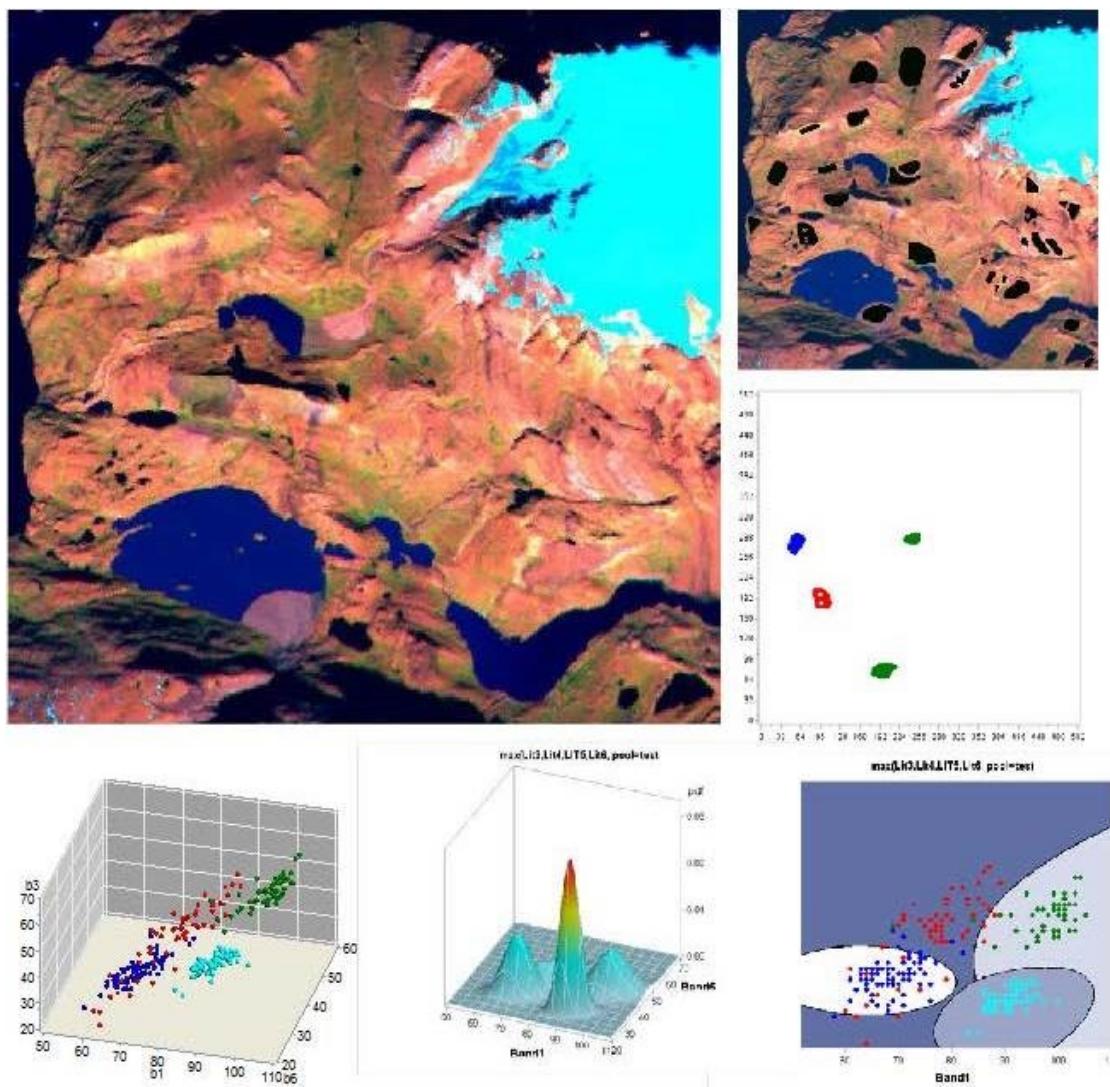


Multivariate Statistics

For the Technical Sciences

Knut Conradsen,
Anders Nymark Christensen,
Allan Aasbjerg Nielsen,
Bjarne Kjær Ersbøll



DTU Compute, Lyngby
2018 Autumn, v. 0.4

Contents

1 Multidimensional variables	3
1.1 Moments of multidimensional random variables	3
1.1.1 The mean value	3
1.1.2 The variance-covariance matrix (dispersion matrix).	5
1.1.3 Correlation	7
1.1.4 Covariance	8
1.2 The multivariate normal distribution	11
1.2.1 Definition and simple properties	11
1.2.2 Independence and contour ellipsoids.	18
1.2.3 Conditional distributions	22
1.2.4 Theorem of reproducibility and the central limit theorem. .	23
1.2.5 Estimation of the parameters in a multivariate normal distribution.	24
1.2.6 The two-dimensional normal distribution.	26
1.3 Correlation and regression	31
1.3.1 The partial correlation coefficient.	31
1.3.2 The multiple correlation coefficient	39
1.3.3 Regression	43
1.4 The partition theorem	47
1.5 The Wishart distribution and the generalized variance	52
1.6 The complex normal distribution and the complex Wishart distribution	57
1.7 On estimation of multidimensional parameters	57
1.7.1 Maximum likelihood estimation	58
1.7.2 Restricted Maximum Likelihood (REML)	69
1.7.3 Profile, partial, marginal, conditional, and quasi likelihood	72
2 The general linear model	74
2.1 Estimation in the general linear model	74
2.1.1 Formulation of the Model.	74
2.1.2 Estimation in the regular case	77
2.1.3 The case of $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ singular	83
2.1.4 Constrained estimation	92

2.1.5	Confidence-intervals for estimated values. Prediction-intervals	97
2.2	Tests in the general linear model	102
2.2.1	Test for a lower dimension of model space	103
2.2.2	Successive testing in the general linear model.	113
3	Regression analysis	123
3.1	Linear regression analysis	123
3.1.1	Notation and model.	123
3.1.2	Correlation and regression.	126
3.1.3	Analysis of assumptions.	128
3.1.4	On "Influence Statistics"	132
3.2	Regression using orthogonal polynomials	138
3.2.1	Definition and formulation of the model.	138
3.2.2	Determination of orthogonal polynomials.	142
3.3	Choice of the "best" regression equation	148
3.3.1	The Problem.	148
3.3.2	Examination of all regressions.	151
3.3.3	Backwards elimination.	152
3.3.4	Forward selection	154
3.3.5	Stepwise regression.	157
3.3.6	Numerical appendix.	159
3.4	Other regression models and solutions	165
3.4.1	Orthogonal regression (linear functional relationship).	165
3.4.2	Regularization and Ridge Regression	168
3.4.3	Non-linear regression and curve fitting.	177
4	Tests in the multidimensional normal distribution	182
4.1	Test for mean value.	182
4.1.1	Hotelling's T^2 in the One-Sample Situation	182
4.1.2	Hotelling's T^2 in the two-sample situation.	188
4.2	The multidimensional general linear model.	194
4.3	Multivariate Analyses of Variance (MANOVA)	207
4.3.1	One-sided multi-dimensional analysis of variance	207
4.3.2	Two-sided multidimensional analysis of variance	210
4.4	Tests regarding variance-covariance matrices	216
4.4.1	Tests regarding a single variance-covariance matrix	216
4.4.2	Test for equality of several variance-covariance matrices	219
5	Discriminant analysis and classification	222
5.1	Discrimination between two populations	224
5.1.1	Bayes and minimax solutions	224
5.1.2	Discrimination between two normal populations	228
5.1.3	Discrimination with unknown parameters	237
5.1.4	Test for best discrimination function	239
5.2	Discrimination between several populations	240

5.2.1	The Bayes solution	240
5.2.2	The Bayes' solution in the case with several normal distributions	242
5.2.3	The case with several normal distributions and unknown parameters	245
5.2.4	Short about kernel estimates and nearest neighbor estimates	248
5.3	Evaluation	252
5.3.1	Some performance measures for a classifier	252
5.3.2	Terminology from non-statistical communities.	254
5.3.3	Comparing classifiers: The ROC curve and McNemar's test.	255
5.4	Feature selection and extraction.	259
5.4.1	Test for further information	259
5.4.2	Principal Component Analysis	263
5.4.3	Canonical Discriminant Analysis	265
6	Principal components, canonical variables and correlations, and factor analysis	270
6.1	Principal components	271
6.1.1	Definition and simple characteristics	271
6.1.2	Estimation and Testing	277
6.2	Canonical variables and correlations	283
6.2.1	Definition and properties	285
6.2.2	Estimation and testing	291
6.3	Factor analysis	305
6.3.1	Model and assumptions	305
6.3.2	Estimation of factor loadings	308
6.3.3	Factor rotation	310
6.3.4	Computation of the factor scores	315
6.3.5	Briefly on maximum likelihood factor analysis	319
6.3.6	Q-mode analysis	322
6.4	PLS – Regression and Projection on Latent Structure	325
6.4.1	Introduction	325
6.4.2	Ordinary least squares regression.	325
6.4.3	Principal components regression	326
6.4.4	Canonical correlation regression	327
6.4.5	Reduced rank regression	328
6.4.6	Covariance maximization	329
6.4.7	Partial least squares regression	330
A	Summary of linear algebra	334
A.1	Vector space	334
A.1.1	Definition of a vector space	334
A.1.2	Direct sum of vector spaces	337
A.2	Linear transformations and matrices	339
A.2.1	Linear transformations	339
A.2.2	Matrices	340

A.2.3	Linear transformations using matrix-formulation	342
A.2.4	Coordinate transformation	343
A.2.5	Rank of a matrix	344
A.2.6	Determinant of a matrix	345
A.2.7	Block matrices	348
A.3	Pseudoinverse or generalised inverse matrix	350
A.4	Eigenvalue problems. Quadratic forms	359
A.4.1	Eigenvalues and eigenvectors for symmetric matrices . . .	359
A.4.2	Singular value decomposition of an arbitrary matrix. Q - and R -mode analysis	365
A.4.3	Quadratic forms and positive semi-definite matrices . . .	368
A.4.4	The general eigenvalue problem for symmetrical matrices	373
A.4.5	The trace of a matrix	377
A.4.6	Differentiation of linear form and quadratic form	378
A.5	Tensor- or Kronecker product of matrices	380
A.6	Inner products and norms	381

Front page illustration

The large image in the upper left corner is a false color composite of a Landsat satellite image from Ymer Ø, Central East Greenland. The small image in the upper right corner shows in black, areas with a known, homogeneous geology (lithological unit), and the image immediately below shows the location of 4 of those that are used as training areas for a discriminant analysis aiming at a complete geological mapping of the entire scene. The bottom row shows to the left a three-dimensional scatter plot of reflectance values for pixels from the 4 units. The color coding is the same as the one used for showing the training areas. The next image show estimated density functions for the 4 populations, and the rightmost image shows the discrimination boundaries for a classifier based on the training data.

A more thorough description may be found in chapter 5.

Preface

This is a beta version of a textbook aiming at representing multivariate statistical methods and tools and at illustrating their use in settings relevant to engineers. The history behind the textbook goes back to the late 1970's where one of us (Knut Conradsen) wrote lecture notes on multivariate statistical analysis in Danish ('En Introduction til Statistik, Vol. 2'). Around the turn of the century, these notes were modified and rewritten in English and published under the title 'Multivariate Statistics - An Introduction', with Bjarne Kjær Ersbøll and Knut Conradsen as authors.

However, the notion of data collecting and of computing has changed dramatically since the first versions appeared. The amount of data involved in decision making has increased tremendously. Satellites are orbiting the Earth and are providing time series of gigabyte size images for online analysis. Controlling greenhouse growing of flowers may involve daily analysis of robot captured images of a million plants. Modern sensors may yield outcome of many thousand variables on each test specimen thus generating many more variables than observations etc.

As a consequence, data science has had to develop rapidly in order to take advantage of those new possibilities: the plethora of data available for solving problems in science and technology and the vastly increased possibilities of actually solving the computational problems in those analyses.

Based on our experience from teaching engineering students at DTU - the Technical University of Denmark - and from the research projects we have been involved in, we have now started a fundamental redesign of the exposition of the basic multivariate statistical tools.

From the preface to earlier versions we quote: "Furthermore the text has been updated with sections that hopefully make the interpretation of output from statistical software packages easier to follow. A special emphasis has been put on facilitating the understanding of output from SAS."

Suggestions for corrections are very welcome.

Knut Conradsen (knco@dtu.dk)
Anders Nymark Christensen (anym@dtu.dk)
Allan Aasbjerg Nielsen (alan@dtu.dk)
Bjarne Kjær Ersbøll (bker@dtu.dk)

|||| Chapter 1

Multidimensional variables

In this chapter we start by generalising statistical measures known from basic statistics to multidimensional random variables. Then we discuss the multivariate normal distribution and distributions derived from it. Finally we shortly describe the special considerations that estimation and testing give rise to.

1.1 Moments of multidimensional random variables

1.1.1 The mean value

Let there be given a *random* (or *stochastic*) *matrix*, i.e. a matrix, where the single elements are random (stochastic) variables:

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

We then define the *mean value*, or the *expectation*, or the *expected value* of \mathbf{X} as

$$E(\mathbf{X}) = \begin{bmatrix} E(X_{11}) & \cdots & E(X_{1k}) \\ \vdots & & \vdots \\ E(X_{n1}) & \cdots & E(X_{nk}) \end{bmatrix} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1k} \\ \vdots & & \vdots \\ \mu_{n1} & \cdots & \mu_{nk} \end{bmatrix} = \boldsymbol{\mu}.$$

|||| **Theorem 1.1**

Let \mathbf{A} be a $n \times k$ matrix of constants. Then

$$E(\mathbf{A} + \mathbf{X}) = \mathbf{A} + E(\mathbf{X}).$$

This theorem follows trivially from the definition as does the following.

|||| **Theorem 1.2**

Let \mathbf{A} and \mathbf{B} be constant matrices, so that \mathbf{AX} and \mathbf{XB} exist. Then

$$\begin{aligned} E(\mathbf{AX}) &= \mathbf{A}E(\mathbf{X}) \\ E(\mathbf{XB}) &= E(\mathbf{X})\mathbf{B} \end{aligned}$$

Finally we have

|||| **Theorem 1.3**

Let \mathbf{X} and \mathbf{Y} be random matrices of the same dimensions. Then

$$E(\mathbf{X} + \mathbf{Y}) = E(\mathbf{X}) + E(\mathbf{Y}).$$

|||| **Remark 1.4**

We have not mentioned that we of course assume, that the involved expected values exist. This is assumed here and in all the following, where these are mentioned.

1.1.2 The variance-covariance matrix (dispersion matrix).

The generalisation of the variance of a stochastic variable is the *variance-covariance matrix* (or *dispersion matrix*) for a multidimensional random (stochastic) variable $\mathbf{X} = (X_1, \dots, X_n)^T$. It is defined by

$$\mathbf{D}(\mathbf{X}) = \boldsymbol{\Sigma} = \mathbb{E}\{(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\},$$

where

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{X}).$$

It should be noted, that $\mathbf{D}(\mathbf{X})$ also often is called the covariance-matrix and is then denoted $\text{Cov}(\mathbf{X})$. However, this is a bit misleading, since it could be misunderstood as the covariance between two (multidimensional) stochastic variables. Another commonly used notation is $\mathbf{V}(\mathbf{X})$. Furthermore, we note that

$$(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T = \begin{bmatrix} X_1 - \mu_1 \\ \vdots \\ X_n - \mu_n \end{bmatrix} (X_1 - \mu_1, \dots, X_n - \mu_n) =$$

$$\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_n - \mu_n) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n - \mu_n)(X_1 - \mu_1) & (X_n - \mu_n)(X_2 - \mu_2) & \cdots & (X_n - \mu_n)^2 \end{bmatrix}$$

i.e. the variance-covariance matrix's (i, j) 'th element is $\text{Cov}(X_i, X_j)$, or

$$\boldsymbol{\Sigma} = \mathbf{D}(\mathbf{X}) = \begin{bmatrix} \mathbf{V}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \mathbf{V}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \mathbf{V}(X_n) \end{bmatrix}.$$

We will often use the following notation

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix},$$

i.e. the variances can be denoted both as σ_i^2 and as σ_{ii} . We note, that $\boldsymbol{\Sigma}$ is symmetric. More interesting is the following

|||| **Theorem 1.5**

The variance-covariance matrix Σ for a multidimensional random variable is positive semidefinite. This is a necessary and sufficient condition.

|||| **Proof**

For any vector y we have

$$\begin{aligned} \mathbf{y}^T \Sigma \mathbf{y} &= \mathbf{y}^T E\{(X - \mu)(X - \mu)^T\} \mathbf{y} \\ &= E\{\mathbf{y}^T (X - \mu)(X - \mu)^T \mathbf{y}\} \\ &= E\{[(X - \mu)^T \mathbf{y}]^T [(X - \mu)^T \mathbf{y}]\} \\ &\geq 0, \end{aligned}$$

since the expression in the curly brackets is ≥ 0 . ■

There exist theorems which are analogous to the ones known from the one dimensional stochastic variables.

|||| **Theorem 1.6**

Let X and Y be independent. Then

$$D(X + Y) = D(X) + D(Y).$$

Let b be a constant. Then we have

$$D(b + X) = D(X).$$

If A is a constant matrix, so that AX exists, then the following holds

$$D(AX) = A D(X) A^T.$$

|||| Proof

The first relation comes from

$$\begin{aligned}\text{Cov}(X_i + Y_i, X_j + Y_j) &= \text{Cov}(X_i, X_j) + \text{Cov}(X_i, Y_j) + \\ &\quad \text{Cov}(Y_i, X_j) + \text{Cov}(Y_i, Y_j) \\ &= \text{Cov}(X_i, X_j) + \text{Cov}(Y_i, Y_j),\end{aligned}$$

since $\text{Cov}(Y_i, X_j) = 0$, because X_j and Y_i are independent. The second relation is trivial. The last one comes from

$$\begin{aligned}D(\mathbf{A}X) &= E\{(\mathbf{A}X - \mathbf{A}\mu)(\mathbf{A}X - \mathbf{A}\mu)^T\} \\ &= E\{\mathbf{A}[X - \mu][X - \mu]^T\mathbf{A}^T\} \\ &= \mathbf{A}E\{[X - \mu][X - \mu]^T\}\mathbf{A}^T \\ &= \mathbf{A}D(X)\mathbf{A}^T \\ &= \mathbf{A}\Sigma\mathbf{A}^T\end{aligned}$$

■

1.1.3 Correlation

If we let

$$\mathbf{V} = \text{diag}\left(\frac{1}{\sigma_1}, \dots, \frac{1}{\sigma_n}\right) = \begin{bmatrix} \sigma_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^{-1} \end{bmatrix}$$

and we "scale" X by \mathbf{V} , we get

$$D(\mathbf{V}X) = \mathbf{V}\Sigma\mathbf{V}^T = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma_1\sigma_2} & \cdots & \frac{\sigma_{1n}}{\sigma_1\sigma_n} \\ \frac{\sigma_{12}}{\sigma_1\sigma_2} & 1 & \cdots & \frac{\sigma_{2n}}{\sigma_2\sigma_n} \\ \vdots & \vdots & & \vdots \\ \frac{\sigma_{1n}}{\sigma_1\sigma_n} & \frac{\sigma_{2n}}{\sigma_2\sigma_n} & \cdots & 1 \end{bmatrix}.$$

We note, that the elements are the correlation coefficients between X 's components, which is why this matrix is also called the *correlation matrix* for X , and we write

$$R(\mathbf{X}) = \begin{bmatrix} 1 & \cdots & \rho_{1n} \\ \vdots & & \vdots \\ \rho_{1n} & \cdots & 1 \end{bmatrix},$$

where

$$\rho_{ij} = \text{Corr}(X_i, X_j) = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{V}(X_i)\text{V}(X_j)}}.$$

|||| **Remark 1.7**

It follows trivially from theorem 1.5, that a correlation matrix R for a multidimensional random variable is also positive semidefinite.

