tilde{\mathbf{X}} - \boldsymbol{\tau}). \quad (6.13)$$

The right hand side of (6.13) is the squared distance from  $\tilde{\mathbf{X}}$  to the space orthogonal to the one generated by matrix  $\tilde{\mathbf{A}}$ . Hence

$$\min_{\tilde{\mathbf{A}}\boldsymbol{\tau}=\mathbf{0}} (\tilde{\mathbf{X}} - \boldsymbol{\tau})'(\tilde{\mathbf{X}} - \boldsymbol{\tau}) = \tilde{\mathbf{X}}' \tilde{\mathbf{A}}'(\tilde{\mathbf{A}}\tilde{\mathbf{A}}')^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{X}} = (\mathbf{A}\bar{\mathbf{X}})'(\mathbf{A}\mathbf{S}\mathbf{A}')^{-1}(\mathbf{A}\bar{\mathbf{X}}).$$

Under  $H_0$ ,  $N^{1/2}(\mathbf{A}\bar{\mathbf{X}} - \mathbf{c}) \sim N_k(\mathbf{0}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$  and

$$\frac{(N - k)T^2}{k(N - 1)} \sim F_{k, N-k}. \quad (6.14)$$

Indeed, consider  $\mathbf{Y}_i = \mathbf{A}\mathbf{X}_i$ ,  $i = 1, \dots, N$ . We have that  $\mathbf{Y}_i \sim N_k(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ . Also  $\bar{\mathbf{Y}} = \mathbf{A}\bar{\mathbf{X}}$  and the corresponding sample covariance matrix is  $\mathbf{A}\mathbf{S}\mathbf{A}'$ . Hence Hotelling's  $T^2$  statistic for testing  $H_0 : \mathbf{A}\boldsymbol{\mu} = \mathbf{c}$  is (6.14).

## 7 Elements of large samples theory

Let  $Y_n, n = 1, \dots$ , 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

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

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

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

The above inequality sometimes is called Markov inequality.

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

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

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

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

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

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

**Theorem 7.1** *If  $X_n - X \xrightarrow{p} 0$ , then  $X_n \Rightarrow X$ .*

**Proof.** For  $\varepsilon > 0$  we have

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

Similarly

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

and hence

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

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

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

It follows that  $\mathbb{P}(X_n \leq x) \rightarrow \mathbb{P}(X \leq x)$  if  $\mathbb{P}(X = x) = 0$ . This shows that  $X_n \Rightarrow X$ . ■

**Theorem 7.2 (Delta theorem)** Let  $\mathbf{X}_n$  be a sequence of  $m \times 1$  random vectors and  $\mathbf{g} : \mathbb{R}^m \rightarrow \mathbb{R}^k$  be a function. Suppose that  $\lambda_n(\mathbf{X}_n - \boldsymbol{\mu}) \Rightarrow \mathbf{Z}$ , where  $\boldsymbol{\mu} \in \mathbb{R}^m$  and  $\lambda_n \rightarrow \infty$ , and that  $\mathbf{g}(\cdot)$  is differentiable at  $\boldsymbol{\mu}$  with  $\nabla \mathbf{g}(\boldsymbol{\mu})$  being the  $m \times k$  matrix of partial derivatives. Then

$$\lambda_n(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})) \Rightarrow [\nabla \mathbf{g}(\boldsymbol{\mu})]' \mathbf{Z}. \quad (7.1)$$

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

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

where  $\varepsilon(\mathbf{x}) \rightarrow 0$  as  $\mathbf{x} \rightarrow \boldsymbol{\mu}$ . Hence

$$\lambda_n(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu})) = [\nabla \mathbf{g}(\boldsymbol{\mu})]'[\lambda_n(\mathbf{X}_n - \boldsymbol{\mu})] + \varepsilon(\mathbf{X}_n)[\lambda_n(\mathbf{X}_n - \boldsymbol{\mu})]. \quad (7.2)$$

Now since  $\lambda_n(\mathbf{X}_n - \boldsymbol{\mu})$  converges in distribution, it follows that  $\lambda_n(\mathbf{X}_n - \boldsymbol{\mu})$  is bounded in probability. Moreover since  $\lambda_n \rightarrow \infty$  it follows that  $\mathbf{X}_n \xrightarrow{p} \boldsymbol{\mu}$ . Hence  $\varepsilon(\mathbf{X}_n) \xrightarrow{p} 0$ , and thus  $\varepsilon(\mathbf{X}_n)[\lambda_n(\mathbf{X}_n - \boldsymbol{\mu})] \xrightarrow{p} \mathbf{0}$ . By Slutsky's theorem the convergence (7.1) follows from (7.2). ■

In particular it follows that if in addition to the assumptions of Theorem 7.2,  $\sqrt{n}(\mathbf{X}_n - \boldsymbol{\mu})$  converges in distribution to normal  $N(\mathbf{0}, \boldsymbol{\Sigma})$ , then  $\sqrt{n}(\mathbf{g}(\mathbf{X}_n) - \mathbf{g}(\boldsymbol{\mu}))$  converges in distribution to normal with zero mean and covariance matrix  $[\nabla \mathbf{g}(\boldsymbol{\mu})]' \boldsymbol{\Sigma} [\nabla \mathbf{g}(\boldsymbol{\mu})]$ . For example let  $X_n$  and  $Y_n$  be two independent sequences of random variables such that  $\sqrt{n}(X_n - \mu_x) \Rightarrow N(0, \sigma_x^2)$  and  $\sqrt{n}(Y_n - \mu_y) \Rightarrow N(0, \sigma_y^2)$ ,  $\mu_y \neq 0$ . Let us find the asymptotic distribution of  $(V_n, W_n)$ , where  $V_n = X_n Y_n$  and  $W_n = X_n / Y_n$ . Consider  $\mathbf{g}(x, y) = (xy, x/y)$ . Note that  $\mathbf{g}(X_n, Y_n) = (V_n, W_n)$ . By Delta Theorem we have that  $\sqrt{n} \begin{bmatrix} V_n - \mu_x \mu_y \\ W_n - \mu_x / \mu_y \end{bmatrix}$  converges in distribution to normal  $N(\mathbf{0}, \boldsymbol{\Sigma})$  with

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

## 7.1 Maximum likelihood method

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

$$\hat{\boldsymbol{\theta}}_N \in \arg \max_{\boldsymbol{\theta} \in \Theta} \log L_N(\boldsymbol{\theta}). \quad (7.3)$$

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

Since  $\log L_N(\boldsymbol{\theta}) = \sum_{i=1}^N \log f(\mathbf{X}_i, \boldsymbol{\theta})$ , 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  $N \rightarrow \infty$  to

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

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

**Theorem 7.3 (Jensen inequality)** *Let  $\phi : \mathbb{R}^m \rightarrow \mathbb{R}$  be a convex function and  $\mathbf{X}$  be an  $m \times 1$  random vector with mean  $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ . Then*

$$\mathbb{E}[\phi(\mathbf{X})] \geq \phi(\boldsymbol{\mu}). \quad (7.4)$$

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

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

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

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

Since  $\mathbb{E}[\boldsymbol{\gamma}'(\mathbf{x} - \boldsymbol{\mu})] = \boldsymbol{\gamma}'(\mathbb{E}[\mathbf{X}] - \boldsymbol{\mu}) = 0$ , the inequality (7.4) follows. ■

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

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

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

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

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

Since

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

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

### 7.1.1 Asymptotic distribution of the ML estimators

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

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

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

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

We need to show that

$$\mathbb{E}_\theta \left\{ \frac{\partial \log f(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_j} \right\} = -\mathbb{E}_\theta \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\mathbf{X}, \boldsymbol{\theta}) \right\}, \quad (7.5)$$

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

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

Consequently

$$\frac{\partial}{\partial \theta_j} \int \frac{\partial \log f(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_i} f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x} = 0. \quad (7.6)$$

By taking the derivative, in the left hand side of (7.6), inside the integral one obtains

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

and hence (7.5) follows.

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

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

where

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

and hence

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

It follows that

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

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

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

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[ \sum_{i=1}^N \log f(\mathbf{X}_i, \hat{\boldsymbol{\theta}}_N) \right] = \mathbf{0}. \quad (7.7)$$

By the Mean Value Theorem we can write

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

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

$$\sqrt{N}(\hat{\theta}_N - \theta^*) = - \left[ \frac{1}{N} \frac{\partial^2}{\partial \theta \partial \theta'} \sum_{i=1}^N \log f(\mathbf{X}_i, \tilde{\theta}_N) \right]^{-1} \left[ \frac{1}{\sqrt{N}} \frac{\partial}{\partial \theta} \sum_{i=1}^N \log f(\mathbf{X}_i, \theta^*) \right]. \quad (7.8)$$

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

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

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

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

Therefore by the CLT we have that  $\frac{1}{\sqrt{N}} \frac{\partial}{\partial \theta} \sum_{i=1}^N \log f(\mathbf{X}_i, \theta^*)$  converges in distribution to normal with zero mean vector and covariance matrix  $\mathbf{I}(\theta^*)$ . Together with (7.8) this implies that

$$\sqrt{N}(\hat{\theta}_N - \theta^*) \Rightarrow N(\mathbf{0}, \mathbf{I}(\theta^*)^{-1}).$$

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

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

Lebesgue Dominated Convergence Theorem:

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

We have

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

provided

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

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

### 7.1.2 Cramer - Rao lower bound

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

$$\text{Var}_\theta[T(\mathbf{X})] \geq i_X(\theta)^{-1} [\partial g(\theta) / \partial \theta]^2, \quad (7.9)$$

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

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

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

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

Indeed, we have that

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

provided the derivative can be taken inside the integral. Then

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

That is,

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

Now by Cauchy inequality we have

$$[\text{Cov}_\theta (T(\mathbf{X}), \frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta))]^2 \leq \text{Var}_\theta [T(\mathbf{X})] \text{Var}_\theta [\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)].$$

Moreover

$$\text{Var}_\theta [\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)] = \mathbb{E}_\theta [(\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta))^2] = i(\theta),$$

and hence the inequality (7.9) follows. Note that, by the independence of  $X_1, \dots, X_n$ ,

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

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

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

$$\mathbf{a}' \Sigma_T \mathbf{a} \geq \mathbf{a}' \mathbf{I}(\boldsymbol{\theta})^{-1} \mathbf{a}, \quad (7.10)$$