1.1.4 Covariance

Let there be given two random variables

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix}$$

with mean values μ and ν . We now define the covariance between \mathbf{X} and \mathbf{Y} as

$$\mathbf{C}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\nu})^T] = \begin{bmatrix} \text{Cov}(X_1, Y_1) & \cdots & \text{Cov}(X_1, Y_q) \\ \vdots & & \vdots \\ \text{Cov}(X_p, Y_1) & \cdots & \text{Cov}(X_p, Y_q) \end{bmatrix}.$$

Then

$$\mathbf{C}(\mathbf{X}, \mathbf{X}) = \mathbf{D}(\mathbf{X})$$

and

$$\mathbf{C}(\mathbf{X}, \mathbf{Y}) = [\mathbf{C}(\mathbf{Y}, \mathbf{X})]^T.$$

Less trivial is

|||| **Theorem 1.8**

Let \mathbf{X} and \mathbf{Y} be as above, and let \mathbf{A} and \mathbf{B} be $n \times p$ and $m \times q$ matrices of constants respectively. Then

$$\mathbf{C}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \mathbf{AC}(\mathbf{X}, \mathbf{Y})\mathbf{B}^T.$$

If \mathbf{U} is a p -dimensional and \mathbf{V} is a q -dimensional random variable the following holds

$$\mathbf{C}(\mathbf{X} + \mathbf{U}, \mathbf{Y}) = \mathbf{C}(\mathbf{X}, \mathbf{Y}) + \mathbf{C}(\mathbf{U}, \mathbf{Y})$$

$$\mathbf{C}(\mathbf{X}, \mathbf{Y} + \mathbf{V}) = \mathbf{C}(\mathbf{X}, \mathbf{Y}) + \mathbf{C}(\mathbf{X}, \mathbf{V}).$$

Finally

$$\mathbf{D}(\mathbf{X} + \mathbf{U}) = \mathbf{D}(\mathbf{X}) + \mathbf{D}(\mathbf{U}) + \mathbf{C}(\mathbf{X}, \mathbf{U}) + \mathbf{C}(\mathbf{U}, \mathbf{X}).$$

|||| **Proof**

According to the definition we have

$$\begin{aligned} \mathbf{C}(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) &= \mathbf{E}[(\mathbf{A}\mathbf{X} - \mathbf{A}\boldsymbol{\mu})(\mathbf{B}\mathbf{Y} - \mathbf{B}\boldsymbol{\nu})^T] \\ &= \mathbf{E}[\mathbf{A}(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\nu})^T \mathbf{B}^T] \\ &= \mathbf{A}\mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\nu})^T]\mathbf{B}^T \\ &= \mathbf{AC}(\mathbf{X}, \mathbf{Y})\mathbf{B}^T. \end{aligned}$$

This proves the first statement. Similarly - if we let $E(\mathbf{U}) = \boldsymbol{\delta}$

$$\begin{aligned} \mathbf{C}(\mathbf{X} + \mathbf{U}, \mathbf{Y}) &= \mathbf{E}[(\mathbf{X} + \mathbf{U} - \boldsymbol{\mu} - \boldsymbol{\delta})(\mathbf{Y} - \boldsymbol{\nu})^T] \\ &= \mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\nu})^T + (\mathbf{U} - \boldsymbol{\delta})(\mathbf{Y} - \boldsymbol{\nu})^T] \\ &= \mathbf{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\nu})^T] + \mathbf{E}[(\mathbf{U} - \boldsymbol{\delta})(\mathbf{Y} - \boldsymbol{\nu})^T] \\ &= \mathbf{C}(\mathbf{X}, \mathbf{Y}) + \mathbf{C}(\mathbf{U}, \mathbf{Y}), \end{aligned}$$

and the corresponding relation with $\mathbf{Y} + \mathbf{V}$ is shown analogously. Finally we have

$$\begin{aligned} \mathbf{D}(\mathbf{X} + \mathbf{U}) &= \mathbf{C}(\mathbf{X} + \mathbf{U}, \mathbf{X} + \mathbf{U}) \\ &= \mathbf{C}(\mathbf{X}, \mathbf{X}) + \mathbf{C}(\mathbf{X}, \mathbf{U}) + \mathbf{C}(\mathbf{U}, \mathbf{X}) + \mathbf{C}(\mathbf{U}, \mathbf{U}). \end{aligned}$$

If $\mathbf{C}(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ then \mathbf{X} and \mathbf{Y} are said to be uncorrelated. This corresponds to

all components of \mathbf{X} being uncorrelated with all components of \mathbf{Y} .

Later, when we consider the multidimensional general linear model we will need the following

||| Theorem 1.9

Let X_1, \dots, X_n be independent, p -dimensional random variables with the same variance-covariance matrix $\Sigma = (\sigma_{ij})$. We let

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

(Note, that the repetition index is the first index and the variable (coordinate) index is the second). If we define

$$\text{vc}(\mathbf{X}) = \begin{bmatrix} X_{11} \\ \vdots \\ X_{n1} \\ \vdots \\ X_{1p} \\ \vdots \\ X_{np} \end{bmatrix}$$

i.e. as the vector consisting of the columns in \mathbf{X} ($\text{vc} = \text{vector of columns}$) we get

$$\mathbf{D}(\text{vc}(\mathbf{X})) = \Sigma \otimes \mathbf{I}_n,$$

where \mathbf{I}_n is the identity matrix of n 'th order.

||| Proof

Follows trivially from the definition of a tensor-product and from the definition of the variance-covariance matrix.

■

1.2 The multivariate normal distribution

The ***multivariate normal distribution*** plays the same important role in the theory of multidimensional variables, as the normal distribution does in the univariate case. We start with

1.2.1 Definition and simple properties

Let X_1, \dots, X_p be mutually independent, $N(0,1)$ distributed variables. We then say that

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix},$$

is standardised (normed) p -dimensional normally distributed, and we write

$$\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}) = N_p(\mathbf{0}, \mathbf{I}),$$

where the last notation is used, if there is any doubt about the dimension. We note, that

$$E(\mathbf{X}) = \mathbf{0}, \quad D(\mathbf{X}) = \mathbf{I}.$$

We define the multivariate normal distribution with general parameters in

|||| Definition 1.10

We say that the p -dimensional random variable \mathbf{X} is normally distributed with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, if \mathbf{X} has the same distribution as

$$\boldsymbol{\mu} + \mathbf{A} \mathbf{U},$$

where \mathbf{A} satisfies

$$\mathbf{A} \mathbf{A}^T = \boldsymbol{\Sigma},$$

and where \mathbf{U} is standardised p -dimensional normally distributed. We write

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where the last notation again is used, if there is any doubt about the dimension.

|||| Remark 1.11

The definition is only valid, if one shows, that $\mathbf{A} \mathbf{A}^T = \mathbf{B} \mathbf{B}^T$ implies that the random variables

$$\boldsymbol{\mu} + \mathbf{A} \mathbf{U} \quad \text{and} \quad \boldsymbol{\mu} + \mathbf{B} \mathbf{V},$$

where \mathbf{U} and \mathbf{V} are standardised normally distributed and not necessarily of the same dimension, have the same distribution. The relation is valid, but we will not pursue this further here. From theorem A.23 follows that for any positive semidefinite matrix Σ there exists a matrix \mathbf{A} with $\mathbf{A} \mathbf{A}^T = \Sigma$, so the expression $\mathbf{N}(\boldsymbol{\mu}, \Sigma)$ makes sense for any positive semidefinite $p \times p$ matrix Σ and any p -dimensional vector $\boldsymbol{\mu}$.

Trivially, we note that

$$\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \Sigma) \quad \Rightarrow \quad \mathbf{E}(\mathbf{X}) = \boldsymbol{\mu} \quad \text{and} \quad \mathbf{D}(\mathbf{X}) = \Sigma$$

i.e. the distribution is parametrised by its mean and variance-covariance matrix.

If Σ has full rank, then the distribution has the density given in

|||| Theorem 1.12

Let $\mathbf{X} \sim \mathbf{N}_p(\boldsymbol{\mu}, \Sigma)$, and let $\text{rk}(\Sigma) = p$. Then \mathbf{X} has the density

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right] \\ &= \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \Sigma}} \exp\left[-\frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}\|^2\right], \end{aligned}$$

where the norm used is the one defined by Σ^{-1} ,

$$\|\mathbf{x} - \boldsymbol{\mu}\|^2 = \|\mathbf{x} - \boldsymbol{\mu}\|_{\Sigma^{-1}}^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

see section A.6.

|||| Proof

Let $\mathbf{U} \sim N_p(\mathbf{0}, \mathbf{I})$. Then \mathbf{U} has the density

$$\begin{aligned} h(\mathbf{u}) &= \prod_{i=1}^p \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u_i^2\right) = \frac{1}{\sqrt{2\pi}^p} \exp\left(-\frac{1}{2}\sum_{i=1}^p u_i^2\right) \\ &= \frac{1}{\sqrt{2\pi}^p} \exp\left(-\frac{1}{2}\mathbf{u}^T \mathbf{u}\right). \end{aligned}$$

We then consider the transformation from $R^p \rightarrow R^p$ given by

$$\mathbf{u} \rightarrow \mathbf{x} = \boldsymbol{\mu} + \mathbf{A} \mathbf{u}$$

where $\mathbf{A} \mathbf{A}^T = \boldsymbol{\Sigma}$. From theorem A.28 it follows that \mathbf{A} is regular. We obtain

$$\mathbf{u} = \mathbf{A}^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$

giving

$$\begin{aligned} \mathbf{u}^T \mathbf{u} &= (\mathbf{x} - \boldsymbol{\mu})^T (\mathbf{A}^{-1})^T \mathbf{A}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}). \end{aligned}$$

Furthermore, since

$$\det(\boldsymbol{\Sigma}) = \det(\mathbf{A} \mathbf{A}^T) = \det(\mathbf{A})^2,$$

i.e.

$$\det(\mathbf{A}^{-1}) = \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}}$$

and the result follows from the theorem on the distribution of transformed random variables. ■

We note that the inverse variance-covariance matrix $\boldsymbol{\Sigma}^{-1}$ is often called the *precision* of the normal distribution.

If $\boldsymbol{\Sigma}$ is not regular, then the distribution is degenerate and has no density. We then introduce the concept of the affine support in the following definition.

|||| Definition 1.13

Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. By the (*affine*) *support* for \mathbf{X} we mean the smallest (side-) sub-space of R^p , where \mathbf{X} is defined with probability 1.

|||| **Remark 1.14**

If we restrict the considerations to the affine support, then X is regularly distributed and has a density as shown in theorem 1.12.

We have different possibilities of determining the support of a p -dimensional normal distribution. Firstly

|||| **Theorem 1.15**

Let $X \sim N_p(\mu, \Sigma)$, and let \mathbf{A} be an $p \times m$ matrix, so that $\mathbf{A} \mathbf{A}^T = \Sigma$. We then let V equal \mathbf{A} 's projection-space, i.e.

$$V = \{v \in \mathbb{R}^p \mid \exists u \in \mathbb{R}^m : v = \mathbf{A} u\}.$$

Then the (affine) support for X is the (side-) sub-space

$$\mu + V = \{\mu + v \mid v \in V\}.$$

|||| **Proof omitted**

Further, we have

|||| **Theorem 1.16**

Let X be as in the previous theorem. Then the subspace V equals the direct sum of the eigen-spaces corresponding to those eigenvalues in Σ which are different from 0.

|||| **Proof omitted**

Finally we have

|||| Theorem 1.17

Let X be as in the previous theorems. Then the subspace V equals the orthogonal complement to the null-space for Σ , i.e.

$$V = \{v | \Sigma v = \mathbf{0}\}^\perp$$

|||| Proof omitted

The three theorems are illustrated in

|||| Example 1.18

We consider

$$X \sim N \left(\begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix} \right) = N(\mu, \Sigma).$$

Since

$$\det \begin{pmatrix} 1 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{pmatrix} = 0,$$

then X is singularly distributed, and we will determine the affine support.

We first seek a matrix A , so $A A^T = \Sigma$, by determining Σ 's eigenvalues and (normed) eigenvectors. These are

$$\begin{aligned} \lambda_1 &= 9 \quad \wedge \quad p_1 = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}, \\ \lambda_2 &= 2 \quad \wedge \quad p_2 = \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}, \\ \lambda_3 &= 0 \quad \wedge \quad p_3 = \begin{bmatrix} \frac{2\sqrt{2}}{3} \\ -\frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \end{bmatrix}. \end{aligned}$$

It now follows that

$$\Sigma = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \\ \frac{2}{3} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \end{bmatrix} \begin{bmatrix} 9 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{2\sqrt{2}}{3} & -\frac{\sqrt{2}}{6} & -\frac{\sqrt{2}}{6} \end{bmatrix}$$

From this we see that we as **A-matrix** can choose

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & -1 & 0 \end{bmatrix} \quad (= \begin{bmatrix} \frac{1}{3} & 0 & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \\ \frac{2}{3} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \end{bmatrix} \begin{bmatrix} \sqrt{9} & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}).$$

If we regard \mathbf{A} as the matrix for a linear projection $\mathbb{R}^3 \rightarrow \mathbb{R}^3$ we then obtain that the projection-space is

$$\begin{aligned} V &= \{\mathbf{A} \mathbf{u} | \mathbf{u} \in \mathbb{R}^3\} \\ &= \{u_1 \mathbf{p}_1 + u_2 \mathbf{p}_2 | u_1 \in \mathbb{R} \wedge u_2 \in \mathbb{R}\}. \end{aligned}$$

It is immediately noted that this is also the direct sum of the eigen-spaces corresponding to the eigenvalues which are different from 0.

The null-space for Σ is given by

$$\Sigma \mathbf{u} = \mathbf{0} \iff \mathbf{u} = t \cdot \mathbf{p}_3.$$

This again gives the same description of V .

The affine support for \mathbf{Y} is then the (side-) sub-space

$$\mu + V = \left\{ \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} + u_1 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix} \mid u_1, u_2 \in \mathbb{R} \right\}.$$

|||| Remark 1.19

From the example the proofs of theorems 1.15-1.17 can nearly be deduced completely.

We now formulate a trivial but useful theorem.

|||| Theorem 1.20

Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$. Then

$$\mathbf{A} \mathbf{X} + \mathbf{b} \sim N(\mathbf{A} \boldsymbol{\mu} + \mathbf{b}, \mathbf{A} \Sigma \mathbf{A}^T),$$

where we implicitly require that the implied matrix-products etc. exist.

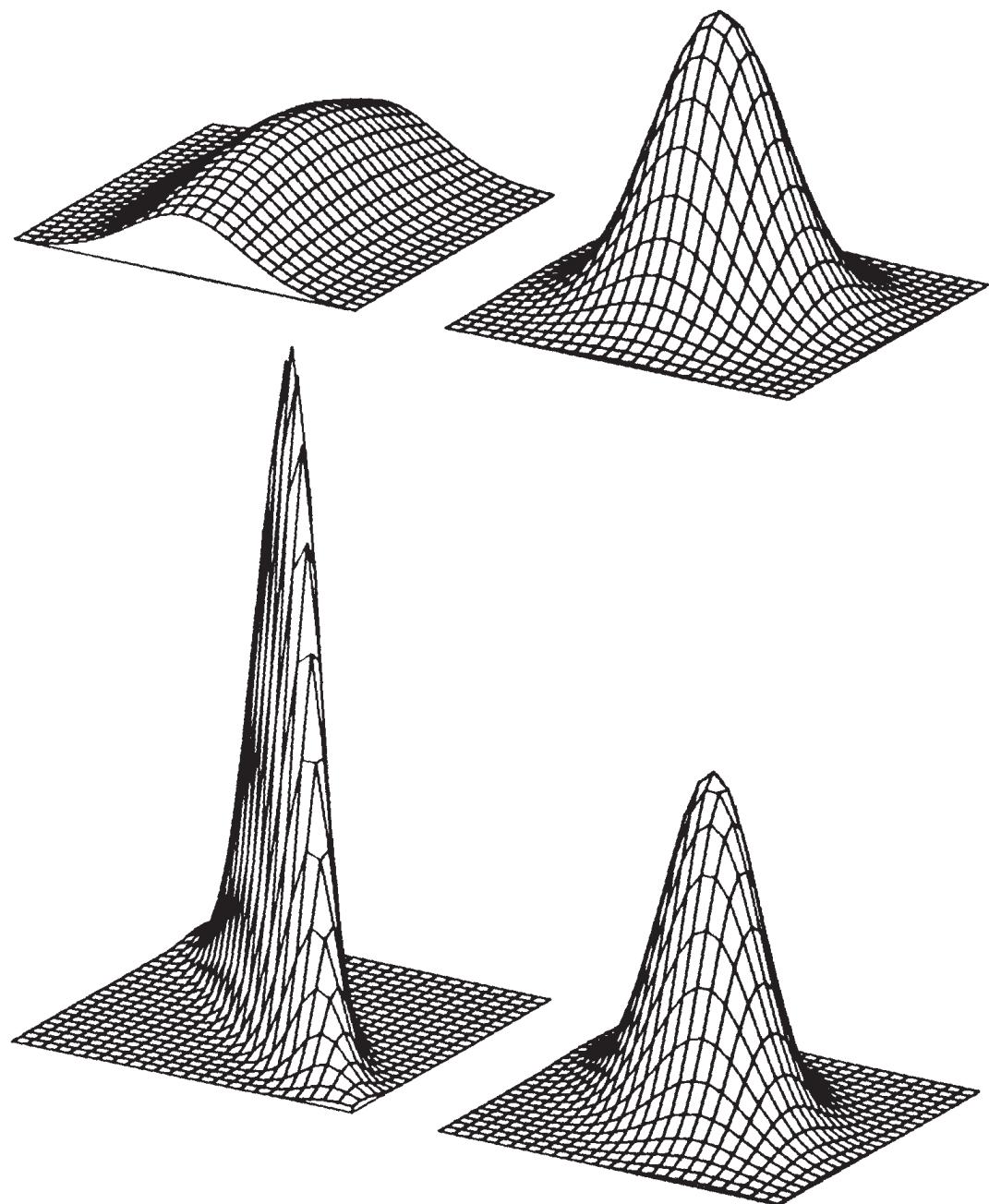


Figure 1.1: Density functions for two-dimensional normal distributions with the variance-covariance matrices:

$$\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}.$$

|||| **Proof**

Trivial from the definition.

■

1.2.2 Independence and contour ellipsoids.

In this section we will give the conditions for *independence* of the normally distributed random variables, and we will prove that the isosets for the density functions are ellipsoids, i.e. *contour ellipsoids*. First we have

|||| **Theorem 1.21**

Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right).$$

Then

$$X_i \sim N(\boldsymbol{\mu}_i, \Sigma_{ii}),$$

and

$$X_1, X_2 \text{ are stochastically independent} \Leftrightarrow \Sigma_{12} = \Sigma_{21}^T = \mathbf{0},$$

where $\mathbf{0}$ is a null matrix.

|||| **Proof**

The first statement follows from the previous theorem. The second follows by proving that the condition $\Sigma_{12} = \mathbf{0}$ assures, that the distribution becomes a product distribution.

■

From the theorem follows that the components in a vector $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ are stochastically independent if Σ is a diagonal matrix. We will now show that independence is just a question of choosing a suitable coordinate-system.

Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\boldsymbol{\Sigma}$ have the ortho-normed eigenvectors $\mathbf{p}_1, \dots, \mathbf{p}_n$. We now consider a coordinate system, with origo in $\boldsymbol{\mu}$ and the vectors $\mathbf{p}_1, \dots, \mathbf{p}_n$ as base-vectors. The coordinates in this system are called \mathbf{y} .

If we let

$$\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_n),$$

we have the following correspondence between the original coordinates \mathbf{x} and the new coordinates \mathbf{y} for any point $\in \mathbb{R}^n$.

$$\mathbf{y} = \mathbf{P}^T(\mathbf{x} - \boldsymbol{\mu}) \Leftrightarrow \mathbf{x} = \mathbf{P}\mathbf{y} + \boldsymbol{\mu},$$

cf. p. 343.

Note: The above relation is a relation between coordinates for a fixed vector viewed in two coordinate-systems.

Let \mathbf{Y} be the new coordinates for \mathbf{X} , then we have

|||| Theorem 1.22

Let $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let \mathbf{Y} be as above. Then

$$\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Lambda}),$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with $\boldsymbol{\Sigma}$'s eigenvalues on the diagonal.

|||| Proof

Follows from theorem 1.20 and theorem A.23.

■

|||| Remark 1.23

By translating and rotating (or reflection of) the original coordinate-system we have obtained that the variance-covariance matrix is a diagonal matrix, i.e. that the components in the stochastic vector are uncorrelated and thereby also independent.

By rescaling the axes we can even obtain that the variance-covariance matrix has zeros or ones on the diagonal. Considering the base-vectors

$$c_1 \mathbf{p}_1, \dots, c_n \mathbf{p}_n,$$

where

$$c_i = \begin{cases} \frac{1}{\sqrt{\lambda_i}} & \text{if } \lambda_i > 0 \\ 1 & \text{if } \lambda_i = 0 \end{cases},$$

cf. the proof of theorem A.24, and calling the coordinates in this system \mathbf{z} , we get the equation

$$\mathbf{z} = \mathbf{C}^T \mathbf{P}^T (\mathbf{x} - \boldsymbol{\mu}) = (\mathbf{P} \mathbf{C})^T (\mathbf{x} - \boldsymbol{\mu}),$$

where $\mathbf{C} = \text{diag}(c_1, \dots, c_n)$.

If we let the \mathbf{z} -coordinates for \mathbf{X} equal \mathbf{Z} we get

$$\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{E}),$$

where

$$\mathbf{E} = (\mathbf{P} \mathbf{C})^T \boldsymbol{\Sigma} \mathbf{P} \mathbf{C} = \mathbf{C}^T \mathbf{P}^T \boldsymbol{\Sigma} \mathbf{P} \mathbf{C} = \mathbf{C}^T \boldsymbol{\Lambda} \mathbf{C}$$

has zeros or ones on the diagonal.

The transformation into the new bases is closely related to the isocurves for the density function for the normal distribution.

As mentioned earlier the density for an $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\begin{aligned} f(\mathbf{x}) &= k \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \\ &= k \cdot \exp\left(-\frac{1}{2}(\|\mathbf{x} - \boldsymbol{\mu}\|)^2\right). \end{aligned}$$

Therefore we have

$$f(\mathbf{x}) = k_1 \Leftrightarrow (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c,$$

where k_1 and c are constants. Since $\boldsymbol{\Sigma}^{-1}$ is positive definite, the isocurves

$$E_c = \{\mathbf{x} | f(\mathbf{x}) = k_1\}$$

will be ellipsoids, cf. theorem A.38. From theorem A.38 is also seen that the major axes in these ellipsoids are the eigenvectors for $\boldsymbol{\Sigma}^{-1}$, but from theorem A.25 we note that they are also eigenvectors for $\boldsymbol{\Sigma}$. In the new coordinates the densities become

$$g(\mathbf{y}) = k \cdot \exp\left(-\frac{1}{2} \sum \frac{1}{\lambda_i} y_i^2\right),$$

where λ_i is the i 'th eigenvalue for Σ , and

$$h(z) = k_1 \cdot \exp\left(-\frac{1}{2}\sum z_i^2\right).$$

The ellipsoids E_i are often called *contour-ellipsoids*. The relation to the Chi-Square (χ^2) distribution is given in the following theorem.

|||| Theorem 1.24

Let \mathbf{P} and \mathbf{C} be as above. Then

$$(\mathbf{X} - \boldsymbol{\mu})^T (\mathbf{P} \mathbf{C}) (\mathbf{P} \mathbf{C})^T (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2(\text{rk } \Sigma).$$

If Σ has full rank p then

$$(\mathbf{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) = \|\mathbf{X} - \boldsymbol{\mu}\|^2 \sim \chi^2(p).$$

|||| Proof

$$(\mathbf{X} - \boldsymbol{\mu})^T (\mathbf{P} \mathbf{C}) (\mathbf{P} \mathbf{C})^T (\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Z}^T \mathbf{Z} = \Sigma \delta_i Z_i^2,$$

where $\delta_i = 1$ if $\lambda_i \neq 0$ and equal to 0 otherwise.

Since the non-degenerate components in \mathbf{Z} are stochastically independent and $N(0,1)$ -distributed the result follows immediately. The last remark comes from

$$\mathbf{P} \mathbf{C} (\mathbf{P} \mathbf{C})^T = \mathbf{P} \mathbf{C} \mathbf{C}^T \mathbf{P}^T = \mathbf{P} \Lambda^{-1} \mathbf{P}^T = \Sigma^{-1}$$

■

|||| Remark 1.25

The result of the theorem is that the probability of an outcome being within the contour ellipsoid can be computed using a χ^2 -distribution.

Examples of these concepts will be given in example 1.33, where we consider the two-dimensional normal distribution.

1.2.3 Conditional distributions

In this section we consider the partitioning of a random variable $\mathbf{X} \sim \mathbf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, into

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}; \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}; \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

We then have

|||| Theorem 1.26

If \mathbf{X}_2 is regularly distributed, i.e. if $\boldsymbol{\Sigma}_{22}$ has full rank, then the distribution of \mathbf{X}_1 conditioned on $\mathbf{X}_2 = \mathbf{x}_2$ is again a normal distribution, and the following holds

$$\begin{aligned} E(\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2) &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ D(\mathbf{X}_1 | \mathbf{X}_2 = \mathbf{x}_2) &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. \end{aligned}$$

If $\boldsymbol{\Sigma}_{22}$ does not have full rank then the conditional distribution is still normal and $\boldsymbol{\Sigma}_{22}^{-1}$ in the above equations should be substituted by a generalised inverse $\boldsymbol{\Sigma}_{22}^-$.

|||| Proof omitted

The proof is technical and is omitted. The result is shown for $p = 2$ in section 1.2.6.

|||| Remark 1.27

It is seen that the conditional dispersion of \mathbf{X}_1 is independent of \mathbf{x}_2 . This result is not valid for all distributions, but is special for the normal distribution. Also we see the conditional mean is an affine function of \mathbf{x}_2 , cf. the discussion in section 1.3.3. Furthermore, we see that the conditional dispersion equals the Schur Complement of the nonconditional dispersion of \mathbf{X}_2

We will not discuss the implications of the theorem here. Instead we refer to the examples in section 1.2.5.

1.2.4 Theorem of reproducivity and the central limit theorem.

Analogous to the theorem of reproducivity for the univariate normal distribution we have

|||| Theorem 1.28 Theorem of reproducivity

Let X_1, \dots, X_k be independent, and let $X_i \sim N(\mu_i, \Sigma_i)$.

Then

$$\sum_{i=1}^k X_i \sim N\left(\sum_{i=1}^k \mu_i, \sum_{i=1}^k \Sigma_i\right).$$

|||| Proof omitted

As in the univariate case, central limit theorems exist, i.e. sums of independent multidimensional stochastic variables are under generel assumptions asymptotically normally distributed. We state an analogue to Lindeberg-Levy's theorem.

|||| Theorem 1.29 Central limit theorem

Let the independent and identically distributed variables X_1, \dots, X_n , have finite first and second moments

$$\mu = E(X_i), \quad \Sigma = D(X_i).$$

Then we have - with $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$ - that

$$\sqrt{n}(\bar{X}_n - \mu)$$

has an $N(\mathbf{0}, \Sigma)$ -distribution as its limiting distribution, and we say that \bar{X}_n is asymptotically $N(\mu, \frac{1}{n}\Sigma)$ distributed.

|||| Proof

This and the previous theorem can be proved from the corresponding univariate theorems by first using a theorem, which characterises the multivariate distribution

(a multidimensional variable is normally distributed if and only if all linear combinations of its components are (univariate) normally distributed; and by using a theorem which characterises a multivariate limiting distribution as limiting distributions of linear combinations of the components (coordinates). However, this is out of the scope of this presentation and the interested reader is referred to the literature e.g. [1], section 2c.5.

■

1.2.5 Estimation of the parameters in a multivariate normal distribution.

We consider a number of observations X_1, \dots, X_n , which are assumed independent and identically $N_p(\mu, \Sigma)$ distributed. We assume there are more observations than the dimension indicates, i.e. that $n > p$. In this section we will give estimates of the parameters μ and Σ .

We introduce the notation

$$\begin{aligned} \mathbf{X}_i &= \begin{bmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{bmatrix} \\ \bar{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix} \\ \mathbf{S} &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = \frac{1}{n-1} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T - \frac{n}{n-1} \bar{\mathbf{X}} \bar{\mathbf{X}}^T. \end{aligned}$$

If we consider the *data-matrix*

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T \\ \vdots \\ \mathbf{X}_n^T \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix},$$

where the i 'th row corresponds to the i 'th observation, we can also write

$$\begin{aligned} \bar{\mathbf{X}} &= \frac{1}{n} \mathbf{X}^T \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \frac{1}{n} \mathbf{X}^T \mathbf{1} \\ (n-1)\mathbf{S} &= \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T = \mathbf{X}^T \mathbf{X} - n \bar{\mathbf{X}} \bar{\mathbf{X}}^T = \mathbf{X}^T \mathbf{X} - \frac{1}{n} \mathbf{X}^T \mathbf{1} \mathbf{1}^T \mathbf{X}. \end{aligned}$$

With this we can now state

|||| Theorem 1.30

Let the situation be as stated above. Then the maximum likelihood estimators for μ and Σ equal

$$\begin{aligned}\hat{\mu} &= \bar{\mathbf{X}} \\ \hat{\Sigma} &= \frac{n-1}{n} \mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \\ &= \frac{1}{n} \mathbf{X}^T \mathbf{X} - \frac{1}{n^2} \mathbf{X}^T \mathbf{1} \mathbf{1}^T \mathbf{X}\end{aligned}$$

$\hat{\mu}$ is an unbiased estimate of μ , and \mathbf{S} is an unbiased estimate of Σ .

|||| Proof omitted

See e.g. [2], chapter 3.

|||| Remark 1.31

Since the empirical variance-covariance matrix \mathbf{S} is an unbiased estimate Σ , and since it only differs from the maximum likelihood estimator by the factor $\frac{n}{n-1}$, we often prefer \mathbf{S} as the estimate. Often one will see the notation $\hat{\Sigma}$ used for \mathbf{S} . One should in each case be aware of what the expression $\hat{\Sigma}$ precisely means.

The distribution of $\hat{\mu}$ comes trivially from theorem 1.2.4. The following holds

$$\hat{\mu} = \bar{\mathbf{X}} \sim N_p(\mu, \frac{1}{n} \Sigma).$$

The distribution of \mathbf{S} is the Wishart distribution, the multivariate analogue to the Chi-Square distribution. It is treated in section 1.5.

We give an example of estimating the parameters in the following section.

1.2.6 The two-dimensional normal distribution.

We now specialise the results from before to two dimensions.

Let $\mathbf{X} = \begin{bmatrix} Y \\ X \end{bmatrix}$ be normally distributed with $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_Y \\ \mu_X \end{bmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_Y^2 & \sigma_{YX} \\ \sigma_{YX} & \sigma_X^2 \end{bmatrix}.$$

Since

$$\det(\boldsymbol{\Sigma}) = \sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2$$

we have for $\det(\boldsymbol{\Sigma}) \neq 0$,

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \begin{bmatrix} \sigma_X^2 & -\sigma_{YX} \\ -\sigma_{YX} & \sigma_X^2 \end{bmatrix}.$$

Introducing the correlation coefficient ρ

$$\rho = \frac{\sigma_{YX}}{\sigma_Y \sigma_X},$$

we get

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{1 - \rho^2} \begin{bmatrix} \frac{1}{\sigma_Y^2} & \frac{-\rho}{\sigma_Y \sigma_X} \\ \frac{-\rho}{\sigma_Y \sigma_X} & \frac{1}{\sigma_X^2} \end{bmatrix},$$

and the density becomes

$$\begin{aligned} f(y, x) = & \\ & \frac{1}{2\pi} \frac{1}{\sigma_Y \sigma_X \sqrt{1 - \rho^2}} \exp \left[-\frac{1}{2} \frac{1}{1 - \rho^2} \left\{ \left[\frac{y - \mu_Y}{\sigma_Y} \right]^2 \right. \right. \\ & \left. \left. - 2\rho \frac{y - \mu_Y}{\sigma_Y} \frac{x - \mu_X}{\sigma_X} + \left[\frac{x - \mu_X}{\sigma_X} \right]^2 \right\} \right]. \end{aligned}$$

The graph is shown in fig. 1.2 It is immediately seen that we have a product distribution i.e. that Y and X are stochastically independent, if $\rho = 0$, i.e. if $\boldsymbol{\Sigma}$ is a diagonal matrix.

The conditional distribution of Y conditioned on $X = x_2$ is proportional to the intersecting curve between the plane through $(0, x, 0)$ parallel to the (1)-(3) plane. If we denote the density as g we have

$$g(\cdot) = c f(\cdot, x),$$

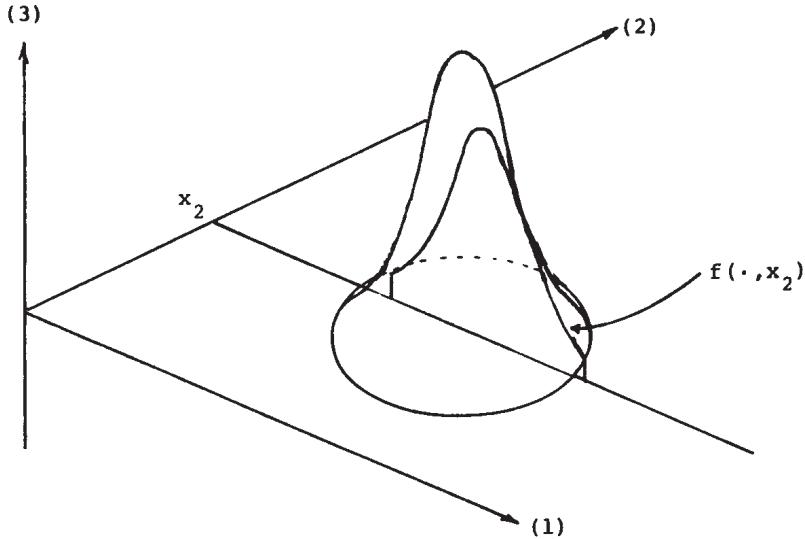


Figure 1.2: The density of a two-dimensional normal distribution.

where c is a normalisation constant. We have

$$\begin{aligned}
 g(y) &= k_1 \cdot \exp \left[-\frac{1}{2} \frac{1}{1-\rho^2} \left\{ \left[\frac{y - \mu_Y}{\sigma_Y} \right]^2 - 2\rho \frac{y - \mu_Y}{\sigma_Y} \frac{x - \mu_X}{\sigma_X} \right\} \right] \\
 &= k_2 \cdot \exp \left[-\frac{1}{2} \frac{1}{1-\rho^2} \left[\frac{y - \mu_Y}{\sigma_Y} - \rho \frac{x - \mu_X}{\sigma_X} \right]^2 \right] \\
 &= k_3 \cdot \exp \left[-\frac{1}{2} \frac{1}{\sigma_Y^2(1-\rho^2)} \left(y - \mu_Y - \rho \frac{\sigma_Y}{\sigma_X} (y - \mu_Y) \right)^2 \right] \\
 &= k_3 \cdot \exp \left[-\frac{1}{2\gamma^2} (y - \xi_1)^2 \right].
 \end{aligned}$$

Note that no bookkeeping has been done with respect to x . It has disappeared into different constants. From the final result we note that the conditional distribution is normal and that

$$k_3 = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}},$$

and finally that

$$E(Y|X=x) = \xi_1 = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X} (x - \mu_X)$$

and

$$V(Y|X=x) = \gamma^2 = \sigma_1^2(1-\rho^2).$$

We have shown the result of theorem 1.26 for the case $n = 2$. Note, that the conditional mean depends linearly (or more correctly: affinely) upon x_2 , and that the conditional variance is independent of x_2 . Further we have

$$V(Y|X=x) \leq V(Y),$$

and the squared coefficient of correlation represents the reduction in variance. i.e. the fraction of Y 's variance, which can be explained by X , since

$$\rho^2 = \frac{V(Y) - V(Y|X=x)}{V(Y)}.$$

In the following example we consider a numerical example which also involves an estimation problem.

||| Example 1.32

In the following table is shown corresponding values of the air's content of airborne particle matter measured in $\frac{\mu\text{g}}{\text{m}^3}$. Two different measuring principles were used, a measure of grey-value (using a so-called OECD instrument) and a weighing principle (using a so-called High Volume Sampler). Among other things the reason for the large deviations is that the measurements using the grey value principle are sensitive to the deviation of the suspended dust particles from "normal dust". In this way, a large content of calcium dust in the air could result in the measurements being systematically too small.