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

## 8 Hypotheses testing

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

**Theorem 8.1 (Neyman - Pearson Lemma)** *Consider simple alternatives  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$  versus  $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$  with pdfs  $f(x, \boldsymbol{\theta}_i)$ ,  $i = 0, 1$ . Then the minimal error rejection region is  $R = \{\mathbf{x} \in \mathbb{R}^N : f(\mathbf{x}, \boldsymbol{\theta}_1) \geq \kappa f(\mathbf{x}, \boldsymbol{\theta}_0)\}$ , where  $\kappa > 0$  is such that  $\int_R f(\mathbf{x}, \boldsymbol{\theta}_0) d\mathbf{x} = \alpha$  with  $f(\mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^N f(x_i, \boldsymbol{\theta})$ .*

**Proof** Consider minimization of

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



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

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

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

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

It follows that the minimum is attained for

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

and

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

## 8.1 Likelihood Ratio Test

Consider

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

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

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

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

For simple alternatives when  $\Theta_0 = \{\boldsymbol{\theta}_0\}$  and  $\Theta_1 = \{\boldsymbol{\theta}_1\}$  we have that

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

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

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

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

Note that

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

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

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) = \sum_{i=1}^N \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i, \boldsymbol{\theta}) \quad (8.1)$$

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

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

Note that  $\frac{\partial}{\partial \theta} \log L(\theta_0) = S(\theta_0)$  and  $\frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_0) = \frac{\partial}{\partial \theta'} S(\theta_0)$ . Hence and since  $S(\hat{\theta}) = \mathbf{0}$ ,

$$\log L(\hat{\theta}) \approx \log L(\theta_0) - [S(\hat{\theta}) - S(\theta_0)]' (\hat{\theta} - \theta_0) + \frac{1}{2} (\hat{\theta} - \theta_0)' \left[ \frac{\partial}{\partial \theta'} S(\theta_0) \right] (\hat{\theta} - \theta_0).$$

Also

$$S(\hat{\theta}) - S(\theta_0) \approx \left[ \frac{\partial}{\partial \theta'} S(\theta_0) \right] (\hat{\theta} - \theta_0).$$

Therefore

$$-2 \log \lambda(\mathbf{X}) = -2 \log L(\theta_0) + 2 \log L(\hat{\theta}) \approx [\sqrt{N}(\hat{\theta} - \theta_0)]' \left[ -\frac{1}{N} \frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_0) \right] [\sqrt{N}(\hat{\theta} - \theta_0)].$$

Assuming  $H_0$ , we have that  $\sqrt{N}(\hat{\theta} - \theta_0) \Rightarrow N(\mathbf{0}, \mathbf{I}(\theta_0)^{-1})$  and by the LLN,  $-\frac{1}{N} \frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_0)$  converges in probability to  $\mathbf{I}(\theta_0)$ . It follows that under  $H_0$ , the statistic  $-2 \log \lambda(\mathbf{X})$  converges in distribution to  $\chi_m^2$ .

In general  $-2 \log \lambda(\mathbf{X}) \Rightarrow \chi_{q-k}^2$  under  $H_0$ , where  $q = \dim \Theta$  and  $k = \dim \Theta_0$ .

### Power of the LR test under local alternatives

Suppose the following so-called parameter drift (local alternatives) for testing  $H_0 : \theta = \theta_0$ :  $\theta_{0,N} = \theta_0 + N^{-1/2} \mathbf{b}$ , where  $\mathbf{b} \in \mathbb{R}^m$  is a fixed vector. Then

$$\sqrt{N}(\hat{\theta} - \theta_0) = \sqrt{N}(\hat{\theta} - \theta_{0,N}) + \sqrt{N}(\theta_{0,N} - \theta_0) = \sqrt{N}(\hat{\theta} - \theta_{0,N}) + \mathbf{b} \Rightarrow N(\mathbf{b}, \mathbf{I}(\theta_0)^{-1}).$$

Hence

$$-2 \log \lambda \approx [\sqrt{N}(\hat{\theta} - \theta_0)]' \mathbf{I}(\theta_0) [\sqrt{N}(\hat{\theta} - \theta_0)]$$

converges in distribution to noncentral chi-square  $\chi_m^2(\delta)$  with the noncentrality parameter  $\delta = \mathbf{b}' \mathbf{I}(\theta_0) \mathbf{b}$ .

## 8.2 Testing equality constraints

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

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

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

Wald test statistic for testing (linear<sup>1</sup>) equality constraints  $H_0 : \mathbf{A}\theta = \mathbf{c}$  against  $H_1 : \mathbf{A}\theta \neq \mathbf{c}$  is

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

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

$$\sqrt{N} \mathbf{A}(\hat{\theta} - \theta_0) = \sqrt{N}(\mathbf{A}\hat{\theta} - \mathbf{c}) \Rightarrow N(\mathbf{0}, \mathbf{A}\mathbf{I}(\hat{\theta})^{-1} \mathbf{A}').$$

It follows that under  $H_0$  the Wald test statistic converges in distribution to  $\chi_q^2$ .

Note that the LRT

$$2[\log L(\hat{\theta}) - L(\tilde{\theta})] \approx \inf_{\mathbf{A}\theta=\mathbf{c}} [\sqrt{N}(\hat{\theta} - \theta)]' \left[ -\frac{1}{N} \frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_0) \right] [\sqrt{N}(\hat{\theta} - \theta)],$$

---

<sup>1</sup>For nonlinear constraints we can use  $\mathbf{A} = \partial \mathbf{a}(\hat{\theta}) / \partial \theta'$ .

and under  $H_0$ ,  $-\frac{1}{N} \frac{\partial^2}{\partial \theta \partial \theta'} \log L(\theta_0) \approx \mathbf{I}(\theta_0)$ . Therefore under  $H_0$ , the LR and Wald test statistics are asymptotically equivalent.

Score function test statistic for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \neq \theta_0$ :

$$N^{-1} S(\theta_0)' \mathbf{I}(\theta_0)^{-1} S(\theta_0).$$

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

$$N^{-1} S(\theta_0)' \mathbf{I}(\theta_0)^{-1} S(\theta_0) \Rightarrow \chi_m^2.$$

In general, under  $H_0$ ,

$$N^{-1} S(\tilde{\theta})' \mathbf{I}(\tilde{\theta})^{-1} S(\tilde{\theta}) \Rightarrow \chi_q^2$$

when testing  $q$  equality constraints.

## 9 Multinomial distribution

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

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

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

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

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

If  $\mathbf{Y} \sim \text{Mult}(N, \mathbf{p})$ , then the covariance matrix  $\text{Cov}(\mathbf{Y}) = N\mathbf{C}$ , where  $c_{ii} = p_i(1 - p_i)$ ,  $i = 1, \dots, m$  and  $c_{ij} = -p_i p_j$ ,  $i \neq j$ . This can be written as  $\mathbf{C} = \mathbf{P} - \mathbf{p}\mathbf{p}'$ , where  $\mathbf{P} = \text{diag}(p_1, \dots, p_m)$  and  $\mathbf{p} = (p_1, \dots, p_m)'$ . Note that  $\mathbf{C}\mathbf{1}_m = \mathbf{0}$  and  $\text{rank}(\mathbf{C}) = m - 1$ .

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

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

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

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

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

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

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

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

Moreover

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

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

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

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

We can write this statistic as

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

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

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

It follows that  $N(\hat{\mathbf{p}} - \mathbf{p}^*)' \mathbf{Q}(\hat{\mathbf{p}} - \mathbf{p}^*) \Rightarrow \chi_{m-1}^2$ .

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

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

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

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

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

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

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

where  $\tilde{\boldsymbol{\theta}}$  is the MLE under  $H_0$ . Asymptotics of the LR test is  $-2 \log \lambda \Rightarrow \chi_{m-1-q}^2$ .

## 10 Logistic regression

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

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki})}, \quad i = 1, \dots, n. \quad (10.1)$$

That is

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

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

We have that

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

It follows that the likelihood function here is

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

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  $\log L(\boldsymbol{\pi}; \mathbf{y})$  can be written as

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

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

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

$$\mathbf{W} = \text{diag}\{m_i \pi_i (1 - \pi_i)\}.$$

Indeed, we have that

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

and hence

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

where

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

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

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

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

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

and

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

It follows that

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

The maximum likelihood (ML) equations are

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

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

Consider

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

and the following so-called deviance function,

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

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

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

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

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

## 11 Exponential family of distributions

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

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

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

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

Let us show that

$$\mathbb{E}_{\boldsymbol{\theta}}(T_j) = \frac{\partial}{\partial \theta_j} A(\boldsymbol{\theta}), \tag{11.1}$$

$$\text{Cov}(T_j, T_k) = \frac{\partial^2}{\partial \theta_j \partial \theta_k} A(\boldsymbol{\theta}). \quad (11.2)$$

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

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

and hence

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

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

## 12 Generalized linear models

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

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

$$\eta_i = \mathbf{x}_i' \boldsymbol{\beta}, \quad i = 1, \dots, n. \quad (12.1)$$

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

$$\eta_i = g(\mu_i), \quad i = 1, \dots, n, \quad (12.2)$$

and

$$\mu_i = g^{-1}(\mathbf{x}_i' \boldsymbol{\beta}), \quad i = 1, \dots, n. \quad (12.3)$$

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

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

$$f_Y(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\} \quad (12.4)$$

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

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

in the form (12.4) with  $\theta = \mu$ ,  $\phi = \sigma$  and

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

If  $\phi$  is known, then  $a(\phi)$  is viewed as a constant,  $c(y, \phi) = c(y)$ , and (12.4) becomes an exponential family in the canonical form with canonical parameter  $\theta$ .

Consider

$$l(y; \theta, \phi) = \log f_Y(y; \theta, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi). \quad (12.5)$$

By the standard theory of the ML we have that

$$\mathbb{E}[\partial l / \partial \theta] = 0, \quad (12.6)$$

$$\mathbb{E}[\partial^2 l / \partial \theta^2] = -\mathbb{E}[(\partial l / \partial \theta)^2]. \quad (12.7)$$

Also by (12.5)

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

and because of (12.6),  $\mathbb{E}[Y - b'(\theta)] = 0$ . Thus  $\mathbb{E}[Y] = b'(\theta)$ , that is (compare with (11.1))

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

Moreover

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

and hence  $b''(\theta) / a(\phi) = \text{Var}(Y) / a^2(\phi)$  and thus (compare with (11.2))

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

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

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

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

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

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

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

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  $\phi$  is known) the model is  $\theta_i = \eta_i$ ,  $i = 1, \dots, n$ , with  $\eta_i$  being linear predictors specified in equation (12.1). In order to compute the ML estimate of  $\beta$  we need to maximize the corresponding log-likelihood function (given in (12.5)), that is to solve the problem

$$\max_{\beta} \sum_{i=1}^n Y_i \mathbf{x}'_i \beta - b(\mathbf{x}'_i \beta). \quad (12.8)$$

When  $b(\cdot)$  is a convex function, the above problem (12.8) is convex. For the binomial and Poisson distributions the corresponding functions  $b(\cdot)$  are convex.



### 13 Moments structural models

Let  $\mathbf{X}_1, \dots, \mathbf{X}_N$  be an iid sample from  $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Let us show that the ML estimators of  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  are  $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$  and  $\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})'(\mathbf{X}_i - \bar{\mathbf{X}})$ . Note that  $\hat{\boldsymbol{\Sigma}} = \frac{n}{N} \mathbf{S}$ , where  $n = N - 1$  and  $\mathbf{S}$  is the sample covariance matrix.

The likelihood function is

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

Up to the constant we can write logarithm  $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  of this function as

$$\begin{aligned} \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{1}{2} N \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^N (\mathbf{X}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) \\ &= -\frac{1}{2} N \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^N \text{tr} \left[ \boldsymbol{\Sigma}^{-1} (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})' \right] \\ &= -\frac{1}{2} N \log |\boldsymbol{\Sigma}| - \frac{1}{2} \text{tr} \boldsymbol{\Sigma}^{-1} \left[ \sum_{i=1}^N (\mathbf{X}_i - \boldsymbol{\mu}) (\mathbf{X}_i - \boldsymbol{\mu})' \right] \\ &= -\frac{1}{2} N \log |\boldsymbol{\Sigma}| - \frac{1}{2} \text{tr} \left\{ \boldsymbol{\Sigma}^{-1} \left[ \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}}) (\mathbf{X}_i - \bar{\mathbf{X}})' + N(\bar{\mathbf{X}} - \boldsymbol{\mu}) (\bar{\mathbf{X}} - \boldsymbol{\mu})' \right] \right\} \\ &= -\frac{1}{2} N \log |\boldsymbol{\Sigma}| - \frac{1}{2} \text{tr} \left[ \boldsymbol{\Sigma}^{-1} \mathbf{A} \right] - \frac{1}{2} N (\bar{\mathbf{X}} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}). \end{aligned} \tag{13.1}$$

where  $\mathbf{A} = \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}}) (\mathbf{X}_i - \bar{\mathbf{X}})'$ .

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

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

Note that function  $f(\lambda) = \lambda - N \log \lambda$  has unique minimum at  $\lambda = N$ . It follows that  $\hat{\boldsymbol{\Sigma}}^{-1} \mathbf{A} = N \mathbf{I}_m$ , and hence  $\hat{\boldsymbol{\Sigma}} = N^{-1} \mathbf{A}$ .

Consider now a model  $\boldsymbol{\mu} \in \Omega$  for the mean vector. The set  $\Omega \subset \mathbb{R}^m$  can be defined by equations or can be given in a parametric form. The Likelihood Ratio (LR) test statistic for testing  $H_0 : \boldsymbol{\mu} \in \Omega$  against the unrestricted alternative  $H_1 : \boldsymbol{\mu} \in \mathbb{R}^m$  is

$$\lambda^{-1} = \frac{\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma} \succ 0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\max_{\boldsymbol{\mu} \in \Omega, \boldsymbol{\Sigma} \succ 0} L(\boldsymbol{\mu}, \boldsymbol{\Sigma})}.$$

That is

$$-2 \log \lambda = 2\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) - 2 \max_{\boldsymbol{\mu} \in \Omega, \boldsymbol{\Sigma} \succ 0} \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Now

$$2\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = -N \log |\hat{\boldsymbol{\Sigma}}| - Nm. \tag{13.2}$$

Also the maximum of  $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  over  $\boldsymbol{\Sigma} \succ \mathbf{0}$  is attained at  $\boldsymbol{\Sigma}^* = \hat{\boldsymbol{\Sigma}} + (\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'$ . Hence

$$\begin{aligned} -2 \log \lambda &= 2\ell(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) - 2 \max_{\boldsymbol{\mu} \in \Omega} \ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}^*) \\ &= -N \log |\hat{\boldsymbol{\Sigma}}| - Nm + N \inf_{\boldsymbol{\mu} \in \Omega} \log |\hat{\boldsymbol{\Sigma}} + (\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'| + Nm \\ &= N \inf_{\boldsymbol{\mu} \in \Omega} \log |\mathbf{I}_m + \hat{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'| \\ &= N \inf_{\boldsymbol{\mu} \in \Omega} \log (1 + (\bar{\mathbf{X}} - \boldsymbol{\mu})' \hat{\boldsymbol{\Sigma}}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu})). \end{aligned}$$

The last equality holds since matrix  $\hat{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})(\bar{\mathbf{X}} - \boldsymbol{\mu})'$  has only one nonzero eigenvalue equal to  $(\bar{\mathbf{X}} - \boldsymbol{\mu})'\hat{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu})$ .

That is, the LRT rejects  $H_0$  for large values of the statistic

$$T^2 = N \inf_{\boldsymbol{\mu} \in \Omega} (\bar{\mathbf{X}} - \boldsymbol{\mu})' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}). \quad (13.3)$$

Suppose that the model is given in the parametric form  $\boldsymbol{\mu} = \mathbf{g}(\boldsymbol{\theta})$ ,  $\boldsymbol{\theta} \in \Theta$ . Here  $\Theta \subset \mathbb{R}^q$  is the parameter space and  $\mathbf{g} = (g_1, \dots, g_m) : \mathbb{R}^q \rightarrow \mathbb{R}^m$ . Then

$$T^2 = N \inf_{\boldsymbol{\theta} \in \Theta} (\bar{\mathbf{X}} - \mathbf{g}(\boldsymbol{\theta}))' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \mathbf{g}(\boldsymbol{\theta})). \quad (13.4)$$

Let  $\boldsymbol{\mu}_0$  be the true (population) value of the mean vector. The model is correct if there is  $\boldsymbol{\theta}_0 \in \Theta$  such that  $\boldsymbol{\mu}_0 = \mathbf{g}(\boldsymbol{\theta}_0)$ . It is said that the model is *identified* at  $\boldsymbol{\theta}_0 \in \Theta$  if  $\mathbf{g}(\boldsymbol{\theta}_0) = \mathbf{g}(\boldsymbol{\theta}^*)$ ,  $\boldsymbol{\theta}^* \in \Theta$ , implies that  $\boldsymbol{\theta}_0 = \boldsymbol{\theta}^*$ .

We assume that functions  $g_i(\boldsymbol{\theta})$ ,  $i = 1, \dots, m$ , are continuously differentiable and denote  $\boldsymbol{\Delta}(\boldsymbol{\theta}) = \partial \mathbf{g}(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}'$  the  $m \times q$  Jacobian matrix of partial derivatives. In particular if the model is linear,  $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} + \mathbf{a}$ , where  $\mathbf{A}$  is a known  $m \times q$  matrix of rank  $q$  and  $\mathbf{a}$  is  $m \times 1$  vector, then  $\boldsymbol{\Delta}(\boldsymbol{\theta}) = \mathbf{A}$  for all  $\boldsymbol{\theta}$ . Note that by the Law of Large Numbers,  $\bar{\mathbf{X}}$  converges w.p.1 to the true value  $\boldsymbol{\mu}_0$  of the mean vector and  $\mathbf{S}$  converges w.p.1 to the covariance matrix  $\boldsymbol{\Sigma}$ . Moreover, by the Central Limit Theorem  $N^{1/2}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)$  converges in distribution to normal  $N(\mathbf{0}, \boldsymbol{\Sigma})$ .

Let  $\hat{\boldsymbol{\theta}}$  be the minimizer of the right hand side of (13.4). It is possible to show that  $\hat{\boldsymbol{\theta}}$  converges w.p.1 to the minimizer of  $(\boldsymbol{\mu}_0 - \mathbf{g}(\boldsymbol{\theta}))' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_0 - \mathbf{g}(\boldsymbol{\theta}))$  over  $\boldsymbol{\theta} \in \Theta$ , provided this minimizer is unique. In particular if  $\boldsymbol{\mu}_0 = \mathbf{g}(\boldsymbol{\theta}_0)$ , for some  $\boldsymbol{\theta}_0 \in \Theta$ , and the model is identified at  $\boldsymbol{\theta}_0$ , then  $\hat{\boldsymbol{\theta}}$  converges w.p.1 to  $\boldsymbol{\theta}_0$ .

Suppose that the null hypothesis  $H_0 : \boldsymbol{\mu}_0 = \mathbf{g}(\boldsymbol{\theta}_0)$  holds, the model is identified at  $\boldsymbol{\theta}_0$ ,  $\boldsymbol{\theta}_0$  is an interior point of the parameter set  $\Theta$ , and denote  $\boldsymbol{\Delta} = \boldsymbol{\Delta}(\boldsymbol{\theta}_0)$ . Then

$$N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (\boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1} [N^{1/2}(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)] + o_p(1) \quad (13.5)$$

and

$$T^2 = N(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{V}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0) + o_p(1), \quad (13.6)$$

where  $\mathbf{V} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta} (\boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1}$ .

We give arguments for the linear model  $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{A}\boldsymbol{\theta} + \mathbf{a}$ . Then  $\hat{\boldsymbol{\theta}}$  is the minimizer of

$$(\bar{\mathbf{X}} - \mathbf{a} - \mathbf{A}\boldsymbol{\theta})' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \mathbf{a} - \mathbf{A}\boldsymbol{\theta}).$$

Hence

$$\hat{\boldsymbol{\theta}} = (\mathbf{A}' \mathbf{S}^{-1} \mathbf{A})^{-1} \mathbf{A}' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \mathbf{a}).$$

Also since  $\boldsymbol{\mu}_0 = \mathbf{A}\boldsymbol{\theta}_0 + \mathbf{a}$ , we have that  $\boldsymbol{\theta}_0 = (\mathbf{A}' \mathbf{S}^{-1} \mathbf{A})^{-1} \mathbf{A}' \mathbf{S}^{-1} (\boldsymbol{\mu}_0 - \mathbf{a})$ . Hence since  $\mathbf{A} = \boldsymbol{\Delta}$ ,

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = (\boldsymbol{\Delta}' \mathbf{S}^{-1} \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta}' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0).$$

Since  $\mathbf{S}$  converges w.p.1 to  $\boldsymbol{\Sigma}$ , this implies (13.5). For a not necessary linear model we can use the approximation  $\mathbf{g}(\boldsymbol{\theta}) \approx \mathbf{g}(\boldsymbol{\theta}_0) + \boldsymbol{\Delta}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ .

It follows from (13.5) that  $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  converges in distribution to normal with zero mean vector and covariance matrix

$$(\boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta})^{-1} \boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta} (\boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta})^{-1} = (\boldsymbol{\Delta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta})^{-1}.$$

It follows from (13.6) that the asymptotic distribution of  $T^2$  is the same as the distribution of the quadratic form  $\mathbf{Y}' \mathbf{V} \mathbf{Y}$ , where  $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ . Consequently under  $H_0$ , statistic  $T^2$  converges in distribution to  $\chi_{m-q}^2$ .

### 13.1 Covariance structural models

Let  $\mathbf{X}_1, \dots, \mathbf{X}_N$  be an iid sample with mean  $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}_i]$  and  $m \times m$  covariance matrix  $\text{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma}$ . A covariance structural model is  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ , where  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  is a symmetric matrix valued function of parameter vector  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ . It is assumed that for every  $\boldsymbol{\theta} \in \Theta$  the matrix  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  is positive semidefinite.

Examples of covariance structures:

- (i)  $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0$ , where  $\boldsymbol{\Sigma}_0$  is a given positive definite matrix.
- (ii)  $\boldsymbol{\Sigma} = \theta \mathbf{I}_m$ ,  $\theta \geq 0$ , sphericity model.
- (iii)  $\boldsymbol{\Sigma} = \text{diag}(\theta_1, \dots, \theta_m)$ ,  $\boldsymbol{\theta} \in \mathbb{R}_+^m$ , uncorrelated variables.
- (iv)  $\boldsymbol{\Sigma}(\boldsymbol{\theta})$  is such that all its diagonal elements are equal to  $\theta$  and off diagonal elements are unrestricted. Equal variances model.

These are examples of linear models

$$\boldsymbol{\Sigma} = \theta_1 \mathbf{G}_1 + \dots + \theta_q \mathbf{G}_q,$$

where  $\mathbf{G}_1, \dots, \mathbf{G}_q$  are known symmetric matrices and

$$\Theta = \{\boldsymbol{\theta} \in \mathbb{R}^q : \theta_1 \mathbf{G}_1 + \dots + \theta_q \mathbf{G}_q \text{ is positive semidefinite}\}.$$

#### Factor analysis model

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

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda} \mathbf{f} + \mathbf{u}, \quad (13.7)$$

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

It follows then that

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{\Lambda} \mathbf{f} + \mathbf{u})(\boldsymbol{\Lambda} \mathbf{f} + \mathbf{u})'] = \boldsymbol{\Lambda} \boldsymbol{\Phi} \boldsymbol{\Lambda}' + \boldsymbol{\Psi}, \quad (13.8)$$

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

$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda} \boldsymbol{\Lambda}' + \boldsymbol{\Psi}. \quad (13.9)$$

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

#### Testing of covariance structural models

Let  $\mathbf{X}_1, \dots, \mathbf{X}_N$  be an iid sample from  $N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Then the log-likelihood ratio of testing  $H_0 : \boldsymbol{\Sigma}_0 \in \Omega$  against unrestricted alternative is

$$N \min_{\boldsymbol{\Sigma} \in \Omega} F(\hat{\boldsymbol{\Sigma}}, \boldsymbol{\Sigma}), \quad (13.10)$$

where

$$F(\mathbf{S}, \mathbf{\Sigma}) = \log |\mathbf{\Sigma}| - \log |\mathbf{S}| + \text{tr}(\mathbf{S}\mathbf{\Sigma}^{-1}) - m \quad (13.11)$$

and

$$\hat{\mathbf{\Sigma}} = \frac{1}{N} \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' = \frac{1}{N} \mathbf{A}$$

is the MLE of  $\mathbf{\Sigma}_0$ .

**Proof** The log-likelihood ratio statistic here is (see (13.1))

$$-2 \log \lambda = 2\ell(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) - 2 \max_{\mathbf{\Sigma} \in \Omega, \boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \mathbf{\Sigma}),$$

where

$$\ell(\boldsymbol{\mu}, \mathbf{\Sigma}) = -\frac{1}{2}N \log |\mathbf{\Sigma}| - \frac{1}{2}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{A}) - \frac{1}{2}N(\bar{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}).$$

Now (see (13.2))

$$2\ell(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) = -N \log |\hat{\mathbf{\Sigma}}| - Nm,$$

and

$$\max_{\boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \mathbf{\Sigma}) = -\frac{1}{2}N \log |\mathbf{\Sigma}| - \frac{1}{2}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{A}).$$

Hence

$$-2 \max_{\mathbf{\Sigma} \in \Omega, \boldsymbol{\mu}} \ell(\boldsymbol{\mu}, \mathbf{\Sigma}) = \min_{\mathbf{\Sigma} \in \Omega} \{N \log |\mathbf{\Sigma}| + \text{tr}(\mathbf{\Sigma}^{-1}\mathbf{A})\}.$$

It follows that

$$-2 \log \lambda = N \min_{\mathbf{\Sigma} \in \Omega} \{ \log |\mathbf{\Sigma}| - \log |\hat{\mathbf{\Sigma}}| + \text{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}) - m \}. \quad (13.12)$$

■

The function  $F(\mathbf{S}, \mathbf{\Sigma})$  has the following properties:  $F(\mathbf{S}, \mathbf{\Sigma}) \geq 0$  and  $F(\mathbf{S}, \mathbf{\Sigma}) = 0$  if and only if  $\mathbf{S} = \mathbf{\Sigma}$ . Consider  $m^2 \times 1$  vectors  $\boldsymbol{\sigma} = \text{vec}(\mathbf{\Sigma})$  and  $\mathbf{s} = \text{vec}(\mathbf{S})$ . We can consider  $F(\mathbf{s}, \boldsymbol{\sigma})$  as a function of  $\mathbf{s}$  and  $\boldsymbol{\sigma}$ . Since for given (fixed)  $\boldsymbol{\sigma}$  the function  $F(\mathbf{s}, \boldsymbol{\sigma})$  attains its minimum (of 0) at  $\mathbf{s} = \boldsymbol{\sigma}$  it follows that

$$\left. \frac{\partial F(\mathbf{s}, \boldsymbol{\sigma})}{\partial \mathbf{s}} \right|_{\mathbf{s}=\boldsymbol{\sigma}} = \mathbf{0}.$$

Similarly

$$\left. \frac{\partial F(\mathbf{s}, \boldsymbol{\sigma})}{\partial \boldsymbol{\sigma}} \right|_{\boldsymbol{\sigma}=\mathbf{s}} = \mathbf{0}.$$

Let us show that the second order Taylor expansion of  $F(\mathbf{s}, \boldsymbol{\sigma})$  at  $(\mathbf{s}, \boldsymbol{\sigma}) = (\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0)$  is

$$F(\mathbf{s}, \boldsymbol{\sigma}) = \frac{1}{2}(\mathbf{s} - \boldsymbol{\sigma})'\mathbf{\Sigma}_0^{-1} \otimes \mathbf{\Sigma}_0^{-1}(\mathbf{s} - \boldsymbol{\sigma}) + o(\|(\mathbf{s} - \boldsymbol{\sigma}_0, \boldsymbol{\sigma} - \boldsymbol{\sigma}_0)\|^2). \quad (13.13)$$

**Proof** Note that for ‘small’ matrix  $\mathbf{H}$ ,

$$|\mathbf{I}_m + \mathbf{H}| = 1 + \text{tr}\mathbf{H} + o(\|\mathbf{H}\|),$$

and hence

$$\log |\mathbf{I}_m + \mathbf{H}| = \text{tr}\mathbf{H} + o(\|\mathbf{H}\|).$$

It follows that

$$\log |\mathbf{\Sigma} + \mathbf{H}| = \log |\mathbf{\Sigma}| + \log |\mathbf{I}_m + \mathbf{\Sigma}^{-1}\mathbf{H}| = \log |\mathbf{\Sigma}| + \text{tr}(\mathbf{\Sigma}^{-1}\mathbf{H}) + o(\|\mathbf{H}\|).$$

We can write this as

$$d \log |\Sigma| = \text{tr}(\Sigma^{-1} d\Sigma). \quad (13.14)$$

Also

$$(\Sigma + d\Sigma)^{-1} = [\Sigma(\mathbf{I} + \Sigma^{-1} d\Sigma)]^{-1} = (\mathbf{I} + \Sigma^{-1} d\Sigma)^{-1} \Sigma^{-1} = \Sigma^{-1} - \Sigma^{-1} (d\Sigma) \Sigma^{-1},$$

and hence

$$d[\text{tr}(\mathbf{S} \Sigma^{-1})] = -\text{tr}[\mathbf{S} \Sigma^{-1} (d\Sigma) \Sigma^{-1}] = -\text{tr}[\Sigma^{-1} \mathbf{S} \Sigma^{-1} d\Sigma]. \quad (13.15)$$

It follows that the differential of  $F$  with respect to  $\Sigma$  is

$$dF = \text{tr} [\Sigma^{-1} (\Sigma - \mathbf{S}) \Sigma^{-1} d\Sigma]. \quad (13.16)$$

Since

$$\text{tr} [\Sigma^{-1} (\Sigma - \mathbf{S}) \Sigma^{-1} d\Sigma] = \text{vec} [\Sigma^{-1} (\Sigma - \mathbf{S}) \Sigma^{-1}]' d\sigma,$$

using (6.2) we can write (13.16) in the following equivalent form

$$dF = [\Sigma^{-1} \otimes \Sigma^{-1} (\sigma - s)]' d\sigma. \quad (13.17)$$

It follows that

$$\left. \frac{\partial^2 F}{\partial \sigma' \partial s} \right|_{s=\sigma} = -\Sigma^{-1} \otimes \Sigma^{-1}. \quad (13.18)$$

Finally let us show that the Hessian matrices

$$\left. \frac{\partial^2 F}{\partial \sigma' \partial \sigma} \right|_{s=\sigma}, \quad -\left. \frac{\partial^2 F}{\partial \sigma' \partial s} \right|_{s=\sigma} \quad \text{and} \quad \left. \frac{\partial^2 F}{\partial s' \partial s} \right|_{s=\sigma} \quad (13.19)$$

are equal to each other. Let  $\sigma(s) = s$  and  $h(s) = F(s, \sigma(s)) = F(s, s)$ . We have that  $h(s) \equiv 0$  and by the chain rule of differentiation

$$\frac{\partial h(s)}{\partial s} = \left. \frac{\partial F(s, \sigma)}{\partial s} \right|_{\sigma=s} + \left. \frac{\partial F(s, \sigma)}{\partial \sigma} \right|_{\sigma=s} \frac{\partial \sigma(s)}{\partial s}.$$