	I	2	5	15	16	16	19	26	24	16	36
Method	II	2	12	4	21	41	14	31	29	31	8
	I	39	42	44	40	42	42	50	51	58	64
	II	30	44	26	60	34	34	14	41	58	47

We consider this data as being observations from independent identically distributed stochastic variables

$$\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \dots, \begin{bmatrix} X_{20} \\ Y_{20} \end{bmatrix}.$$

We will examine whether we can assume the distribution is normal with parameters (μ, Σ) . If the distribution is normal, we find the estimates

$$\hat{\mu} = \begin{bmatrix} \hat{\mu}_X \\ \hat{\mu}_Y \end{bmatrix} = \begin{bmatrix} \bar{X} \\ \bar{Y} \end{bmatrix} = \begin{bmatrix} 32.35 \\ 29.05 \end{bmatrix},$$

and

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma}_X^2 & \hat{\sigma}_{XY} \\ \hat{\sigma}_{XY} & \hat{\sigma}_Y^2 \end{bmatrix} = \begin{bmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{bmatrix} = \begin{bmatrix} 311 & 182 \\ 182 & 279 \end{bmatrix},$$

where $\hat{\Sigma}$ is the unbiased estimate of Σ . Specially we have

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

We now want to check if the observations can be assumed to come from a normal distribution with parameters $(\hat{\mu}, \hat{\Sigma})$. To do that we first estimate the contour ellipses.

The eigenvalues and eigenvectors for $\hat{\Sigma}$ are

$$\hat{\lambda}_1 = 477.613 \quad \text{and} \quad \hat{p}_1 = \begin{bmatrix} 0.736 \\ 0.678 \end{bmatrix}$$

and

$$\hat{\lambda}_2 = 112.676 \quad \text{and} \quad \hat{p}_2 = \begin{bmatrix} -0.678 \\ 0.736 \end{bmatrix}.$$

If we choose the coordinate system with origo in $\hat{\mu}$ and with \hat{p}_1 and \hat{p}_2 as base vectors, the contour ellipsoids have equations of the form

$$\frac{z_1^2}{\hat{\lambda}_1} + \frac{z_2^2}{\hat{\lambda}_2} = c,$$

or

$$\frac{z_1^2}{477.613} + \frac{z_2^2}{112.676} = c,$$

where the new coordinates are given by

$$\mathbf{P} z = (\mathbf{p}_1 \ \mathbf{p}_2) z = x - \hat{\mu}.$$

In figure 1.2 we show the observations and 3 contour ellipses corresponding to the c -values $c_1 = \chi^2(2)_{0.40} = 1.02$, $c_2 = \chi^2(2)_{0.80} = 3.22$ and $c_3 = \chi^2(2)_{0.95} = 5.99$. This has the effect (see theorem 1.24) that in the normal distribution with parameters $(\hat{\mu}, \hat{\Sigma})$ we have the probabilities 40%, 80% and 95% of getting observations within the inner, the middle and the outer ellipse. For the areas between the ellipses resp. outside these, we have the probabilities 40%, 40%, 15% and 5%. These numbers can be compared to the corresponding observed relative probabilities 40%, 30%, 30% and 0%. The fit is - if not overwhelming - at least acceptable.

If one wants a more precise result, one can perform a χ^2 -test. It would then be reasonable to divide the plane further according to the eigenvectors. In the case shown, this would result in 4×4 areas with estimated probabilities of 10%, 10%, 3.75% and 1.25%. One can then compute the usual χ^2 test-statistic:

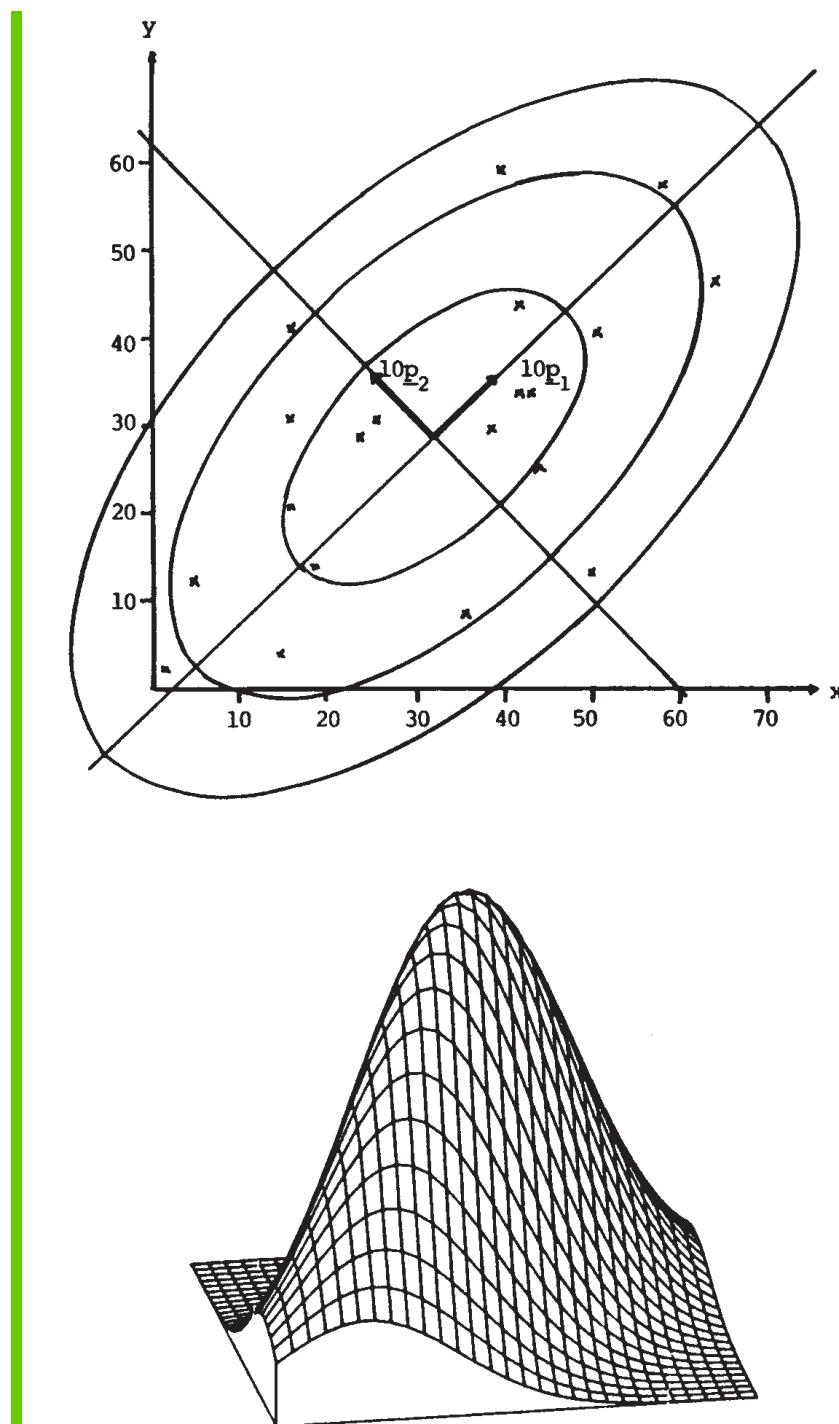


Figure 1.32: Estimated contour ellipses and estimated density function corresponding to the data in example 1.32

$$\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

and compare it with a $\chi^2(n - 6)$ distribution (we have estimated 5 parameters). In

the present case there are not really enough observations to perform this analysis. The correlation coefficient is estimated at

$$\hat{\rho} = \frac{182}{\sqrt{311 \cdot 279}} = 0.62,$$

and the conditional variances are estimated at

$$\begin{aligned}\hat{V}(X|Y=y) &= 311(1 - \hat{\rho}^2) = 192 \\ \hat{V}(Y|X=x) &= 279(1 - \hat{\rho}^2) = 172.\end{aligned}$$

We see, that the conditional variances have been reduced by 38% corresponding to $\rho^2 = 0.38$. That the conditional variance of e.g. an OECD-measurement for given High Volume Sampler measurement is substantially less than the unconditional variance seems rather reasonable. If we eg. find, that the amount of suspended matter measured using a High Volume Sampler is found as e.g. $2 \frac{\mu g}{m^3}$, we would not expect to get results from the OECD-instrument, which deviate grossly. This corresponds to a small conditional variance. If the result from the High Volume Sampler is unknown, then we must expect a measurement from the OECD-instrument that can lie anywhere in its natural range of variation - corresponding to a larger unconditional variance.

1.3 Correlation and regression

In this section we will discuss the meaning of parameters in a multidimensional normal distribution in greater detail. First we will try to generalise the properties of the correlation coefficient seen in the previous section.

1.3.1 The partial correlation coefficient.

The starting point is the formula for the conditional distributions in a multidimensional normal distribution. Let $Z \sim N_p(\mu, \Sigma)$, and let the variables be partitioned as follows

$$Z = \begin{bmatrix} Y \\ X \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}; \quad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix},$$

where Y consists of the m first elements in Z and X the following k elements. Then the conditional dispersion of Y for given $X = x$ is, as was shown in theorem 1.26, equal to

$$D(Y|X=x) = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}.$$

By the *partial correlation coefficient* between Y_i and Y_j , $i, j \leq m$, conditioned on (or: for given) $X = x$ we will understand the correlation in the conditional distribution of Y given that $X = x$. It is denoted by $\rho_{y_i y_j | x_1, \dots, x_k}$.

Let

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{1p} & \cdots & \sigma_p^2 \end{bmatrix}$$

and

$$\Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{1m} & \cdots & a_{mm} \end{bmatrix},$$

we now have

$$\rho_{Y_i Y_j | X_1, \dots, X_k} = \frac{a_{ij}}{\sqrt{a_{ii}} \sqrt{a_{jj}}}.$$

For the special case of

$$\mathbf{Z} = \begin{bmatrix} Y_1 \\ Y_2 \\ X \end{bmatrix}; \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_{y_1} \\ \mu_{y_2} \\ \mu_x \end{bmatrix}; \quad \Sigma = \begin{bmatrix} \Sigma_{y_1 y_1} & \Sigma_{y_1 y_2} & \Sigma_{y_1 x} \\ \Sigma_{y_2 y_1} & \Sigma_{y_2 y_2} & \Sigma_{y_2 x} \\ \Sigma_{x y_1} & \Sigma_{x y_2} & \Sigma_{xx} \end{bmatrix},$$

being three dimensional we have with

$$\Sigma = \begin{bmatrix} \sigma_{y_1}^2 & \rho_{y_1 y_2} \sigma_{y_1} \sigma_{y_2} & \rho_{y_1 x} \sigma_{y_1} \sigma_x \\ \rho_{y_1 y_2} \sigma_{y_1} \sigma_{y_2} & \sigma_{y_2}^2 & \rho_{y_2 x} \sigma_{y_2} \sigma_x \\ \rho_{y_1 x} \sigma_{y_1} \sigma_x & \rho_{y_2 x} \sigma_{y_2} \sigma_x & \sigma_x^2 \end{bmatrix},$$

that

$$\begin{aligned} & \Sigma_{yy} - \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy} \\ &= \begin{bmatrix} \sigma_{y_1}^2 & \rho_{y_1 y_2} \sigma_{y_1} \sigma_{y_2} \\ \rho_{y_1 y_2} \sigma_{y_1} \sigma_{y_1} & \sigma_{y_2}^2 \end{bmatrix} - \frac{1}{\sigma_x^2} \begin{bmatrix} \rho_{y_1 x}^2 \sigma_{y_1}^2 \sigma_x^2 & \rho_{y_1 x} \rho_{y_2 x} \sigma_{y_1} \sigma_{y_2} \sigma_x^2 \\ \rho_{y_1 x} \rho_{y_2 x} \sigma_{y_1} \sigma_{y_2} \sigma_x^2 & \rho_{y_2 x}^2 \sigma_{y_2}^2 \sigma_x^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{y_1}^2 (1 - \rho_{y_1 x}^2) & \sigma_{y_1} \sigma_{y_2} (\rho_{y_1 y_2} - \rho_{y_1 x} \rho_{y_2 x}) \\ \sigma_{y_1} \sigma_{y_2} (\rho_{y_1 y_2} - \rho_{y_1 x} \rho_{y_2 x}) & \sigma_{y_2}^2 (1 - \rho_{y_2 x}^2) \end{bmatrix}. \end{aligned}$$

From this follows that the partial correlation coefficient between Y_1 and Y_2 conditioned on X is

$$\rho_{y_1 y_2 | x} = \frac{\rho_{y_1 y_2} - \rho_{y_1 x} \rho_{y_2 x}}{\sqrt{(1 - \rho_{y_1 x}^2)(1 - \rho_{y_2 x}^2)}}.$$

For the p -dimensional vector \mathbf{Z} we therefore find

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik} \rho_{jk}}{\sqrt{(1 - \rho_{ik}^2)(1 - \rho_{jk}^2)}}. \quad (**)$$

Since it is possible to find conditional distributions for given X_{m+1}, \dots, X_p by successive conditionings we can therefore determine partial correlation coefficients of higher order by successive use of (**). E.g. we find

$$\rho_{ij|kl} = \frac{\rho_{ij|k} - \rho_{il|k} \cdot \rho_{jl|k}}{\sqrt{(1 - \rho_{il|k}^2) \cdot (1 - \rho_{jl|k}^2)}},$$

here we have first conditioned on X_k and then conditioned on X_l .

In section 1.2.6 we saw that the (squared) correlation coefficient is a measure of the reduction in variance if we condition on one of the variables. Since the partial correlation coefficients are just correlations in conditional distributions we can use the same interpretation here. We have e.g. that $\rho_{ij|kl}^2$ gives the fraction of X_i 's variance for given $X_k = x_k$ and $X_l = x_l$ which is explained by X_j . It should be emphasised that *these interpretations are strongly dependent on the assumption of normality*. For the general case the conditioned variances will depend on the values with which they are conditioned (i.e. depend on x_k and x_l).

When estimating the *partial correlations* one just estimates the variance-covariance matrix and then computes the partial correlations as shown. If the estimate of the variance-covariance matrix is a maximum-likelihood estimator then the estimates of the partial correlations computed in this way will also be maximum likelihood estimates.

We will now illustrate the concepts in

||| Example 1.33

(Data are from [3]).

In the table below correlation coefficients between 3- and 28-day strengths for Portland Cement and the content of minerals C₃S (Alit, Tricalciumsilicat Ca₃SiO₅) and C₃A (Aluminat, Tricalciumaluminat, Ca₃Al₂O₆), and the degree of fine-grainedness (BLAINE) are given. The correlations are estimated using 51 corresponding observations.

	C ₃ S	C ₃ A	BLAINE	Strength 3	Strength 28
C ₃ S	1	-0.309	0.091	0.158	0.344
C ₃ A	-0.309	1	0.192	0.120	-0.166
BLAINE	0.091	0.192	1	0.745	0.320
Strength 3	0.158	0.120	0.745	1	0.464
Strength 28	0.344	-0.166	0.320	0.464	1

The correlation matrix for 5 cement variables.

It should be noted that C_3S constitutes about 35-60% of normal portland clinkers and C_3A is about 5-18% of clinker. The BLAINE is a measure of the specific surface so that a large BLAINE corresponds to a very fine-grained cement.

We will be especially interested in the relationship between C_3A content in clinker and the two strengths. It is commonly accepted cf. the following figure, that a large content of C_3A gives a larger 3-day strength which is also in correspondence with $\hat{\rho}_{C_3A, Strength3} = 0.120$. The problem is that this larger 3-day strength for cement with large content of C_3A only depends on C_3A 's larger degree of hydratization (the faster the water reacts with the cement the faster it will have greater strength).

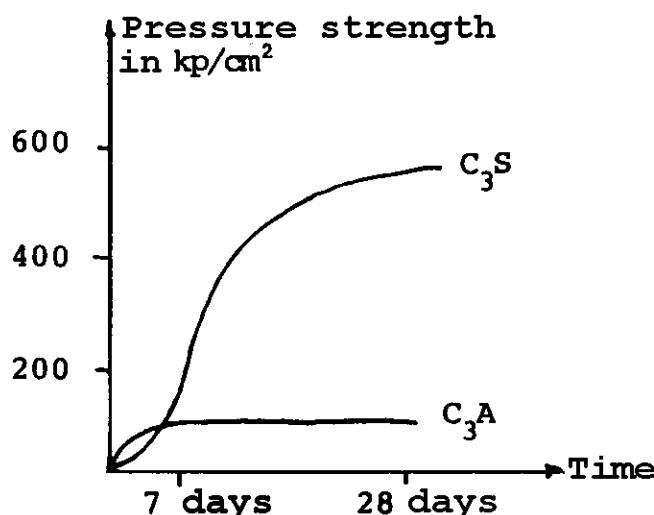


Figure A: Strength by pressure test at ordinary temperature of paste of C_3S and C_3A seasoned for different amounts of time. (from [4]).

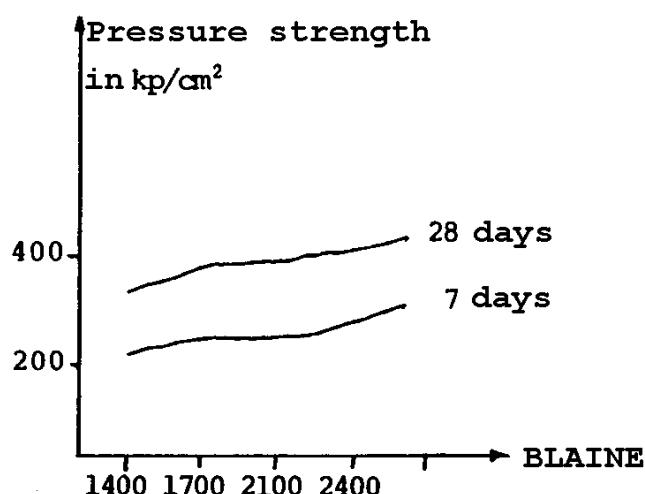


Figure B: Pressure strengths for different fine-grainedness of the cement. (from [4]).