Since  $\partial h(s)/\partial s = 0$  and  $\partial \sigma(s)/\partial s = \mathbf{I}$ , it follows that

$$\left. \frac{\partial F(s, \sigma)}{\partial \sigma} \right|_{\sigma=s} = -\left. \frac{\partial F(s, \sigma)}{\partial s} \right|_{\sigma=s},$$

and hence

$$\left. \frac{\partial^2 F(s, \sigma)}{\partial \sigma' \partial \sigma} \right|_{s=\sigma} = -\left. \frac{\partial^2 F(s, \sigma)}{\partial \sigma' \partial s} \right|_{s=\sigma} = \left. \frac{\partial^2 F(s, \sigma)}{\partial s' \partial s} \right|_{s=\sigma}.$$

Now since  $F(\sigma_0, \sigma_0) = 0$  and  $\partial F(\sigma_0, \sigma_0)/\partial s = 0$ ,  $\partial F(\sigma_0, \sigma_0)/\partial \sigma = 0$ , the second order Taylor expansion at  $(\sigma_0, \sigma_0)$  gives

$$\begin{aligned} F(s, \sigma) &= \frac{1}{2}(s - \sigma_0)' \mathbf{Q}_{11}(s - \sigma_0) + (s - \sigma_0)' \mathbf{Q}_{12}(\sigma - \sigma_0) + \frac{1}{2}(\sigma - \sigma_0)' \mathbf{Q}_{22}(\sigma - \sigma_0) \\ &\quad + o(\|(s - \sigma_0, \sigma - \sigma_0)\|^2), \end{aligned}$$

where

$$\mathbf{Q}_{11} = \left. \frac{\partial^2 F(s, \sigma)}{\partial s' \partial s} \right|_{s=\sigma=\sigma_0}, \quad \mathbf{Q}_{12} = \left. \frac{\partial^2 F(s, \sigma)}{\partial s' \partial \sigma} \right|_{s=\sigma=\sigma_0}, \quad \mathbf{Q}_{22} = \left. \frac{\partial^2 F(s, \sigma)}{\partial \sigma' \partial \sigma} \right|_{s=\sigma=\sigma_0}.$$

By (13.18) and (13.19) we have that  $\mathbf{Q}_{11} = \mathbf{Q}_{22} = -\mathbf{Q}_{12} = \Sigma_0^{-1} \otimes \Sigma_0^{-1}$ , and hence (13.13) follows. ■

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

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

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

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

and hence

$$[\boldsymbol{\Gamma}]_{ij,kl} = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}. \quad (13.20)$$

In a matrix form equations (13.20) can be written as

$$\boldsymbol{\Gamma} = 2\mathbf{M}_m(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0), \quad (13.21)$$

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

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

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

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

It follows that

$$\boldsymbol{\Gamma} = 2\mathbf{M}_m(\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0)\mathbf{M}_m. \quad (13.22)$$

By using above derivations it is possible to show that if the null hypothesis  $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$  is correct and the model is identified at  $\boldsymbol{\theta}_0 \in \Theta$ , then (under mild regularity conditions) the test statistic  $N \min_{\boldsymbol{\theta} \in \Theta} F(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$  converges in distribution to the quadratic form  $\mathbf{Y}'\mathbf{W}\mathbf{Y}$ , where  $\mathbf{Y} \sim N_{m^2}(\mathbf{0}, \boldsymbol{\Gamma})$ ,

$$\mathbf{W} = \mathbf{V} - \mathbf{V}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{V}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{V} \text{ and } \mathbf{V} = \frac{1}{2}(\boldsymbol{\Sigma}_0^{-1} \otimes \boldsymbol{\Sigma}_0^{-1}),$$

and  $\boldsymbol{\Delta} = \partial\boldsymbol{\sigma}(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}'$  is that the  $m^2 \times q$  Jacobian matrix. Note that a sufficient condition for local identification at  $\boldsymbol{\theta}_0$  is that the Jacobian matrix  $\boldsymbol{\Delta}$  has full column rank  $q$ . In case of normally distributed population, when  $\boldsymbol{\Gamma}$  is of the form (13.22), it follows that under  $H_0$ , the test statistic  $N \min_{\boldsymbol{\theta} \in \Theta} F(\mathbf{S}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$  tends in distribution to chi-square with  $m(m+1)/2 - q$  degrees of freedom.

Also it follows that

$$N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = (\boldsymbol{\Delta}'\mathbf{V}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{V} \left[ N^{1/2}(\mathbf{s} - \boldsymbol{\sigma}_0) \right] + o_p(1).$$

Hence  $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$  converges in distribution to normal with zero mean vector and the covariance matrix

$$(\boldsymbol{\Delta}'\mathbf{V}\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}'\mathbf{V}\boldsymbol{\Gamma}\mathbf{V}\boldsymbol{\Delta}(\boldsymbol{\Delta}'\mathbf{V}\boldsymbol{\Delta})^{-1}.$$

In particular if the population has normal distribution, then this covariance matrix becomes  $(\boldsymbol{\Delta}'\mathbf{V}\boldsymbol{\Delta})^{-1}$ .

## 14 Principal components analysis

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

$$\boldsymbol{\Sigma} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}' = \lambda_1\mathbf{e}_1\mathbf{e}_1' + \dots + \lambda_m\mathbf{e}_m\mathbf{e}_m', \quad (14.1)$$

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

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

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

By (14.1) we have that

$$\mathbf{w}'\boldsymbol{\Sigma}\mathbf{w} = \mathbf{w}'\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'\mathbf{w} = \mathbf{v}'\boldsymbol{\Lambda}\mathbf{v} = \lambda_1v_1^2 + \dots + \lambda_mv_m^2.$$

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

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

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

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

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

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

this means that is we want to solve the problem

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

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

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

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

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

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

$$(\mathbf{A} + d\mathbf{A})(\mathbf{e}_i + d\mathbf{e}_i) = (\lambda_i + d\lambda_i)(\mathbf{e}_i + d\mathbf{e}_i) \quad (14.4)$$

and

$$(\mathbf{A} + d\mathbf{A})(\mathbf{e}_i + d\mathbf{e}_i) = \mathbf{A}\mathbf{e}_i + (d\mathbf{A})\mathbf{e}_i + \mathbf{A}d\mathbf{e}_i + (d\mathbf{A})d\mathbf{e}_i. \quad (14.5)$$

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

$$(d\mathbf{A})\mathbf{e}_i + \mathbf{A}d\mathbf{e}_i = (d\lambda_i)\mathbf{e}_i + \lambda_i d\mathbf{e}_i. \quad (14.6)$$

Furthermore up to high order terms

$$(\mathbf{e}_i + d\mathbf{e}_i)'(\mathbf{e}_j + d\mathbf{e}_j) = (d\mathbf{e}_i)'\mathbf{e}_j + \mathbf{e}_i'd\mathbf{e}_j + \mathbf{e}_i'\mathbf{e}_j. \quad (14.7)$$

It follows that for  $i = j$

$$\mathbf{e}_i'd\mathbf{e}_i = 0, \quad (14.8)$$

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

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

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

$$d\lambda_i = \mathbf{e}_i'(d\mathbf{A})\mathbf{e}_i. \quad (14.10)$$

It is also possible to write (14.10) as

$$d\lambda_i = \text{tr}(\mathbf{e}_i\mathbf{e}_i'd\mathbf{A}). \quad (14.11)$$

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

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

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

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

This implies that

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

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

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

Now let  $\mathbf{X}_1, \dots, \mathbf{X}_N$  be an iid sample from  $N_m(\mu, \Sigma)$  and  $\mathbf{S}$  be the corresponding sample covariance matrix. Let  $\lambda_1 \geq \dots \geq \lambda_m$  be the eigenvalues and  $\mathbf{e}_1, \dots, \mathbf{e}_m$  be a corresponding



set of orthonormal eigenvectors of  $\mathbf{\Sigma}$ , and  $\ell_1 \geq \dots \geq \ell_m$  be the eigenvalues and  $\mathbf{q}_1, \dots, \mathbf{q}_m$  be a corresponding set of orthonormal eigenvectors of  $\mathbf{S}$ .

Suppose that  $\lambda_i$  has multiplicity one. Let us show that  $N^{1/2}(\ell_i - \lambda_i)$  and  $N^{1/2}(\mathbf{q}_i - \mathbf{e}_i)$  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  $N^{1/2}(\mathbf{q}_i - \mathbf{e}_i)$  is

$$\sum_{j=1, j \neq i}^m \frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \mathbf{e}_j \mathbf{e}_j'. \quad (14.16)$$

By the Delta Theorem 7.2 and (14.10) we have that

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

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

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

and hence

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

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

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

Similarly, by (14.15)

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

where

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

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

$$\frac{[\text{vec}(\mathbf{e}_i \mathbf{e}_i')]' \mathbf{\Gamma} [\text{vec}(\mathbf{e}_i \mathbf{e}_i')]}{(\lambda_i - \lambda_j)^2} = \frac{2\text{tr}[(\mathbf{e}_i \mathbf{e}_i') \mathbf{M}_m \mathbf{\Sigma} (\mathbf{e}_i \mathbf{e}_i') \mathbf{M}_m \mathbf{\Sigma}]}{(\lambda_i - \lambda_j)^2}. \quad (14.17)$$

Also  $\mathbf{M}_m(\mathbf{e}_i \mathbf{e}_j') = \frac{1}{2}[(\mathbf{e}_i \mathbf{e}_j') + (\mathbf{e}_j \mathbf{e}_i')]$  and  $\mathbf{M}_m(\mathbf{e}_i \mathbf{e}_k') = \frac{1}{2}[(\mathbf{e}_i \mathbf{e}_k') + (\mathbf{e}_k \mathbf{e}_i')]$ . It follows that the right hand side of (14.17) is equal to

$$\frac{\text{tr}[(\mathbf{e}_i \mathbf{e}_j') + (\mathbf{e}_j \mathbf{e}_i')] \mathbf{\Sigma} [(\mathbf{e}_i \mathbf{e}_k') + (\mathbf{e}_k \mathbf{e}_i')] \mathbf{\Sigma}}{2(\lambda_i - \lambda_j)^2}. \quad (14.18)$$

Moreover, we have that  $\mathbf{e}_j' \mathbf{\Sigma} \mathbf{e}_k = \lambda_j \mathbf{e}_j' \mathbf{e}_k$  equals 0 if  $j \neq k$ , and  $\lambda_j$  if  $j = k$ . Therefore the expression in (14.18) 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  $a_j \mathbf{e}_j$  is  $\frac{\lambda_i \lambda_j}{(\lambda_i - \lambda_j)^2} \mathbf{e}_j \mathbf{e}_j'$ , and  $a_j \mathbf{e}_j$  is asymptotically uncorrelated with  $a_k \mathbf{e}_k$  for  $j \neq k$ . Formula (14.16) follows.