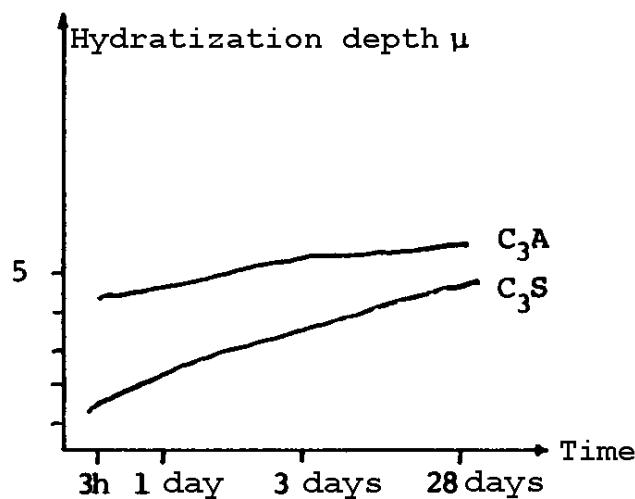


Figure C: Degree of hydratation for cement minerals and their dependence on time (from [4]).

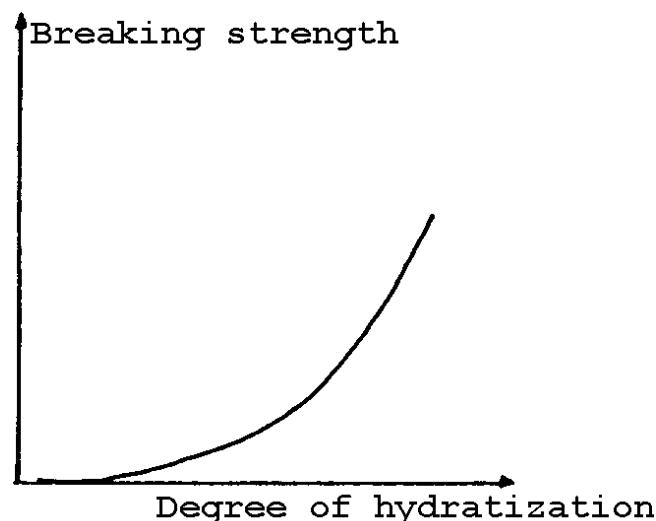


Figure D: Relationship between degree of hydratation and strength (from [4]).

C_3A 's far greater hydratization after 3 days is seen in figure C, and the degree of hydratization and its influence on the strengths has been sketched in figure D

If we look at the correlation matrix we also see that the content of C_3A is positively correlated with the BLAINE i.e. cements with a very high content of C_3A will usually be very fine-grained and as it is seen in figure B this should also help increase the strength.

Finally we see that the 28-day strength is slightly negatively correlated with the content of C_3A . This does not seem strange if we consider the temporal dependence of C_3S 's and C_3A 's as seen in e.g. in figure A, even though the finer grain (for cement with large content of C_3A) should also be seen in the 28-day strength, figure B.

In order to separate the different characteristics of C₃A from the effects which arise from a C₃A-rich cement seems to be easier to grind and therefore often is seen in a bit more fine-grained form. Therefore, we will estimate the conditional correlations for fixed value of BLAINE. These are seen in the table below

	C ₃ S	C ₃ A	Strength 3	Strength 28
C ₃ S	1	-0.333	0.137	0.333
C ₃ A	-0.333	1	-0.035	-0.246
Strength 3	0.137	-0.035	1	0.358
Strength 28	0.333	-0.246	0.358	1

Correlation matrix for 4 cement variables conditioned on BLAINE.

We see that the partial correlation coefficient between 3-day strength and C₃A for given fine-grainedness is negative (note the unconditioned correlation coefficient was positive). This implies that we for fixed fine-grainedness must expect that cements with a high content of C₃A will tend to have lower strengths. This might indicate that the large 3-day strength for cements with high content of C₃A rather depends on these cements having a large BLAINE (that they are crushed somewhat easier) than that C₃A hydrates quickly!

We see a corresponding effect on the correlation between C₃A and 28-day strength. Here the unconditional correlation is -0.168 and the partial correlation for fixed BLAINE has become -0.246.

||| Remark 1.34

The example above shows that one has to be very cautious in the interpretation of correlation coefficients. It would be directly misleading e.g. to say that a large content of C₃A assures a large 3-day strength. First of all it is not possible to conclude anything about the relation between two variables just by looking at their correlation. What you can conclude is that there seems to be a tendency that a high content of C₃A and a high 3-day strength appear at the same time. The reason for this could be that they both depend on a third but unknown factor without there having to be any direct relation between the two variables. Secondly we also see that going from unconditioned to partial correlations can even give a change of sign corresponding to an effect which is the opposite of that we get by a direct analysis. The reason for this is a correlation with a 3rd factor in this case BLAINE which disturbs the picture.

In many situations we would like to test if the correlation coefficient can be assumed to be 0. You can then use

|||| Theorem 1.35

Let $R = R_{ij|m+1\dots p}$ be the empirical partial correlation coefficient between Z_i and Z_j conditioned on (or: for given) Z_{m+1}, \dots, Z_p . It is assumed to be computed from the unbiased estimates of the variance-covariance matrix and from n observations. Then

$$\frac{R}{\sqrt{1-R^2}} \sqrt{n-2-(p-m)} \sim t(n-2-(p-m)),$$

if $\rho_{ij|m+1\dots p} = 0$.

|||| Proof omitted

|||| Remark 1.36

The number $(p - m)$ is the number of variables which are fixed (conditioned upon). The degrees of freedom are therefore equal to the number of observations minus 2 minus the number of fixed variables. *The theorem is also valid if $p - m = 0$ i.e. if we have the case of an unconditional correlation coefficient.*

We continue example 1.33 in

|||| Example 1.37

Let us investigate whether the value of $r_{24|3}$ is significantly different from 0. We find with $r_{24|3} = R$:

$$\begin{aligned} \frac{R}{\sqrt{1-R^2}} \sqrt{n-2-(p-m)} &= \frac{-0.035}{\sqrt{1-0.035^2}} \cdot \sqrt{51-2-(5-4)} \\ &= -0.243 = t(48)_{40\%}. \end{aligned}$$

A hypothesis that $\rho_{24|3}$ is 0 will therefore be accepted using a test at level α for $\alpha < 80\%$. (Note: this is by nature a two-sided test.)

If we wish to test other values of ρ or to determine confidence intervals we can use

|||| Theorem 1.38

Assume the situation is as in the previous theorem. We consider the hypothesis

$$H_0 : \rho_{ij|m+1,\dots,p} = \rho_0$$

versus

$$H_1 : \rho_{ij|m+1,\dots,p} \neq \rho_0.$$

We let

$$Z = \frac{1}{2} \log \frac{1 + R_{ij|m+1,\dots,p}}{1 - R_{ij|m+1,\dots,p}}$$

and

$$z_0 = \frac{1}{2} \log \frac{1 + \rho_0}{1 - \rho_0}.$$

Under H_0 we will have

$$(Z - z_0) \cdot \sqrt{n - (p - m) - 3} \text{ approx. } \sim N(0, 1).$$

|||| Proof omitted

|||| Example 1.39

Let us determine a 95% confidence interval for $\rho_{24|3}$ in example 1.37. We have

$$\begin{aligned} P\{-1.96 < (Z - z) \cdot \sqrt{51 - (5 - 4) - 3} < 1.96\} &\simeq 95\% \\ \Leftrightarrow P\{-1.96 - 6.86Z < -6.86z < 1.96 - 6.86Z\} &\simeq 95\% \\ \Leftrightarrow P\{Z - 0.29 < z < Z + 0.29\} &\simeq 95\%. \end{aligned}$$

The relationship between z and $\rho_{24|3} = \rho$ is

$$z = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} \Leftrightarrow \rho = \frac{e^{2z} - 1}{e^{2z} + 1}$$

The observed value of Z is

$$Z = \frac{1}{2} \log \frac{1 - 0.035}{1 + 0.035} = -0.03501.$$

The limits for z become

$$[-0.3250, 0.2549].$$

The corresponding limits for $\rho_{24|3}$ are

$$\left[\frac{e^{-0.6500} - 1}{e^{-0.6500} + 1}, \frac{e^{0.5098} - 1}{e^{0.5098} + 1} \right] = [-0.31, 0.25].$$

1.3.2 The multiple correlation coefficient

The partial correlation coefficient is one possible generalisation of the correlation between two variables. The partial correlations are mostly intended to describe the degree of relationship (correlation, covariance) between two variables. Instead we will now consider the formula on p. 28

$$\rho^2 = \frac{V(Y) - V(Y|X=x)}{V(Y)},$$

This is the "degree of reduction in variation" interpretation of the (squared) correlation coefficient. This we now seek to generalise. We again consider the partition of the p -dimensionally normally distributed vector Z in an m -dimensional vector Y and a $(p-m) = k$ -dimensional vector X , and the resulting partitioning of the parameters i.e.

$$Z = \begin{bmatrix} Y \\ X \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}; \quad \Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}.$$

We now define the multiple correlation coefficient between Y_i , $i = 1, \dots, m$ and X as the maximal correlation between Y_i and a linear combination of X 's elements. It is denoted $\rho_{y_i|x}$.

Any linear combination proportional to $\beta_i^T X$ will of course have the same correlation with Y_i . It can be shown that

$$\beta_i^T X = (\Sigma_{yx} \Sigma_{xx}^{-1})_i X,$$

where β_i^T is the i 'th row in the matrix $\Sigma_{yx} \Sigma_{xx}^{-1}$. This matrix appears in the expression for the conditional mean of Y given X . As stated before this is

$$E(Y|X=x) = \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mu_x) = \mu_y + \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_m^T \end{bmatrix} (x - \mu_x).$$

It can also be shown that

$$\inf_{\alpha} V(Y_i - \alpha^T X) = V(Y_i - \beta_i^T X),$$

i.e. the considered linear combination minimises the variance of $(Y_i - \alpha^T X)$.
cf. section 1.3.3

We now have the following important theorem

|||| Theorem 1.40

We consider the situation above. Let σ_i be the i 'th column in Σ_{xy} , i.e. σ_i^T is the i 'th row in Σ_{yx} . Further, let σ_{ii} denote the i 'th diagonal element, i.e. the variance of Y_i

Then

$$\rho_{y_i|x} = \frac{\sqrt{\sigma_i^T \Sigma_{xx}^{-1} \sigma_i}}{\sqrt{\sigma_{ii}}}.$$

If we let

$$\Sigma_i = \begin{bmatrix} \sigma_{ii} & \Sigma_i^T \\ \Sigma_i & \Sigma_{xx} \end{bmatrix},$$

then

$$1 - \rho_{y_i|x}^2 = \frac{\det \Sigma_i}{\sigma_{ii} \det \Sigma_{xx}} = \frac{V(Y_i|X)}{V(Y_i)},$$

|||| Proof

The proofs to the claims before the theorem are quite simple. One just has to use a Lagrange multiplier and also use that the variance-covariance matrix is positive semidefinite. What is claimed in the theorem then follows by using the formula for the conditional variance-covariance structure (p. 22) on Σ_i by use of the matrix formulas in section A.2.7.

■

|||| Remark 1.41

In the theorem we have obtained a large number of characteristics for the multiple correlation coefficient and since

$$\rho_{y_i|x}^2 = \frac{V(Y_i) - V(Y_i|\mathbf{X})}{V(Y_i)},$$

we note that we have generalised the property of reduction in variance. It is important to note that we can see from the determinant formula that it is possible to compute the multiple correlation coefficient from the correlation matrix by using the same formulas as when computing it from the variance-covariance matrix.

With regard to the estimation of multiple correlation coefficients the same remark as on p. 33 regarding the estimation of partial coefficients holds.

In the next example we continue example 1.37.

|||| Example 1.42

To get an impression of to which degree the content of C₃A and C₃S in example 1.37 can explain the variation in e.g. 3-day strength we can compute the multiple correlation coefficient between strength day 3 and (C₃S, and C₃A). We find

$$1 - \hat{\rho}_{4|12}^2 = \frac{\det \begin{bmatrix} 1 & 0.158 & 0.120 \\ 0.158 & 1 & -0.309 \\ 0.120 & -0.309 & 1 \end{bmatrix}}{1 \cdot \det \begin{bmatrix} 1 & -0.309 \\ -0.309 & 1 \end{bmatrix}}$$

where the indices of the variables correspond to those used in example 1.33. We find

$$\hat{\rho}_{4|12}^2 = 1 - 0.9435 = 0.0565.$$

The data therefore indicate that only about 6% of the variation in the strength of the cement (from samples which have been collected the way these data have been collected) can be explained by variations in C₃S- and C₃A- content alone.

If the multiple correlation coefficient is 0 (i.e. if $\sigma_i = 0$) it is not difficult to determine the distribution of $\hat{\rho}_{y_i|x}^2$. We give the results in the slightly changed form in

|||| Theorem 1.43

Let $R = \hat{\rho}_{y_i|x}$ be the empirical multiple correlation coefficient between Y_i and $\mathbf{X} = (Z_{m+1}, \dots, Z_p)$ based upon n observations. Then

$$\frac{R^2}{1 - R^2} \cdot \frac{n - (p - m) - 1}{p - m} \sim F(p - m, n - (p - m) - 1),$$

if $\rho_{y_i|x} = \rho_{y_i|z_{m+1}, \dots, z_p} = 0$.

|||| Proof omitted

|||| Remark 1.44

The number $p - m$ is equal to the number of variables in \mathbf{X} , i.e. the number of variables we condition on.

Theorem 1.43 can be used in testing the hypotheses

$$H_0 : \rho_{y_i|x} = 0 \quad \text{against} \quad H_1 : \rho_{y_i|x} \neq 0.$$

We reject the null hypothesis for large values of the test statistic. This is illustrated in

|||| Example 1.45

Consider the situation in example 1.42. We now want to examine if it can be assumed that the multiple correlation between X_4 and (X_1, X_2) is 0. (Note that $p = 3$ and $m = 1$.) We find the statistic

$$\frac{R^2}{1 - R^2} \frac{51 - (3 - 1) - 1}{3 - 1} = \frac{0.0565}{0.9435} \cdot \frac{48}{2} = 1.44.$$

Since

$$F(2, 48)_{0.90} = 2.42,$$

we will at least accept a hypothesis that $\rho_{4|12} = 0$ for any level $\alpha < 10\%$. With the available data it cannot be rejected that $\rho_{4|12} = 0$. This does not mean that it is not different from 0 (which it probably is), only that we cannot be sure using the available data because the true (but unknown) value of $\rho_{4|12}$ is probably rather small.

We shall not consider tests for other values of $\rho_{y_i|x}$.

1.3.3 Regression

We start the section with some remarks on errors of estimators and predictors. If we consider an estimator $\hat{\theta}$ of an unknown parameter θ (a fixed number) the **Mean Squared Error** of $\hat{\theta}$ is

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \text{MSE}(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2] \\ &= V(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2\end{aligned}$$

i.e. the $\text{MSE}(\hat{\theta})$ is equal to the variance of $\hat{\theta}$ plus the squared bias of $\hat{\theta}$ thus relating the MSE to the **precision** (low variance) and the **accuracy** (low bias) of the estimation. If the estimator is unbiased we see that the MSE equals the variance of the estimator.

For a **predictor** $\hat{Y} = g(\mathbf{X})$ of a random variable Y the **Mean Squared (Prediction) Error** is

$$\begin{aligned}\text{MSE}(\hat{Y}) &= \text{MSE}(g(\mathbf{X}), Y) = E[(\hat{Y} - Y)^2] \\ &= E[(g(\mathbf{X}) - Y)^2]\end{aligned}$$

where the mean is taken with respect to the joint distribution of Y and $g(\mathbf{X}) = \hat{Y}$. If Y and \hat{Y} ($= g(\mathbf{X})$) have the same mean, we obtain

$$\text{MSE}(\hat{Y}) = V(\hat{Y} - Y) = V(\hat{Y}) + V(Y) - 2\text{Cov}(\hat{Y}, Y).$$

A reasonable condition for finding a good predictor is of course to minimize the MSE of the predictor. The solution is given in the following theorem.

|||| **Theorem 1.46**

The *Minimum Mean Squared (Prediction) Error* predictor of Y based on X is

$$g(X) = E(Y|X)$$

Since $E(E(Y|X)) = E(Y)$ the *prediction variance* is

$$V(\hat{Y} - Y) = E(V(Y|X))$$

|||| **Proof**

A consequence of basic results on conditional means.

■

In the case of normally distributed random variables we use the term regression for the above conditional mean. More specifically we proceed as follows.

Let $\begin{bmatrix} Y \\ X \end{bmatrix}$ be a stochastic vector. By the term *regression* of Y on X we mean the function given by

$$g(x) = E(Y|X = x),$$

i.e. the conditional mean as a function of the conditioned variable.

Let $\begin{bmatrix} Y \\ X \end{bmatrix}$ be normally distributed with parameters

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma'_1 \\ \sigma_1 & \Sigma_{22} \end{bmatrix}.$$

Then theorem 1.26 shows that

$$g(x) = E(Y|X = x) = \mu_1 + \sigma'_1 \Sigma_{22}^{-1} (x - \mu_2),$$

i.e. the regression is linear (affine). The prediction variance is - since $V(Y|X)$ is independent of X and therefore $E(V(Y|X)) = V(Y|X)$ -

$$\begin{aligned} V(Y - g(x)) &= \sigma_{11} - \sigma'_1 \Sigma_{22}^{-1} \sigma_1 \\ &= \sigma_{11}(1 - \rho_{g|x_1, \dots, x_n}^2) \end{aligned}$$

We now specialise to two dimensions.

Let $\begin{bmatrix} Y \\ X \end{bmatrix}$ be normally distributed with parameters

$$\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix}.$$

Then the regression of Y on X is given by

$$E(Y|X = x) = \mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x),$$

and the regression of X on Y is given by

$$E(X|Y = y) = \mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y).$$

Let us assume that we have measurements $\begin{bmatrix} Y_1 \\ X_1 \end{bmatrix}, \dots, \begin{bmatrix} Y_n \\ X_n \end{bmatrix}$.

The maximum likelihood estimates for the slopes are obtained by using the maximum likelihood estimators for the parameters in the formula. Then

$$\begin{aligned} \hat{\rho} &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}} = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}, \\ \hat{\sigma}_x^2 &= \frac{1}{n} \sum(X_i - \bar{X})^2, \\ \hat{\sigma}_y^2 &= \frac{1}{n} \sum(Y_i - \bar{Y})^2, \end{aligned}$$

and we see e.g. that the estimates of the slope in the expression for the regression of Y on X becomes

$$\hat{\rho} \frac{\hat{\sigma}_y}{\hat{\sigma}_x} = \frac{SP_{xy}}{SS_x}.$$

This gives the empirical regression equation

$$\hat{E}(Y|X = x) = \bar{Y} + \frac{SP_{xy}}{SS_x} (x - \bar{X}),$$

i.e. precisely the same result as we obtain in the one dimensional linear regression analysis. However, there the assumptions are completely different since we assume that the values of the independent variable are deterministic values. In the present context we assume that they are observations of a normally distributed variable which is correlated with the dependent variable. Concerning the estimation it is not important which of the two models one works with but the interpretation of the results are of course dependent hereon. We now continue with example 1.47.

||| Example 1.47

In this example we will determine the linear relations from a measurement by one of the two methods stated in example 1.32 to the other measurement.

We find the regressions

$$\begin{aligned}\hat{E}(X_1|X_2 = x_2) &= \bar{x}_1 + \hat{\rho} \frac{s_1}{s_2} (x_2 - \bar{x}_2) \\ &= 0.65x_2 + 13.43\end{aligned}$$

and

$$\begin{aligned}\hat{E}(X_2|X_1 = x_1) &= \bar{x}_2 + \hat{\rho} \frac{s_2}{s_1} (x_1 - \bar{x}_1) \\ &= 0.58x_1 + 10.14.\end{aligned}$$

These lines are shown in figure 1.47.

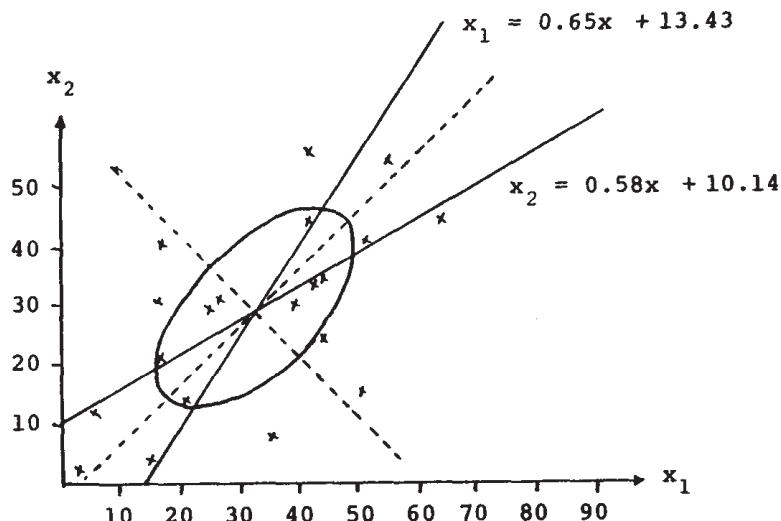


Figure 1.47

If we wish to check if there might be some sort of relation between X_1 and X_2 we can examine the correlation coefficient. It has been found to be

$$\hat{\rho} = \frac{182}{\sqrt{311 \cdot 279}} = 0.617,$$

i.e.

$$\hat{\rho}^2 = 0.380.$$

The test statistic for a test of the hypothesis $\rho = 0$ is, cf. section 1.3.1, with $p = m = 2$

$$t = \frac{0.617}{\sqrt{1 - 0.380}} \sqrt{20 - 2} = 3.32 > t(18)_{0.995}.$$

Using a test at level $\alpha > 1\%$ we must reject the hypothesis and we assume that $\rho \neq 0$, is different from 0. I.e. we now assume there exists a linear relationship between the methods of measurements in the two cases and it is estimated by the two regressions. We can then find estimates of the errors etc. in the usual fashion.

In the figure we have also shown a contour-ellipse and its main axes. It can be shown that the first axis is the line which is obtained by minimizing the orthogonal squared distance to the points. On the other hand the regression equations are found by *minimizing the vertical and horizontal distances respectively*. The first main axis is therefore also called the *orthogonal regression*. In chapter 3 we will return to this concept.

1.4 The partition theorem

In this section we will consider a stochastic variable $X \sim N(\mu, \Sigma)$, where Σ is regular of order n . We will consider the inner product defined by Σ^{-1} and the corresponding norm i.e.

$$(x|y) = x^T \Sigma^{-1} y$$

and

$$\|x\| = \sqrt{(x|x)} = \sqrt{x^T \Sigma^{-1} x}$$

Now let the sub-spaces U_1, \dots, U_k be orthogonal (with respect to this inner product) so that

$$\mathbb{R}^n = U_1 \oplus \dots \oplus U_k.$$

We let $\dim U_i = n_i$ and call the projection onto U_i for p_i . The corresponding projection matrix is called C_i .

Using the notation mentioned above the following is valid

|||| Theorem 1.48

(The partition theorem) If we let

$$\mathbf{Y}_i = p_i(\mathbf{X} - \boldsymbol{\mu}), \quad i = 1, \dots, k$$

and

$$K_i = \|\mathbf{Y}_i\|^2 = \|p_i(\mathbf{X} - \boldsymbol{\mu})\|^2, \quad i = 1, \dots, k,$$

then

$$\mathbf{X} - \boldsymbol{\mu} = \sum_{i=1}^k \mathbf{Y}_i$$

and

$$\|\mathbf{X} - \boldsymbol{\mu}\|^2 = \sum_{i=1}^k K_i.$$

Furthermore $\mathbf{Y}_1, \dots, \mathbf{Y}_k$ are stochastically independent and normally distributed and K_1, \dots, K_k are stochastically independent and $\chi^2(n_i)$ -distributed variables.

|||| Proof

We have that $\mathbf{Y}_i = \mathbf{C}_i(\mathbf{x} - \boldsymbol{\mu})$. Therefore

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_k \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_k \end{bmatrix} (\mathbf{x} - \boldsymbol{\mu}).$$

From this we obtain

$$\mathbf{D}(\mathbf{Y}) = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_k \end{bmatrix} \cdot \boldsymbol{\Sigma} \cdot (\mathbf{C}_1^T, \dots, \mathbf{C}_k^T) = (\mathbf{C}_i \boldsymbol{\Sigma} \mathbf{C}_j^T)_{(i,j)}.$$

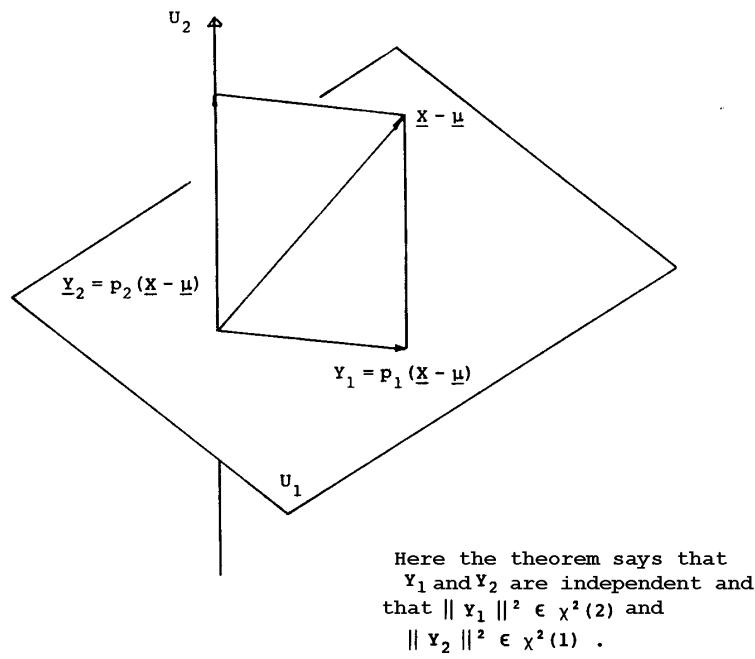


Figure 1.4: Here the theorem says that Y_1 and Y_2 are independent and that $\|Y_2\|^2 \in \chi^2(1)$.

Now for $i \neq j$ it follows from the lemma on page 384 that

$$\mathbf{C}_i \boldsymbol{\Sigma} \mathbf{C}_j^T = \mathbf{0}.$$

From this it follows that the components of \mathbf{Y} are stochastically independent (because \mathbf{Y} is normally distributed).

We must now determine the distribution of $\|p_i(\mathbf{X} - \mu)\|^2$. We have that \mathbf{X} can be written

$$\mathbf{X} = \mu + \mathbf{A}\mathbf{Z}$$

where $\mathbf{Z} \sim N(0, \mathbf{I})$ and $\mathbf{A}\mathbf{A}^T = \boldsymbol{\Sigma}$. From this it follows that

$$\begin{aligned} \|p_i(\mathbf{X} - \mu)\|^2 &= \|p_i(\mathbf{A}\mathbf{Z})\|^2 = \|\mathbf{C}_i \mathbf{A} \mathbf{Z}\|^2 \\ &= \mathbf{Z}^T \mathbf{A}^T \mathbf{C}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{C}_i \mathbf{A} \mathbf{Z} = \mathbf{Z}^T \mathbf{D}_i \mathbf{Z}. \end{aligned}$$

Now

$$\begin{aligned} \mathbf{D}_i \mathbf{D}_i &= \mathbf{A}^T \mathbf{C}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{C}_i \mathbf{A} \mathbf{A}^T \mathbf{C}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{C}_i \mathbf{A} \\ &= \mathbf{A}^T \mathbf{C}_i^T \mathbf{C}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \mathbf{C}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{C}_i \mathbf{A} \\ &= \mathbf{A}^T \mathbf{C}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{C}_i \mathbf{A} \\ &= \mathbf{D}_i, \end{aligned}$$

i.e. \mathbf{D}_i is idempotent. In the above we have used the lemma A.51 repeatedly. It is obvious that $\text{rk } (\mathbf{D}_i) = n_i$. Now, since

$$\begin{aligned} \mathbf{D}_i &= \mathbf{A}^T \mathbf{C}_i^T \mathbf{A}^{T-1} \mathbf{A}^{-1} \mathbf{C}_i \mathbf{A} \\ &= (\mathbf{A}^{-1} \mathbf{C}_i \mathbf{A})^T (\mathbf{A}^{-1} \mathbf{C}_i \mathbf{A}), \end{aligned}$$

then \mathbf{D}_i is positive semidefinite (cf. theorem A.33 p. 368) therefore there exists an orthogonal (and even orthonormal) matrix \mathbf{P}' (theorem A.23) so that

$$\mathbf{P}^T \mathbf{D}_i \mathbf{P} = \Lambda_i \quad \text{or} \quad \mathbf{D}_i = \mathbf{P} \Lambda_i \mathbf{P}^T,$$

where Λ_i is a diagonal matrix with rank n_i . Since \mathbf{D}_i is idempotent we obtain

$$\mathbf{P} \Lambda_i \mathbf{P}^T = \mathbf{P} \Lambda_i \mathbf{P}^T \mathbf{P} \Lambda_i \mathbf{P}^T = \mathbf{P} \Lambda_i^2 \mathbf{P}^T,$$

or $\Lambda_i = \Lambda_i^2$. Therefore Λ_i has n_i 1's and $n - n_i$ 0's on the diagonal. Therefore

$$\begin{aligned} \mathbf{Z}^T \mathbf{D}_i \mathbf{Z} &= \mathbf{Z}^T \mathbf{P} \Lambda_i \mathbf{P}^T \mathbf{Z} = (\mathbf{P}^T \mathbf{Z})^T \Lambda_i (\mathbf{P}^T \mathbf{Z})^T \\ &= \mathbf{V}^T \Lambda_i \mathbf{V} \\ &= \underbrace{V_1^2 + \cdots + V_n^2}_{n_i \text{ components } \neq 0}. \end{aligned}$$

Since $\mathbf{V} \sim N(\mathbf{0}, \mathbf{P}^T \mathbf{P}) = N(\mathbf{0}, \mathbf{I})$ it is seen that

$$\mathbf{Z}^T \mathbf{D}_i \mathbf{Z} = \|p_i(\mathbf{X} - \boldsymbol{\mu})\|^2 \sim \chi^2(n_i).$$

■

||| Example 1.49

Let X_1, \dots, X_n be independent and $N(\boldsymbol{\mu}, \sigma^2)$ -distributed. Then

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}).$$

We consider the subspace U_1 given by

$$\mathbf{x} \in U_1 \Leftrightarrow x_1 = \dots = x_n,$$

and the orthogonal subspace to U_1 (with respect to $\sigma^2 \mathbf{I}$) called U_2 . (This concept of orthogonality corresponds to the usual one). Now the identity

$$\sum (x_i - y)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - y)^2,$$

shows that the projection onto U_1 is given by

$$p_1(\mathbf{x}) = \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix},$$

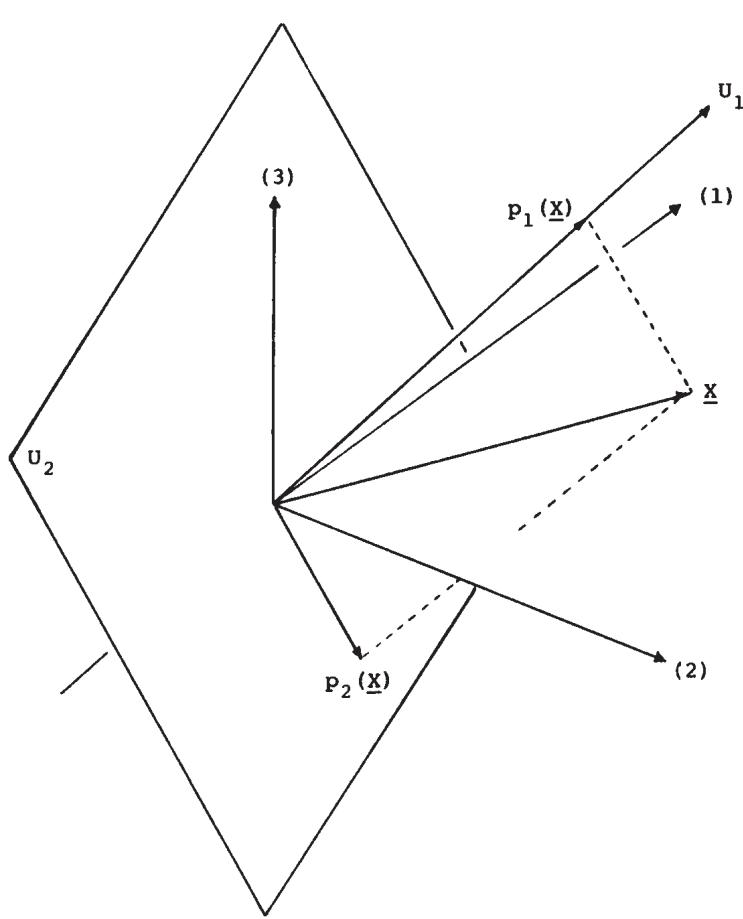


Figure 1.49:

which means

$$p_2(\mathbf{x}) = \mathbf{x} - p_1(\mathbf{x}) = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}.$$

Since $\dim U_1 = 1$ and $\dim U_2 = n - 1$ we find from the partition theorem that

$$p_1(\mathbf{X} - \boldsymbol{\mu}) \quad \text{and} \quad \|p_2(\mathbf{X} - \boldsymbol{\mu})\|^2$$

are stochastically independent. $p_1(\mathbf{X} - \boldsymbol{\mu})$ is normally distributed and $\|p_2(\mathbf{X} - \boldsymbol{\mu})\|^2$ is $\chi^2(n - 1)$ distributed.

Since

$$p_1(\mathbf{X} - \boldsymbol{\mu}) = \begin{bmatrix} \bar{X} - \mu \\ \vdots \\ \bar{X} - \mu \end{bmatrix},$$

and

$$\|p_2(\mathbf{X} - \boldsymbol{\mu})\|^2 = \frac{1}{\sigma^2} \sum_1^n (X_i - \bar{X})^2,$$

we again find the results of the distribution of \bar{X} and $(n - 1)S^2 = \frac{1}{\sigma^2} \sum (X_i - \bar{X})^2$.

1.5 The Wishart distribution and the generalized variance

In the one dimensional case a number of sample-distributions are derived from the normal distribution. The most important of these is the χ^2 -distribution, which corresponds to the sum of squared normally distributed data. Its multi-dimensional analog is the Wishart distribution. We give the definition by means of the density.

||| Definition 1.50

Let \mathbf{V} be a continuously distributed random $p \times p$ -matrix, which is symmetrical and positive semi-definite with probability 1. Then \mathbf{V} is said to be **Wishart distributed** with parameters (n, Σ) , ($n \geq p$), if the density for \mathbf{V} is

$$f(\mathbf{v}) = c \cdot [\det(\mathbf{v})]^{\frac{1}{2}(n-p-1)} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{v} \cdot \Sigma^{-1})\right),$$

for \mathbf{v} positive definite and 0 otherwise. Here Σ is a positive definite $p \times p$ -matrix, and c is the constant given by

$$\frac{1}{c} = 2^{\frac{1}{2}np\pi p(p-1)/4} (\det \Sigma)^{\frac{1}{2}n} \prod_{i=1}^p \Gamma\left(\frac{1}{2}(n+1-i)\right).$$

Abbreviated we write

$$\mathbf{V} \sim W(n, \Sigma) = W_p(n, \Sigma).$$

where the latter version is used whenever there is doubt about the dimension.

We now give a remark about the mean and variance of the components in a Wishart distribution.

|||| Theorem 1.51

Let $\mathbf{V} = (V_{ij})$ be Wishart distributed $W(n, \Sigma)$, where $\Sigma = (\sigma_{ij})$. Then it holds that

$$\begin{aligned}\mathbb{E}(V_{ij}) &= n\sigma_{ij} \\ \text{V}(V_{ij}) &= n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}) \\ \text{Cov}(V_{ij}, V_{kl}) &= n(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}).\end{aligned}$$

|||| Proof omitted

The analogy with the χ^2 -distribution is seen in

|||| Theorem 1.52

Let $\mathbf{X}_i \sim N_p(\mathbf{0}, \Sigma)$, $i = 1, \dots, n$, be independent and regularly distributed. Then for $n \geq p$ it holds that

$$\mathbf{Y} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T \sim W(n, \Sigma).$$

|||| Proof omitted

|||| Remark 1.53

If $n < p$ then \mathbf{Y} as it is defined in the theorem does not have a density function. However, we still choose to say, that \mathbf{Y} is Wishart distributed with parameters (n, Σ) .

Corresponding remarks hold if Σ is singular. Using this convention the theorem holds without the restriction $n \leq p$.