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

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

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

## 15 Spherical and elliptical distributions

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

### Examples

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

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

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

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

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

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

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

Spherical distributions can be generated in the following way. Let  $X_1, \dots, X_m$  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  $\mathbf{X} = (X_1, \dots, X_m)'$  is

$$f(\mathbf{x}) = \int_0^\infty (2\pi z)^{-m/2} \exp\left(-\frac{1}{2}z^{-1}\mathbf{x}'\mathbf{x}\right) dG(z).$$

This is scale mixture of normal distributions. In particular, if  $Z$  can have two possible values  $\sigma_1^2$  and  $\sigma_2^2$  with respective probabilities  $1 - \varepsilon$  and  $\varepsilon$ , then this is the  $\varepsilon$ -contaminated normal distribution. If  $Z \sim n/\chi_n^2$ , then  $\mathbf{X}$  has  $m$ -variate  $t$ -distribution with  $n$  degrees of freedom.

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

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

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

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

$$\partial\phi(\mathbf{0})/\partial\mathbf{t} = i\mathbb{E}[\mathbf{X}] \tag{15.1}$$

and

$$\partial^2\phi(\mathbf{0})/\partial\mathbf{t}\partial\mathbf{t}' = -\mathbb{E}[\mathbf{X}\mathbf{X}'] = -\boldsymbol{\mu}\boldsymbol{\mu}' - \text{Cov}(\mathbf{X}). \tag{15.2}$$

If  $\mathbf{X}$  has spherical distribution, then for any orthogonal matrix  $\mathbf{T}$  we have

$$\phi(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}'\mathbf{X})] = \mathbb{E}[\exp(i\mathbf{t}'\mathbf{T}\mathbf{X})] = \mathbb{E}[\exp(i(\mathbf{T}'\mathbf{t})'\mathbf{X})] = \phi(\mathbf{T}'\mathbf{t}).$$

It follows that  $\phi(\mathbf{t})$  is a function of  $\mathbf{t}'\mathbf{t}$ , i.e.,

$$\phi(\mathbf{t}) = \psi(\mathbf{t}'\mathbf{t}) \quad (15.3)$$

for some function  $\psi(\cdot)$ .

If  $\mathbf{X} \sim E_m(\boldsymbol{\mu}, \mathbf{V})$ , then  $\mathbf{X} = \boldsymbol{\mu} + \mathbf{V}^{1/2}\mathbf{Y}$ , where  $\mathbf{Y}$  has spherical distribution. Hence the characteristic function of  $\mathbf{X}$  can be written as

$$\phi(\mathbf{t}) = \mathbb{E} \left[ \exp(i\mathbf{t}'(\boldsymbol{\mu} + \mathbf{V}^{1/2}\mathbf{Y})) \right] = \exp(i\mathbf{t}'\boldsymbol{\mu}) \mathbb{E}[\exp(i\mathbf{t}'\mathbf{V}^{1/2}\mathbf{Y})],$$

and since  $\mathbf{Y}$  has spherical distribution we have by (15.3) that

$$\mathbb{E}[\exp(i\mathbf{t}'\mathbf{V}^{1/2}\mathbf{Y})] = \psi((\mathbf{V}^{1/2}\mathbf{t})'(\mathbf{V}^{1/2}\mathbf{t})) = \psi(\mathbf{t}'\mathbf{V}\mathbf{t}).$$

That is, the characteristic function of  $\mathbf{X} \sim E_m(\boldsymbol{\mu}, \mathbf{V})$  can be represented in the form

$$\phi(\mathbf{t}) = \exp(i\mathbf{t}'\boldsymbol{\mu})\psi(\mathbf{t}'\mathbf{V}\mathbf{t}) \quad (15.4)$$

for some function  $\psi(\cdot)$ . It follows from (15.4) together with (15.1) and (15.2), that if  $\mathbf{X} \sim E_m(\boldsymbol{\mu}, \mathbf{V})$ , then  $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$  and  $\text{Cov}(\mathbf{X}) = \alpha\mathbf{V}$ , where  $\alpha = -2\psi'(0)$ . In particular this implies that

$$\text{Corr}(X_i, X_j) = \frac{v_{ij}}{\sqrt{v_{ii}v_{jj}}}. \quad (15.5)$$

Let  $\mathbf{X} \sim E_m(\boldsymbol{\mu}, \mathbf{V})$  be partitioned  $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$  with the corresponding partitioning of  $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$  and  $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix}$ . Then it follows from (15.4) that the characteristic function of  $\mathbf{X}_1$  is

$$\phi_1(\mathbf{t}_1) = \exp(i\mathbf{t}_1'\boldsymbol{\mu}_1)\psi(\mathbf{t}_1'\mathbf{V}_{11}\mathbf{t}_1). \quad (15.6)$$

**Theorem 15.1** *Let  $\mathbf{X} = (X_1, \dots, X_m)'\sim E_m(\boldsymbol{\mu}, \mathbf{V})$  and  $\mathbf{V} = \text{diag}(v_{11}, \dots, v_{mm})$ . If  $X_1, \dots, X_m$  are all independent, then  $\mathbf{X}$  has multivariate normal distribution.*

**Proof.** Without loss of generality we can assume that  $\boldsymbol{\mu} = \mathbf{0}$ , and hence the the characteristic function of  $\mathbf{X}_1$  is

$$\phi(\mathbf{t}) = \psi(\mathbf{t}'\mathbf{V}\mathbf{t}) = \psi\left(\sum_{i=1}^m t_i^2 v_{ii}\right).$$

Denote  $u_i = t_i v_{ii}^{1/2}$ . Since  $X_1, \dots, X_m$  are independent, we have that

$$\phi(\mathbf{t}) = \mathbb{E}[\exp(i\mathbf{t}'\mathbf{X})] = \mathbb{E}\left[\prod_{i=1}^m e^{it_i X_i}\right] = \prod_{i=1}^m \mathbb{E}[e^{it_i X_i}] = \prod_{i=1}^m \phi_i(t_i),$$

where  $\phi_i(t_i)$  is the characteristic function of  $X_i$ ,  $i = 1, \dots, m$ . It follows by (15.6) that

$$\psi\left(\sum_{i=1}^m u_i^2\right) = \prod_{i=1}^m \psi(u_i^2). \quad (15.7)$$

In turn equation (15.7) implies that  $\psi(z) = e^{kz}$  for some  $k$ . That is  $\phi(\mathbf{t}) = \exp(k\mathbf{t}'\mathbf{V}\mathbf{t})$ . ■

## 15.1 Multivariate cumulants

Consider a random variable  $X$ . Let

$$\log \mathbb{E}[e^{tX}] = \sum_{n=1}^{\infty} \kappa_n \frac{t^n}{n!} \quad (15.8)$$

be Taylor expansion of its log-moments generating function (note that for  $t = 0$  this function is 0). The coefficient  $\kappa_n$  is called  $n$ -th cumulant of  $X$ . Since  $\mathbb{E}[e^{tX}]$  may not exist, it is preferable to define cumulants in terms of the characteristic function as

$$\log \mathbb{E}[e^{itX}] = \sum_{n=1}^{\infty} \kappa_n \frac{(it)^n}{n!}. \quad (15.9)$$

We have that

$$\frac{\partial \log \mathbb{E}[e^{tX}]}{\partial t} = \frac{1}{\mathbb{E}[e^{tX}]} \mathbb{E}[X e^{tX}],$$

and hence  $\kappa_1 = \mathbb{E}[X]$ . Denote  $\mu_i = \mathbb{E}[X^i]$  the  $i$ -th moment of  $X$ . Then

$$\begin{aligned} \kappa_1 &= \mu_1 = \mathbb{E}[X], \\ \kappa_2 &= \mu_2 - \mu_1^2 = \text{Var}(X), \\ \kappa_3 &= \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3, \\ \kappa_4 &= \mu_4 - 4\mu_1\mu_3 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4. \end{aligned}$$

Skewness of  $X$  is defined as

$$\gamma_1 = \frac{\kappa_3}{\kappa_2^{3/2}}, \quad (15.10)$$

kurtosis of  $X$  is defined as

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2}. \quad (15.11)$$

If distribution of  $X$  is symmetrical around its mean, then  $\gamma_1 = 0$ . If  $X \sim N(\mu, \sigma^2)$ , then  $\mu_4 = 3\mu^2$ , and hence its kurtosis  $\gamma_2 = 0$ .

If  $X$  and  $Y$  are two independent random variables, then

$$\log \mathbb{E}[e^{it(X+Y)}] = \log \mathbb{E}[e^{itX}] + \log \mathbb{E}[e^{itY}],$$

and hence cumulants of  $X + Y$  are equal to the sum of the respective cumulants of  $X$  and  $Y$ .