A nearly trivial implication of the above now is

|||| **Theorem 1.54**

Let $\mathbf{V}_1, \dots, \mathbf{V}_k$ be independent random $p \times p$ -matrices, which are $W(n_i, \Sigma)$ -distributed. Then it holds

$$\mathbf{V} = \mathbf{V}_1 + \dots + \mathbf{V}_k \sim W(n_1 + \dots + n_k, \Sigma).$$

One of the main theorems in the theory of sampling functions of normally distributed random variables is that \bar{X} and S^2 are independent and that S^2 is $\sigma^2 \chi^2 / f$ -distributed with 1 degree of freedom less than the number of observations. This theorem has its multidimensional analog in

|||| **Theorem 1.55**

Let $X_i \sim N_p(\mu, \Sigma)$, $i = 1, \dots, n$, be stochastically independent. We let

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i, \\ \mathbf{S} &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T.\end{aligned}$$

Then

$$\bar{X} \sim N_p(\mu, \frac{1}{n} \Sigma)$$

and

$$\mathbf{S} \sim W(n-1, \frac{1}{n-1} \Sigma).$$

Furthermore, \bar{X} and \mathbf{S} are stochastically independent.

|||| **Proof omitted**

We will now consider some results on marginal distributions. We have that

|||| Theorem 1.56

Let \mathbf{V} be Wishart distributed with parameters (n, Σ) . We consider the partitioning

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

It then holds that

$$\mathbf{V}_{ii} \sim W(n, \Sigma_{ii}).$$

Further, it holds that

|||| Theorem 1.57

We again consider the above situation. If Σ_{12} and Σ_{21} are 0-matrices, then \mathbf{V}_{11} and \mathbf{V}_{22} are stochastically independent.

|||| Proof

for the theorems. They follow readily by considering the corresponding partitions of normally distributed vectors, which produce the Wishart distributions.

■

Since the multidimensional normal distribution can be defined independent of the coordinate system, then it is not surprising that something similar holds for the Wishart distribution. Because change form coordinates in one coordinate system to coordinates in another is performed by manipulating matrices we have the following

|||| Theorem 1.58

Let $\mathbf{V} \sim W_p(n, \Sigma)$ and let \mathbf{A} be an arbitrary fixed $r \times p$ -matrix. Then

$$\mathbf{A} \mathbf{V} \mathbf{A}^T \sim W_r(n, \mathbf{A} \Sigma \mathbf{A}^T).$$

||| Proof

As indicated above one just has to consider the normally distributed vectors which result in V and then transform them. The resultat then follows readily.

■

We now conclude the chapter by introducing a different generalisation from the one-dimensional variance to the multidimensional case than the variance-covariance matrix.

||| Definition 1.59

Let the p -dimensional vector X have the variance-covariance matrix Σ . By the term *the generalized variance* of X we mean the determinant of the variance-covariance matrix, i.e.

$$\text{gen.var.}(X) = \det(\Sigma).$$

||| Remark 1.60

In section A.2.6 we established that the determinant of a matrix corresponds to the volume relationship of the corresponding linear projection, i.e. it is a intuitively sensible measure of the "size" of a matrix.

If we have observations X_1, \dots, X_n , then we define the *empirical generalised variance* in a straight forward way from the empirical variance-covariance matrix

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T,$$

as $\det(\mathbf{S})$

In the normal case we can establish the distribution of the empirical generalised variance. We have

|||| **Theorem 1.61**

Let $X_i \sim N_p(\mu, \Sigma)$, $i = 1, \dots, n$, be stochastically independent. Then the empirical generalised variance follows the same distribution as

$$\frac{\det \Sigma}{(n-1)p} \cdot Z_1 \dots Z_p,$$

where Z_1, \dots, Z_p are stochastically independent and $Z_i \sim \chi^2(n-i)$.

|||| **Proof omitted**

For $p = 1$ and 2 it is possible to find the density of the empirical generalised variance. However, for larger values of p this density involves integrals, which cannot readily be written as known functions, but for $n \rightarrow \infty$ we do have

|||| **Theorem 1.62**

Let \mathbf{S} be as above (in the normal case). Then it holds that

$$\sqrt{n-1} \left(\frac{\det(\mathbf{S})}{\det(\Sigma)} - 1 \right) \quad \text{asymptotically} \quad \sim N(0, 2p).$$

1.6 The complex normal distribution and the complex Wishart distribution

Work in progress...

1.7 On estimation of multidimensional parameters

In this section we present the most important results on estimation of multi-dimensional parameters needed in the sequel. Basic knowledge on estimation of one dimensional parameters is assumed, including concepts as consistency, efficiency and sufficiency.

1.7.1 Maximum likelihood estimation

We consider a (multivariate) random variable X with a frequency function $p(x; \theta)$, where θ is an unknown k -dimensional parameter. Based on the outcome of X we want to estimate θ . In analogy with the one-dimensional case, we introduce the basic concepts in

||| Definition 1.63

The *likelihood function* for θ is

$$L(\theta) = p(x; \theta),$$

and the *maximum likelihood estimator*, short *ML estimator*, $\hat{\theta}$ is the parameter value that maximizes this expression, i.e.

$$L(\hat{\theta}) = \max_{\theta} L(\theta)$$

or

$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta)$$

A mathematically more stringent definition would use the supremum instead of the maximum, i.e.

$$L(\hat{\theta}) = \sup_{\theta} L(\theta).$$

In our setting the distinction is mainly of a technical nature, which we shall not pursue.

The *log likelihood function* is

$$l(\theta) = \log L(\theta)$$

Often X will consist of identically distributed, independent observations X_1, \dots, X_n , each with frequency function $f(x; \theta)$. Then

$$p(x; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

and

$$l(\boldsymbol{\theta}) = l_n(\boldsymbol{\theta}) = \log \prod_{i=1}^n f(x_i; \boldsymbol{\theta}) = \sum_{i=1}^n \log f(x_i; \boldsymbol{\theta})$$

In this case we may use the notation $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_n$ in order to emphasize the number of observations that are entering the maximum likelihood estimator.

If the support of the frequency function $\{x \mid f(x; \boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$, we may determine the maximum likelihood estimators by solving the *likelihood equations*

$$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} l(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_k} l(\boldsymbol{\theta}) \end{bmatrix} = \mathbf{0}$$

or

$$\frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}) = 0, \quad i = 1, \dots, k$$

The gradient vector $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$ is called the *score vector*.

||| Remark 1.64

Under mild regularity conditions, the expected value of the score vector is equal to $\mathbf{0}$. To see this, we consider

$$\begin{aligned} E_{(X|\boldsymbol{\theta})} \left\{ \frac{\partial}{\partial \theta_i} l(\boldsymbol{\theta}) \right\} &= \int_X \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i} p(x; \boldsymbol{\theta}) dx \\ &= \int_X \frac{1}{p(x; \boldsymbol{\theta})} \frac{\partial p(x; \boldsymbol{\theta})}{\partial \theta_i} p(x; \boldsymbol{\theta}) dx \\ &= \frac{\partial}{\partial \theta_i} \int_X p(x; \boldsymbol{\theta}) dx \\ &= 0 \end{aligned}$$

since the integral is equal to 1. The mild assumptions needed shall allow interchanging differentiation and integration.

||| Example 1.65

Consider the independent $N(\mu, 1)$ distributed random variables X_1, \dots, X_n . Then

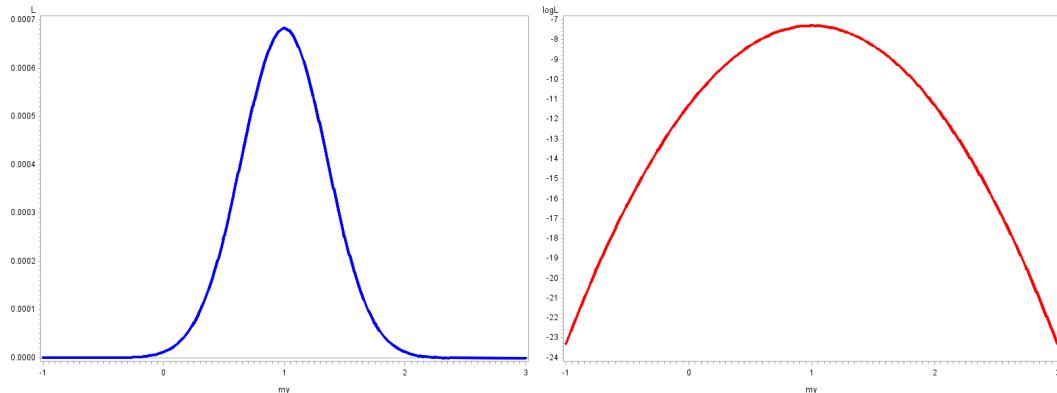
$$l_n(\mu) = \log(2\pi)^{-n/2} - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2$$

and the likelihood equation becomes

$$\frac{\partial}{\partial \mu} l_n(\mu) = \sum_{i=1}^n (x_i - \mu) = 0$$

Thus the maximum likelihood estimator of μ is the average

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$



The likelihood function and the log likelihood function for normally distributed random variables with $\hat{\mu} = 1$.

||| Example 1.66

Consider the independent $U(0, \theta)$ distributed random variables X_1, \dots, X_n . Here the support of the distribution $f(x_i; \theta)$ is the interval $[0, \theta]$, i.e. depending on the unknown parameter θ . Hence we shall not consider the log likelihood but approach the maximization directly. Introducing

$$I_{[0,\theta]}(x) = \begin{cases} 1, & \text{for } 0 \leq x \leq \theta \\ 0, & \text{elsewhere} \end{cases}$$

we get

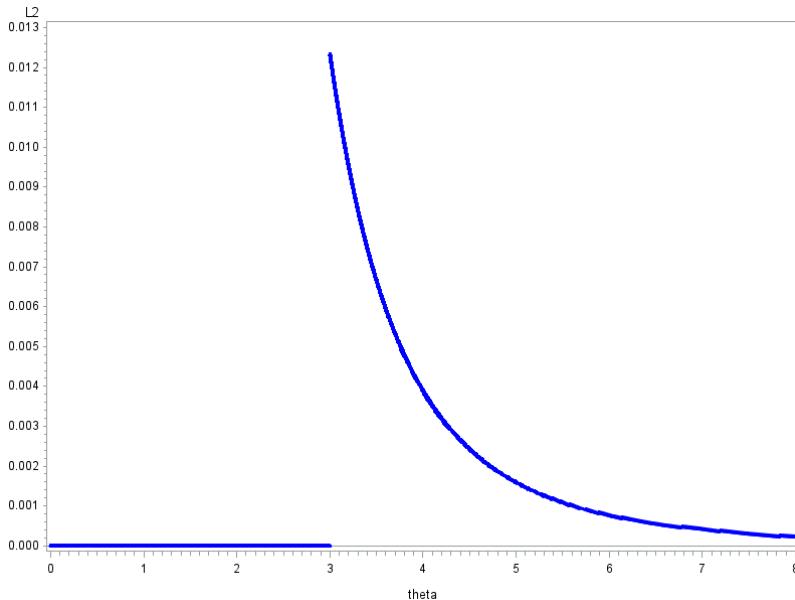
$$L(\theta) = \prod_{i=1}^n \frac{1}{\theta} I_{[0,\theta]}(x_i) = \theta^{-n} \prod_{i=1}^n I_{[0,\theta]}(x_i)$$

It follows that $L(\theta)$ is different from 0 (and the equal to θ^{-n}) exactly when all x'_i 's are in the interval $[0, \theta]$. Since the x'_i 's are positive, this is the case if and only if the maximum value $x_{(n)} \in [0, \theta]$. Therefore

$$L(\theta) = \theta^{-n} I_{[0,\theta]}(x_{(n)}) = \theta^{-n} I_{[x_{(n)}, \infty]}(\theta)$$

The last equality sign is simply expressing that $0 \leq x_{(n)} \leq \theta$ if $\theta \in [x_{(n)}, \infty]$. This function has its maximum in $x_{(n)}$ and thus the maximum likelihood estimator for θ is

$$\hat{\theta} = X_{(n)} = \max_{i=1, \dots, n} X_i$$



The likelihood function for uniformly distributed random variables with $\hat{\theta} = 3$.

|||| Remark 1.67

The notation $\hat{\theta}$ may cause some ambiguity between the ML estimator, which is a random variable, and the ML estimate, which is a (vector of) number(s). To be more specific, we may consider i.i.d. random variables X_1, \dots, X_4 , each $\sim N(\mu, 1)$. As shown above, the ML estimator for μ is $\hat{\mu} = \bar{X} = (1/4) \sum_{i=1}^4 X_i$, which again is a random variable and by the way normally distributed $\sim N(\mu, 1/n)$. If the specific outcomes from observing X_1, \dots, X_4 are $x_1 = 2, x_2 = 1.5, x_3 = 2, x_4 = 2.5$, then the *ML estimate* of μ is

$$\hat{\mu} = \bar{x} = \frac{1}{4} \sum_{i=1}^4 x_i = \frac{1}{4} (2 + 1.5 + 2 + 2.5) = 2$$

This number \bar{x} is thus the observed value of the random variable \bar{X} , but we use the expression $\hat{\mu}$ for both. Normally we distinguish between a random variable and its observed value by using an upper case letter like e.g. X for the random variable and the corresponding lower case letter x for the observed value.

|||| Definition 1.68

The *(Fisher) information matrix* is the dispersion (variance-covariance) matrix of the score vector, i.e.

$$I(\boldsymbol{\theta}) = E_{(X|\boldsymbol{\theta})} \left\{ (\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta})) (\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}))^T \right\} = D_{(X|\boldsymbol{\theta})} \{ \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \}$$

i.e. the matrix with the $(i, j)^{\text{th}}$ element

$$i_{ij}(\boldsymbol{\theta}) = E_{(X|\boldsymbol{\theta})} \left\{ \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} \right\} = Cov_{(X|\boldsymbol{\theta})} \left\{ \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i}, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} \right\}$$

||| Remark 1.69

If we are in the case with i.i.d. random variables we will sometimes use the notation

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{I}_n(\boldsymbol{\theta}) = n\mathbf{I}_1(\boldsymbol{\theta})$$

where the $(i, j)^{\text{th}}$ element in $\mathbf{I}_1(\boldsymbol{\theta})$ is

$$i_{1,ij}(\boldsymbol{\theta}) = E_{(X|\boldsymbol{\theta})} \left\{ \frac{\partial \log f(X_1; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log f(X_1; \boldsymbol{\theta})}{\partial \theta_j} \right\}$$

||| Theorem 1.70

Under the regularity conditions given in the following two theorems, this will be equal to

$$i_{ij}(\boldsymbol{\theta}) = -E_{(X|\boldsymbol{\theta})} \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\boldsymbol{\theta}) \right\}$$

i.e.

$$\mathbf{I}(\boldsymbol{\theta}) = -E_{(X|\boldsymbol{\theta})} \{ \nabla_{\boldsymbol{\theta}}^2 l(\boldsymbol{\theta}) \} = -E_{(X|\boldsymbol{\theta})} \{ \mathbf{H}(\boldsymbol{\theta}) \}$$

i.e. the Fisher information is equal to minus the expected value of the **Hessian matrix** of the log likelihood function. The negative value of the Hessian matrix is called the **observed information matrix**.

||| Proof

We immediately get

$$\begin{aligned} \frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\boldsymbol{\theta}) &= \frac{\partial}{\partial \theta_i} \left\{ \frac{\partial}{\partial \theta_j} \log p(\mathbf{x}; \boldsymbol{\theta}) \right\} \\ &= \frac{\partial}{\partial \theta_i} \left\{ \frac{1}{p(\mathbf{x}; \boldsymbol{\theta})} \frac{\partial p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} \right\} \\ &= \frac{1}{(p(\mathbf{x}; \boldsymbol{\theta}))^2} \left\{ \frac{\partial^2 p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} p(\mathbf{x}; \boldsymbol{\theta}) - \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \log p(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} \right\} \end{aligned}$$

Now

$$\begin{aligned} E_{(X|\theta)} \left\{ \frac{\mathbf{1}}{p(x; \theta)} \frac{\partial^2 p(x; \theta)}{\partial \theta_i \partial \theta_j} \right\} &= \int_X \frac{\partial^2 p(x; \theta)}{\partial \theta_i \partial \theta_j} dx \\ &= \frac{\partial^2}{\partial \theta_i \partial \theta_j} \int_X p(x; \theta) dx \\ &= 0 \end{aligned}$$

which is equal to 0 since the integral equals 1. From this, the result follows immediately. ■

Often one must solve the likelihood equations iteratively. **Newton-Raphson's algorithm** is a commonly used iterative optimization algorithm. To be more specific, if z is k -dimensional and we have k equations

$$\mathbf{g}(z) = \begin{bmatrix} g_1(z) \\ \vdots \\ g_k(z) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

we iteratively compute

$$z_{n+1} = z_n - J^{-1}(z_n) \mathbf{g}(z_n)$$

where $J(z)$ is the **Jacobian**

$$J(z) = \frac{\partial(g_1, \dots, g_k)}{\partial(z_1, \dots, z_k)} = \begin{bmatrix} \frac{\partial g_1}{\partial z_1} & \dots & \frac{\partial g_1}{\partial z_k} \\ \vdots & & \vdots \\ \frac{\partial g_k}{\partial z_1} & \dots & \frac{\partial g_k}{\partial z_k} \end{bmatrix}$$

Now, the Jacobian for the score vector is the Hessian matrix

$$\nabla^2 l(\theta) = H(\theta) = \begin{bmatrix} \frac{\partial^2 l}{\partial \theta_1^2} & \dots & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_k} \\ \vdots & & \vdots \\ \frac{\partial^2 l}{\partial \theta_k \partial \theta_1} & \dots & \frac{\partial^2 l}{\partial \theta_k^2} \end{bmatrix}$$

and the Newton Raphson iteration becomes

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \mathbf{H}^{-1}(\hat{\theta}_n) \nabla l(\hat{\theta}_n)$$

Sometimes the Hessian matrices will fluctuate undesirably, especially if the starting value $\hat{\theta}_0$ is far from the optimum. R. A. Fisher instead introduced the method of scoring by replacing the Hessian matrix with its expected value, giving *the scoring equations*

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \mathbf{I}^{-1}(\hat{\theta}_n) \nabla l(\hat{\theta}_n)$$

Since the information matrix is positive definite if we have not overparametrized, it will often ‘behave’ better than the Hessian. Other modifications have been used in order to improve the convergence properties, but that is beyond the scope of the present representation.

||| **Theorem 1.71**

Cramér-Rao's inequality. Let X_1, \dots, X_n be independent, identically distributed random variables with frequency function $f(x; \theta)$, where the unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. We assume that the support set of f does not depend on θ . Furthermore, assume that for all (i, j) and for all $\theta \in \text{int}(\Theta)$ we have

1. $\frac{\partial f(x; \theta)}{\partial \theta_i}$ exists for all x
2. $E_{(X|\theta)} \left\{ \frac{\partial \log f(X_1; \theta)}{\partial \theta_i} \right\} = 0$
3. $E_{(X|\theta)} \left\{ \left(\frac{\partial \log f(X_1; \theta)}{\partial \theta_i} \right)^2 \right\} < \infty$
4. $\det I(\theta) \neq 0$