Consider now random vector  $\mathbf{X} = (X_1, \dots, X_m)'$ . 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  $\mathbf{X} \sim E_m(\boldsymbol{\mu}, \mathbf{V})$  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  $\mathbf{X} \sim E_m(\boldsymbol{\mu}, \mathbf{V})$  are

$$\kappa_{1111}^{ijkl} = \kappa(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}).$$

By the CLT we have that  $\mathbf{U}_N = N^{1/2}(\mathbf{S} - \boldsymbol{\Sigma})$  converges in distribution to normal with zero mean and covariances

$$\text{Cov}(u_{ij}, u_{kl}) = \kappa_{1111}^{ijkl} + \kappa_{11}^{ik}\kappa_{11}^{jl} + \kappa_{11}^{il}\kappa_{11}^{jk}.$$

If  $\mathbf{X}$  has normal distribution, then  $\kappa = 0$  and

$$\text{Cov}(u_{ij}, u_{kl}) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}.$$

Denote by  $\boldsymbol{\Gamma}_N$  the corresponding  $m^2 \times m^2$  covariance matrix. For elliptical distribution,  $N^{1/2}(\mathbf{s} - \boldsymbol{\sigma})$  converges in distribution to normal with zero mean and  $m^2 \times m^2$  covariance matrix  $\boldsymbol{\Gamma}$  with

$$\boldsymbol{\Gamma} = (1 + \kappa)\boldsymbol{\Gamma}_N + \kappa\boldsymbol{\sigma}\boldsymbol{\sigma}'.$$

## 16 Correlation analysis

### 16.1 Partial correlation

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

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

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

$$\text{Corr}(X, Y|Z) = \text{Corr}\left(X - \frac{\text{Cov}(X, Z)}{\text{Var}(Z)}Z, Y - \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)}Z\right).$$

Without loss of generality we can assume that  $\text{Var}(X) = \text{Var}(Y) = \text{Var}(Z) = 1$ . Then

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

In similar way partial correlation between random variables  $X$  and  $Y$  given random variables  $Z_1, Z_2, \dots, Z_n$ , is defined. That is, suppose that  $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z_1] = \dots = \mathbb{E}[Z_n] = 0$ , and  $\text{Var}(X) = \text{Var}(Y) = \text{Var}(Z_1) = \dots = \text{Var}(Z_n) = 1$ . Consider the problem

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

Solution of this problem is  $\boldsymbol{\beta} = \boldsymbol{\Sigma}_Z^{-1}\boldsymbol{\Sigma}_{ZX}$ . Hence

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

## 16.2 Canonical correlation analysis

Consider random vectors  $\mathbf{X} = (X_1, \dots, X_p)'$  and  $\mathbf{Y} = (Y_1, \dots, Y_q)'$ . Let  $\boldsymbol{\mu}_1 = \mathbb{E}[\mathbf{X}]$  and  $\boldsymbol{\mu}_2 = \mathbb{E}[\mathbf{Y}]$ , and  $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$  be the covariance matrix of  $(\mathbf{X}', \mathbf{Y}')'$ . Consider random variables  $U = \mathbf{a}'\mathbf{X}$  and  $V = \mathbf{b}'\mathbf{Y}$  for some vectors  $\mathbf{a} \in \mathbb{R}^p$  and  $\mathbf{b} \in \mathbb{R}^q$ . We want to solve the problem

$$\max_{\mathbf{a}, \mathbf{b}} \text{Corr}(U, V). \quad (16.1)$$

Suppose for the moment that  $\boldsymbol{\Sigma}_{11} = \mathbf{I}_p$  and  $\boldsymbol{\Sigma}_{22} = \mathbf{I}_q$ . Then  $\text{Cov}(U, V) = \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}$  and  $\text{Var}(U) = \mathbf{a}'\mathbf{a}$ ,  $\text{Var}(V) = \mathbf{b}'\mathbf{b}$ . Hence problem (16.1) becomes

$$\max_{\mathbf{a}, \mathbf{b}} \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\mathbf{a}}\sqrt{\mathbf{b}'\mathbf{b}}}. \quad (16.2)$$

Note that for a given vector  $\mathbf{w}$ , the maximum of  $\mathbf{w}'\mathbf{b}$  subject to  $\|\mathbf{b}\| = 1$  is attained at  $\bar{\mathbf{b}} = \mathbf{w}/\|\mathbf{w}\|$ . Therefore for given  $\mathbf{a}$  the maximum in (16.2) is attained at  $\mathbf{b} = \boldsymbol{\Sigma}_{21}\mathbf{a}$ . Hence with respect to  $\mathbf{a}$  problem (16.2) becomes

$$\max_{\mathbf{a}} \left\{ \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\mathbf{a}}{\sqrt{\mathbf{a}'\mathbf{a}}\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\mathbf{a}}} = \sqrt{\frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\mathbf{a}}{\mathbf{a}'\mathbf{a}}} \right\} \quad (16.3)$$

Optimal solution  $\bar{\mathbf{a}}$  of problem (16.3) is given by the eigenvector of matrix  $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}$  corresponding to its largest eigenvalue  $\lambda_1$ , and the maximum in (16.1) is equal to  $\sqrt{\lambda_1}$ . Similar the optimal  $\bar{\mathbf{b}}$  is given by the eigenvector of matrix  $\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}$  corresponding to its largest eigenvalue  $\lambda_1$ . Note that

$$\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{21}\bar{\mathbf{a}} = \lambda_1\boldsymbol{\Sigma}_{21}\bar{\mathbf{a}},$$

and hence  $\bar{\mathbf{b}} = \boldsymbol{\Sigma}_{21}\bar{\mathbf{a}}$ .

In general let  $\mathbf{c} = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{a}$  and  $\mathbf{d} = \boldsymbol{\Sigma}_{22}^{1/2}\mathbf{b}$ . Then

$$\text{Corr}(U, V) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b}}} = \frac{\mathbf{c}'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\mathbf{d}}{\sqrt{\mathbf{c}'\mathbf{c}}\sqrt{\mathbf{d}'\mathbf{d}}}.$$

Hence the maximum is attained at  $\bar{\mathbf{c}}$  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  $\lambda_1$ , and at  $\bar{\mathbf{d}}$  given by the eigenvector of  $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$  corresponding to its largest eigenvalue  $\lambda_1$ , and

$$\bar{\mathbf{d}} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\bar{\mathbf{c}}.$$

We have that

$$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}\bar{\mathbf{c}} = \lambda_1\bar{\mathbf{c}}$$

and  $\bar{\mathbf{c}} = \boldsymbol{\Sigma}_{11}^{1/2}\bar{\mathbf{a}}$ . Hence

$$\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\bar{\mathbf{a}} = \lambda_1\bar{\mathbf{a}}, \quad (16.4)$$

and similarly

$$\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\bar{\mathbf{b}} = \lambda_1\bar{\mathbf{b}}. \quad (16.5)$$

Let  $\mathbf{a}_1 = \bar{\mathbf{a}}$  and  $\mathbf{b}_1 = \bar{\mathbf{b}}$ , and  $U_1 = \mathbf{a}_1'\mathbf{X}$  and  $V_1 = \mathbf{b}_1'\mathbf{Y}$ . At the second stage we want to find  $U_2 = \mathbf{a}_2'\mathbf{X}$  and  $V_2 = \mathbf{b}_2'\mathbf{Y}$  such that  $\text{Cov}(U_2, U_1) = 0$ ,  $\text{Cov}(V_2, V_1) = 0$  and  $\text{Corr}(U_2, V_2)$  is maximized. Consider  $\mathbf{c}_2 = \boldsymbol{\Sigma}_{11}^{1/2}\mathbf{a}_2$  and  $\mathbf{d}_2 = \boldsymbol{\Sigma}_{22}^{1/2}\mathbf{b}_2$ . Then

$$\text{Cov}(U_2, U_1) = \mathbf{a}_2'\boldsymbol{\Sigma}_{11}\mathbf{a}_1 = \mathbf{c}_2'\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{11}^{-1/2}\mathbf{c}_1 = \mathbf{c}_2'\mathbf{c}_1.$$

Hence  $\text{Cov}(U_2, U_1) = 0$  iff  $\mathbf{c}'_2 \mathbf{c}_1 = 0$ . Therefore the second stage problem is

$$\max_{\mathbf{c}} \frac{\mathbf{c}' \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1/2} \mathbf{d}}{\sqrt{\mathbf{c}' \mathbf{c}} \sqrt{\mathbf{d}' \mathbf{d}}} \text{ s.t. } \mathbf{c}' \mathbf{c}_1 = 0.$$

The maximum is attained at  $\bar{\mathbf{c}}$  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  $\lambda_2$ . And so on.

## 17 Classification problem

Consider an  $m \times 1$  random vector  $\mathbf{X}$  of measurements. We want to classify  $\mathbf{X}$  into one of two population  $\pi_1$  or  $\pi_2$ . Let  $p_1(\mathbf{x})$  and  $p_2(\mathbf{x})$  be respective densities of populations  $\pi_1$  and  $\pi_2$ . Suppose that the probability that an observation comes from population  $\pi_i$  is  $q_i$ ,  $i = 1, 2$ . Consider regions  $R_1 \subset \mathbb{R}^m$  and  $R_2 = \mathbb{R}^m \setminus R_1$ . If  $\mathbf{X} \in R_1$  we classify  $\mathbf{X}$  as from  $\pi_1$ , and if  $\mathbf{X} \in R_2$  we classify  $\mathbf{X}$  as from  $\pi_2$ . Then the probability of misclassification of an observation from  $\pi_1$  is

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

Similarly the probability of misclassification of an observation from  $\pi_2$  is  $\int_{R_1} p_2(\mathbf{x}) d\mathbf{x}$ . The expected loss of misclassification is

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

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

Note that

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

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

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

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

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

It follows that the expected loss is minimized if

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

Or equivalently

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

and

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

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

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

and

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

When

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

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

### 17.1 Classification with normally distributed populations

Suppose that the populations  $\pi_1$  and  $\pi_2$  have multivariate normal distributions with equal covariance matrices, i.e.,  $\pi_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ ,  $i = 1, 2$ . Then

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

and

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

Hence the optimal region is

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

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

$$R_1 = \left\{ \mathbf{x} : \mathbf{x}' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq \frac{1}{2} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \kappa \right\}. \quad (17.1)$$

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

### 17.2 Fisher discriminant analysis

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

$$\max_{\mathbf{d} \in \mathbb{R}^m} \left\{ g(\mathbf{d}) = \frac{(\mathbf{d}' \boldsymbol{\mu}_1 - \mathbf{d}' \boldsymbol{\mu}_2)^2}{\mathbf{d}' \boldsymbol{\Sigma}_1 \mathbf{d} + \mathbf{d}' \boldsymbol{\Sigma}_2 \mathbf{d}} \right\}. \quad (17.2)$$

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

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

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

Hence the optimal solution  $\bar{\mathbf{d}}$  of problem (17.2) is defined by the equation

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

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

$$\bar{\mathbf{d}} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \quad (17.3)$$

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



### 17.3 Several populations

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

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

and

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

where

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

and

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

Consider the following optimization problem

$$\max_{\mathbf{d} \in \mathbb{R}^m} \left\{ g(\mathbf{d}) = \frac{\mathbf{d}'\mathbf{M}\mathbf{d}}{\mathbf{d}'\boldsymbol{\Omega}\mathbf{d}} \right\}. \quad (17.4)$$

The optimal solution of the above problem (17.4) is defined by the equation

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

where  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$  are eigenvalues of  $\boldsymbol{\Omega}^{-1}\mathbf{M}$ . That is, the optimal solution  $\mathbf{d}_1$  of problem (17.4) is the eigenvector of  $\boldsymbol{\Omega}^{-1}\mathbf{M}$  corresponding to its largest eigenvalue.

Next maximize  $g(\mathbf{d})$  subject to  $\mathbf{d}'\boldsymbol{\Omega}\mathbf{d}_1 = 0$ . The solution of this problem is given by eigenvector  $\mathbf{d}_2$  of  $\boldsymbol{\Omega}^{-1}\mathbf{M}$  corresponding to the eigenvalue  $\lambda_2$ . By continuing this process we obtain discriminant functions  $\mathbf{d}_i'\mathbf{X}$ ,  $i = 1, \dots, r-1$ . Note that  $\text{rank}(\mathbf{M}) \leq r-1$ , and hence  $\lambda_r = \dots = \lambda_m = 0$ .

#### 17.3.1 Mahalanobis distance

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

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

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

### 17.4 Bayes and Logistic Regression classifiers

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

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

where  $q_1 = \text{Prob}(Y = 1)$  and  $q_2 = \text{Prob}(Y = -1)$ . We classify  $\mathbf{X}$  in  $\pi_1$  if  $\text{Prob}(Y = 1|\mathbf{X} = \mathbf{x}) > \text{Prob}(Y = -1|\mathbf{X} = \mathbf{x})$ , which is equivalent to  $p_1(\mathbf{x})q_1 > p_2(\mathbf{x})q_2$ .

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

$$\log \text{odd}(\mathbf{x}) = \beta_0 + \beta' \mathbf{x}. \quad (17.5)$$

We classify  $\mathbf{X}$  in  $\pi_1$  if  $\text{odd}(\mathbf{x}) > 1$ . This is equivalent to  $\beta_0 + \beta' \mathbf{x} > 0$ .

Note that

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

In case of normal distributions with the same covariance matrix  $\Sigma$  we have that (see section 17.1)

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

In that case (assuming  $q_1 = q_2$ ) equation (17.5) holds with  $\beta = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ .

## 18 Support Vector Machines

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

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

$$\max_{\beta_0, \beta, \|\beta\|=1} c \quad (18.6)$$

$$\text{subject to } y_i(\beta_0 + \beta' \mathbf{x}_i) \geq c, \quad i = 1, \dots, n. \quad (18.7)$$

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

$$\min_{\beta_0, \beta} \|\beta\|^2 \quad (18.8)$$

$$\text{subject to } y_i(\beta_0 + \beta' \mathbf{x}_i) \geq 1, \quad i = 1, \dots, n. \quad (18.9)$$

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

$$y_i(\beta_0 + \beta' \mathbf{x}_i) \geq c - \xi_i, \quad i = 1, \dots, n, \quad (18.10)$$

or

$$y_i(\beta_0 + \beta' \mathbf{x}_i) \geq c(1 - \xi_i), \quad i = 1, \dots, n, \quad (18.11)$$

where  $\xi_i \geq 0$ ,  $i = 1, \dots, n$ , and  $\sum_{i=1}^n \xi_i \leq \text{const}$ . Similar to (18.8)–(18.9) formulation (18.11) leads to the following optimization problem

$$\min_{\beta_0, \beta, \xi} \|\beta\|^2 \quad (18.12)$$

$$\text{subject to } y_i(\beta_0 + \beta' \mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1, \dots, n, \quad (18.13)$$

$$\xi_i \geq 0, \quad i = 1, \dots, n, \quad (18.14)$$

$$\sum_{i=1}^n \xi_i \leq \text{const}. \quad (18.15)$$

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

$$\min_{\beta_0, \beta} \sum_{i=1}^n \delta(-y_i(\beta_0 + \beta' \mathbf{x}_i)), \quad (18.16)$$

where  $\delta(t) = 1$  if  $t \geq 0$ , and  $\delta(t) = 0$  if  $t < 0$ . Problem (18.16) is a difficult combinatorial problem. Note that  $\delta(t) \leq [1 + t]_+$ , where  $[a]_+ = \max\{0, a\}$ . Therefore we can approximate problem (18.16) by the following *convex* problem

$$\min_{\beta_0, \beta} \sum_{i=1}^n [1 - y_i(\beta_0 + \beta' \mathbf{x}_i)]_+ + c\|\beta\|^2. \quad (18.17)$$

Equivalently we can formulate problem (18.17) as

$$\min_{\beta_0, \beta, \xi} \quad \frac{1}{2}\|\beta\|^2 + \gamma \sum_{i=1}^n \xi_i \quad (18.18)$$

$$\text{s.t. } y_i(\beta_0 + \beta' \mathbf{x}_i) \geq 1 - \xi_i, \quad i = 1, \dots, n, \quad (18.19)$$

$$\xi_i \geq 0, \quad i = 1, \dots, n, \quad (18.20)$$

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

The Lagrangian of the above problem is

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

The Lagrangian dual of problem (18.18)–(18.20) is the problem

$$\max_{\lambda \geq 0, \mu \geq 0} \min_{\beta_0, \beta, \xi \geq 0} L(\beta_0, \beta, \xi, \lambda, \mu). \quad (18.21)$$

The corresponding Lagrangian-Wolfe dual is obtained by employing optimality conditions for the problem of minimization of  $L(\beta_0, \beta, \xi, \lambda, \mu)$  in (18.21). That is, by setting derivatives of the Lagrangian to zero, with respect to  $\beta, \beta_0, \xi$ , we have

$$\beta = \sum_{i=1}^n \lambda_i y_i \mathbf{x}_i \quad (18.22)$$

$$0 = \sum_{i=1}^n \lambda_i y_i \quad (18.23)$$

$$\lambda_i = \gamma - \mu_i, \quad i = 1, \dots, n, \quad (18.24)$$

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

$$\max_{\lambda} \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \quad (18.25)$$

$$\text{s.t.} \quad 0 \leq \lambda_i \leq \gamma, \quad i = 1, \dots, n, \quad (18.26)$$

$$\sum_{i=1}^n \lambda_i y_i = 0. \quad (18.27)$$

We also have the following complementarity conditions for problem (18.18)–(18.20):

$$\lambda_i [y_i(\beta_0 + \beta' \mathbf{x}_i) - (1 - \xi_i)] = 0, \quad i = 1, \dots, n, \quad (18.28)$$

$$\mu_i \xi_i = 0, \quad i = 1, \dots, n. \quad (18.29)$$

Given solution  $\bar{\lambda}$  of problem (18.25)–(18.27) the optimal  $\beta$  can be computed using equation (18.22), that is

$$\bar{\beta} = \sum_{i=1}^n \bar{\lambda}_i y_i \mathbf{x}_i. \quad (18.30)$$

The complementarity conditions (18.29) mean that  $\xi_i = 0$  if  $\mu_i > 0$ , and similarly for the complementarity conditions (18.28). By (18.24) we have that  $\mu_i > 0$  if  $\lambda_i < \gamma$ . Therefore by using equation (18.28), for  $0 < \bar{\lambda}_i < \gamma$  the optimal  $\beta_0$  can be computed by solving  $y_i f(\mathbf{x}_i) = 1$ , where  $f(\mathbf{x}) = \beta_0 + \beta' \mathbf{x}$ .

Suppose now that we want to make classification by using feature vectors  $h(\mathbf{x}_i)$ ,  $i = 1, \dots, n$ , where  $h : \mathbb{R}^p \rightarrow \mathbb{R}^q$ . We can approach this by solving the corresponding dual problem with replacing  $\mathbf{x}_i$  with  $h(\mathbf{x}_i)$ ,  $i = 1, \dots, n$ . That is the objective function in (18.25) is replaced by

$$\sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j h(\mathbf{x}_i)' h(\mathbf{x}_j). \quad (18.31)$$

Consequently, by using  $\beta = \sum_{i=1}^n \lambda_i y_i h(\mathbf{x}_i)$  (see (18.30)), the classification is performed according to the sign of

$$f(\mathbf{x}) = \beta_0 + \beta' h(\mathbf{x}) = \beta_0 + \sum_{i=1}^n \lambda_i y_i h(\mathbf{x})' h(\mathbf{x}_i). \quad (18.32)$$

Both expressions (18.31) and (18.32) are defined by the so-called kernel function

$$K(\mathbf{x}, \mathbf{z}) = h(\mathbf{x})' h(\mathbf{z}) = \sum_{j=1}^q h_j(\mathbf{x}) h_j(\mathbf{z}).$$

In terms of the kernel function the objective function (18.31) can be written as

$$\sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \quad (18.33)$$

and the classifier (18.32) as

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^n \lambda_i y_i K(\mathbf{x}, \mathbf{x}_i). \quad (18.34)$$

For example

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

defines a quadratic separation.

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

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

## 18.1 Kernel PCA

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

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

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

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

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

$$\mathbf{e}_s = \frac{1}{\lambda_s N} \sum_{i=1}^N \alpha_{is} \mathbf{Z}_i, \quad (18.35)$$

where  $\alpha_{is} = \mathbf{Z}_i' \mathbf{e}_s$ ,  $s = 1, \dots, q$ ,  $i = 1, \dots, N$ . It follows by (18.35) that

$$\alpha_{is} = \frac{1}{\lambda_s N} \mathbf{Z}_i' \left( \sum_{j=1}^N \alpha_{js} \mathbf{Z}_j \right) = \frac{1}{\lambda_s N} \sum_{j=1}^N \alpha_{js} \mathbf{Z}_i' \mathbf{Z}_j = \frac{1}{\lambda_s N} \sum_{j=1}^N \alpha_{js} \mathbf{h}(\mathbf{X}_i)' \mathbf{h}(\mathbf{X}_j). \quad (18.36)$$

Consider kernel function  $K(\mathbf{x}, \mathbf{z}) = \mathbf{h}(\mathbf{x})' \mathbf{h}(\mathbf{z})$ . In terms of the kernel function equation (18.36) can be written as

$$\sum_{j=1}^N \alpha_{js} K(\mathbf{X}_i, \mathbf{X}_j) = \lambda_s N \alpha_{is}. \quad (18.37)$$

Consider  $N \times N$  matrix  $\mathbf{K}$  with components  $\mathbf{K}_{ij} = K(\mathbf{X}_i, \mathbf{X}_j)$ ,  $i, j = 1, \dots, N$ . Equation (18.37) can be written as

$$\mathbf{K} \boldsymbol{\alpha}_s = \lambda_s N \boldsymbol{\alpha}_s, \quad s = 1, \dots, q, \quad (18.38)$$

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

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

That is  $\boldsymbol{\alpha}_s' \mathbf{K} \boldsymbol{\alpha}_s = \lambda_s^2 N^2$ . Because of (18.38) 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  $\mathbf{K}$  corresponding to its largest eigenvalues. This will give us vectors  $\boldsymbol{\alpha}_s$  and numbers  $\lambda_s$ . For a data point  $\mathbf{X} \in \mathbb{R}^m$  its  $s$ -PCA component is  $\mathbf{e}_s' \mathbf{h}(\mathbf{X})$ . By (18.35) we have

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

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

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