Assume furthermore that $\check{\theta} = \check{\theta}(X_1, \dots, X_n) = (\check{\theta}_1, \dots, \check{\theta}_k)^T$ is an unbiased estimator for θ satisfying

5. $E_{(X|\theta)} \left\{ \check{\theta}_i \frac{\partial}{\partial \theta_j} \log \prod_{i=1}^n f(X_i; \theta) \right\} = E_{(X|\theta)} \left\{ \check{\theta}_i \frac{\partial l_n(\theta)}{\partial \theta_j} \right\} = \delta_{ij}$
6. $E_{(X|\theta)} \left\{ \check{\theta}_i^2 \right\} < \infty$

Then the dispersion matrix of $\check{\theta}$ satisfies

$$D(\check{\theta}(X_1, \dots, X_n)) \geq I^{-1}(\theta) = \frac{1}{n} I_1^{-1}(\theta)$$

i.e. the matrix

$$D(\check{\theta}(X_1, \dots, X_n)) - I^{-1}(\theta)$$

is positive semi-definite.

||| **Proof omitted**

See e.g. Witting and Nölle (1970)

|||| **Remark 1.72**

In ‘regular’ cases the assumptions are not too restrictive. As earlier, it is important that integration wrt x and differentiation wrt θ may be interchanged. If this is the case, then 5. becomes

$$\begin{aligned} E_{(X|\theta)} \left\{ \check{\theta}_i \frac{\partial l_n(\theta)}{\partial \theta_j} \right\} &= \int_X \check{\theta}_i \frac{1}{p(x;\theta)} \frac{\partial p(x;\theta)}{\partial \theta_j} p(x;\theta) dx \\ &= \frac{\partial}{\partial \theta_j} \int_X \check{\theta}_i p(x;\theta) dx \\ &= \frac{\partial \theta_i}{\partial \theta_j} \\ &= \delta_{ij}. \end{aligned}$$

where δ_{ij} is Kronecker’s δ ($= 0$ for $i \neq j$ and $= 1$ for $i = j$).

|||| **Theorem 1.73**

Let X_1, \dots, X_n be independent, identically distributed random variables with frequency function $f(x; \theta)$, where the unknown parameter $\theta \in \Theta \subseteq \mathbb{R}^k$. We assume that the support of f does not depend on θ . Furthermore, assume that for all (i, j) and for all $\theta \in \text{int}(\Theta)$ we have

1. $\theta_1 \neq \theta_2 \Rightarrow f(\bullet, \theta_1) \neq f(\bullet, \theta_2)$
2. $\frac{\partial^2 f(x, \theta)}{\partial \theta_i \partial \theta_j}$ exists and is continuous.
3. $E_{(X|\theta)} \left\{ \frac{1}{f(X_1; \theta)} \frac{\partial f(X_1; \theta)}{\partial \theta_i} \right\} = 0$
4. $E_{(X|\theta)} \left\{ \frac{1}{f(X_1; \theta)} \frac{\partial^2 f(X_1; \theta)}{\partial \theta_i \partial \theta_j} \right\} = 0$

and let there exist a neighborhood \mathcal{U} around θ and a function $M(x, \theta)$ for which $E_{(X|\theta)} \{ M(X_1, \theta) \} < \infty$ so that

5. $\left| \frac{\partial^2 \log f(x; \theta)}{\partial \theta_i \partial \theta_j} \right|_{\theta=\theta^*} \leq M(x, \theta) \quad \forall \theta^* \in \mathcal{U}$
6. $\det I(\theta) \neq 0$

If the maximum likelihood estimator $\hat{\theta}_n$ is consistent, i.e. $\hat{\theta}_n \xrightarrow{P} \theta$, then

$$\sqrt{n} (\hat{\theta}_n - \theta) \xrightarrow{d} N_k(0, I^{-1}(\theta))$$

|||| **Proof omitted**

See e.g. Witting and Nölle (1970)

|||| **Remark 1.74**

The requirement on the consistency of the maximum likelihood estimator is less restrictive than it may appear. Assumption 1 is sufficient for ensuring that with a probability approaching 1 as $n \rightarrow \infty$, there exists consistent solutions of the likelihood equations, and if there is a unique root of the likelihood equation, then this root is consistent, cf. Hunter (2014), who use a version of the assumptions that involve the third order derivatives

$$\frac{\partial^3 l(\theta)}{\partial \theta_i \partial \theta_j \partial \theta_k}$$

We shall not go into any further detail on these aspects. The theorem does also maintain its validity with simple modifications of the asymptotic parameters in the case with independent, but not necessarily identically distributed random variables X_1, \dots, X_n .

1.7.2 Restricted Maximum Likelihood (REML)

We consider a multivariate normally distributed random variable

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \mathbf{z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where \mathbf{x} and \mathbf{z} are known (design) matrices, $\boldsymbol{\theta}$ a vector of fixed (constant, unknown) parameters, and $\boldsymbol{\gamma}$ a vector of unknown random variables called random effects parameters. $\boldsymbol{\varepsilon}$ is the error term. We assume that $\boldsymbol{\gamma}$ and $\boldsymbol{\varepsilon}$ are normally distributed with

$$E \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad D \begin{bmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{E} \end{bmatrix}$$

This model is called a *mixed model*, and we shall see some special cases in the forthcoming chapters. A very thorough treatment also of the matrix algebra involved is provided by Searle, Casella, and McCulloch (1992). The basic highlights are given in Duchateau, Janssen, and Rowlands (1998).

An immediate consequence of the above is

$$\mathbf{Y} \sim N_n (\mathbf{x}\boldsymbol{\theta}, \boldsymbol{\Sigma}) = N_n (\mathbf{x}\boldsymbol{\theta}, \mathbf{z}\boldsymbol{\Gamma}\mathbf{z}^T + \mathbf{E})$$

We assume that $\mathbf{x} = [x_1 \ \cdots \ x_k]$ has rank r . Let the full rank $n \times (n - r)$ matrix $\mathbf{K} = [k_1 \ \cdots \ k_{n-r}]$ be such that

$$\mathbf{K}^T \mathbf{x} = \begin{bmatrix} k_1^T x_1 & \cdots & k_1^T x_k \\ \vdots & & \vdots \\ k_{n-r}^T x_1 & \cdots & k_{n-r}^T x_k \end{bmatrix} = \mathbf{0}.$$

Then

$$\mathbf{K}^T \mathbf{Y} \sim N_{n-r} (\mathbf{0}, \mathbf{K}^T \Sigma \mathbf{K})$$

Thus, the distribution of the $n - r$ dimensional vector $\mathbf{K}^T \mathbf{Y}$ does not depend on the vector $\boldsymbol{\theta}$ of fixed parameters, only the parameters describing the random parts, the so-called *variance components*. By analyzing this distribution using ordinary maximum likelihood methods, we obtain the *restricted maximum likelihood estimators (REML estimators) for the variance component parameters*.

Initially we state an important property of \mathbf{x} and \mathbf{K} . The projection matrix for projecting orthogonally on the column space of \mathbf{x} is

$$\mathbf{H} = \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T.$$

Since all the columns in \mathbf{K} are orthogonal to the columns in \mathbf{x} , we have the projection matrix

$$\mathbf{M} = \mathbf{K}(\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T$$

||| Example 1.75

To elucidate the above, we consider n independent $N(\mu, \sigma^2)$ distributed random variables Y_i . Organized as a multivariate observation we get

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \mathbf{1}\mu + \boldsymbol{\varepsilon}.$$

i.e.

$$\mathbf{x} = \mathbf{1}$$

We see that the $n \times (n - 1)$ matrix

$$\mathbf{K} = \begin{bmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & & -\frac{1}{n} \\ \vdots & & & \\ -\frac{1}{n} & -\frac{1}{n} & 1 - \frac{1}{n} & \\ -\frac{1}{n} & -\frac{1}{n} & & -\frac{1}{n} \end{bmatrix}$$

satisfies the condition that $\mathbf{K}^T \mathbf{x} = 0$. The distribution of the $(n - 1)$ dimensional random variable $\mathbf{K}^T \mathbf{Y}$ is

$$\mathbf{K}^T \mathbf{Y} \sim N_{n-1}(\mathbf{0}, \mathbf{K}^T \Sigma \mathbf{K}) = N_{n-1}(\mathbf{0}, \sigma^2 \mathbf{K}^T \mathbf{K})$$

The likelihood function of σ^2 (depending on the random variables $\mathbf{K}^T \mathbf{Y}$) becomes

$$L(\sigma^2 \mid \mathbf{K}^T \mathbf{Y}) = \frac{1}{(2\pi\sigma^2)^{(n-1)/2} \sqrt{|\mathbf{K}^T \mathbf{K}|}} \exp\left\{-\frac{1}{2\sigma^2} \mathbf{Y}^T \mathbf{K} (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{Y}\right\}$$

Taking the log we obtain

$$l(\sigma^2 \mid \mathbf{K}^T \mathbf{Y}) = c - \frac{n-1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \mathbf{Y}^T \mathbf{K} (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{Y}$$

where c does not depend on σ^2 . The maximum of this is obtained for

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-1} \mathbf{Y}^T \mathbf{K} (\mathbf{K}^T \mathbf{K})^{-1} \mathbf{K}^T \mathbf{Y} \\ &= \frac{1}{n-1} \mathbf{Y}^T (\mathbf{I} - \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T) \mathbf{Y} \\ &= \frac{1}{n-1} (\mathbf{Y}^T \mathbf{Y} - \frac{1}{n} (\mathbf{Y}^T \mathbf{x})^2) \\ &= \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \end{aligned}$$

i.e. the usual unbiased estimator of σ^2 is the restricted maximum likelihood estimator of σ^2 .

1.7.3 Profile, partial, marginal, conditional, and quasi likelihood

In this section we briefly mention other modifications of maximum likelihood estimation.

We consider a random variable Z with a frequency function f_Z depending on a vector parameter ϕ

$$Z = \begin{bmatrix} X \\ Y \end{bmatrix}, \quad \phi = \begin{bmatrix} \theta \\ \eta \end{bmatrix}$$

with frequency function

$$f_Z(z; \phi) = f_Z(x, y; \theta, \eta).$$

The (unrestricted) likelihood function is thus

$$L(\theta, \eta) = f_Z(x, y; \theta, \eta).$$

We assume that η is a nuisance parameter, possibly of high dimension and of no interest in the analysis.

The *profile likelihood function* for estimating θ is defined by replacing *boldsymbol* η with an estimate of $\hat{\eta}$, i.e.

$$L_{prf}(\theta) = L(\theta, \hat{\eta}(\theta))$$

where

$$\hat{\eta}(\theta) = \operatorname{argmax}_{\eta} L(\theta, \eta)$$

The *profile maximum likelihood* estimator is thus

$$\hat{\theta} = \operatorname{argmax}_{\theta} L_{prf}(\theta) = \operatorname{argmax}_{\theta} L(\theta, \hat{\eta}(\theta))$$

We factorize the frequency function

$$f_Z(z | \phi) = f_Z(x, y; \theta, \eta) = f_X(x | \phi) f_{Y|X}(y | x; \phi)$$

into the marginal distribution of X and the conditional distribution of Y given X . If the marginal distribution only depends on θ , i.e.

$$f_X(x | \phi) = f_X(x | \theta)$$

it may be advantageous to base the analysis of θ on the marginal distribution alone leading to a *marginal maximum likelihood estimator* of θ . If the conditional distribution depends on θ alone, i.e.

$$f_{Y|X}(y | x; \phi) = f_{Y|X}(y | x; \theta)$$

this leads to a *conditional maximum likelihood estimator* of θ . In a slightly different setting we may consider a (partially) sufficient statistic S for the nuisance parameters η , i.e. the frequency function splits into a product like

$$f_Z(z | \phi) = f_Z(z; \theta, \eta) = f_{Z|S}(z | S(z); \theta) f_S(S(z); \eta)$$

It immediately follows that conditioning the distribution of Z on S results in a frequency function which does not depend on the nuisance parameters.

$$f_{Z|S}(z | s; \phi) = f_{Z|S}(z | s; \theta)$$

also leading to a *conditional maximum likelihood estimator* of θ .

Sometimes we will deliberately misspecify the likelihood function. As an example we may consider a linear regression model, where the observations are not independent having the same variance. However, we may decide to analyze the simpler model where we assume that the errors are independent and normally distributed with the same variance. This will lead to what is called the *quasi maximum likelihood estimators*, in this case coinciding with the normal least squares estimate of the regression parameter. The concept is used in the so-called linear exponential family, and we shall refer to the literature for more specific results.

In a Bayesian set-up, we consider prior distributions for the parameters and integrate those out leading to different concepts of marginal or integrated likelihood. We shall not go into any detail with this but refer to the rich literature.

|||| Chapter 2

The general linear model

In this chapter we will formulate a model which is a natural generalisation of the variance and regression analysis models known from introductory statistics. The theorems and definitions will to a large extent be interpreted geometrically in order to give a more intuitive understanding of problems.

2.1 Estimation in the general linear model

We first give a description of the model in

2.1.1 Formulation of the Model.

We consider an n -dimensional stochastic variable $\mathbf{Y} \in N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is assumed known. Consider the norm given by $\boldsymbol{\Sigma}^{-1}$ i.e.

$$\|\mathbf{x}\|^2 = \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x}.$$

The norm $(\sigma^2 \boldsymbol{\Sigma})^{-1}$ defined by the inverse variance-covariance matrix is given by

$$\|\mathbf{x}\|_{\sigma^2}^2 = \frac{1}{\sigma^2} \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} = \frac{1}{\sigma^2} \|\mathbf{x}\|^2.$$

The two norms are seen to be proportional and they result in the same concept of orthogonality. We will now consider a number of problems in connection with the estimation and testing of the mean value $\boldsymbol{\mu}$ in cases where $\boldsymbol{\mu}$ is a known linear function of unknown parameters i.e.

$$\boldsymbol{\mu} = \mathbf{x} \boldsymbol{\theta}$$

or

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix},$$

where \mathbf{x} is assumed known.

Geometrically this can be expressed such that we assume the expected value of the stochastic vector \mathbf{Y} is contained in a subspace M of R^n . M is the image of R^k corresponding to the linear projection \mathbf{x} . The dimension of M is $\text{rg}(\mathbf{x}) \leq k$. The situation is depicted in the following figure.

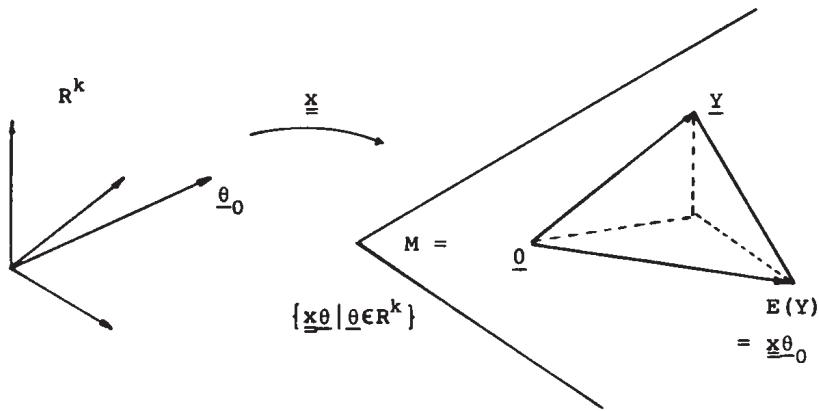


Figure 2.1: Geometrical sketch of the general linear model.

We will call such a model, where the unknown mean value μ is a (known) linear function of the parameter θ a (general) linear model. This is also valid without the assumption \mathbf{Y} has to be normally distributed.

||| Example 2.1

Consider an ordinary one-dimensional regression analysis model i.e. we have observations

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $E(\varepsilon_i) = 0$. This model can be written

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

or

$$\mathbf{Y} = \mathbf{x} \boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

i.e. the model is linear in the meaning stated above.

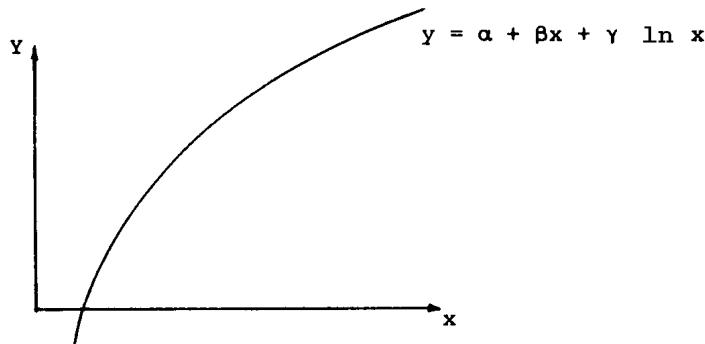
Another example is

||| Example 2.2

We now consider a situation, where

$$Y_i = \alpha + \beta x_i + \gamma \ln x_i + \varepsilon_i, \quad i = 1, \dots, n$$

and still we have $E(\varepsilon_i) = 0$.



Even in this case we have a linear model which is

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \ln x_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & \ln x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

We note that the term linear has nothing to do with $E(Y|X) = \alpha + \beta x + \gamma \ln x$ being linear in the independent variable x , rather that $E(Y|x)$ considered as a function of the unknown parameter $(\alpha, \beta, \gamma)'$ should be linear. If we had had a model such as

$$Y_i = \alpha + \beta \ln(\gamma x_i + \delta) + \varepsilon_i,$$

where α, β, γ and δ are the unknown parameters it would not be possible to write

$$Y = X \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \varepsilon$$

with the known x -matrix and we would therefore not have a linear model.

2.1.2 Estimation in the regular case

We will first formulate the result of estimating θ in

|||| Theorem 2.3

Let x and θ be given as in the preceding section and let $Y \in N_n(x\theta, \sigma^2\Sigma)$, where Σ is positive definite. Then the maximum likelihood estimator $\hat{\theta}$ for θ is given by $x\hat{\theta}$ being the projection (with respect to Σ) onto M , $\hat{\theta}$ is a solution to the so-called *normal equation(s)*

$$(x'\Sigma^{-1}x)\hat{\theta} = x'\Sigma^{-1}y.$$

If x has full rank k , then

$$\hat{\theta} = (x'\Sigma^{-1}x)^{-1}x'\Sigma^{-1}y,$$

and being a linear combination of normally distributed variables $\hat{\theta}$ is also normally distributed with parameters

$$\begin{aligned} E(\hat{\theta}) &= \theta \\ D(\hat{\theta}) &= \sigma^2(x'\Sigma^{-1}x)^{-1}. \end{aligned}$$

It is especially noted that $\hat{\theta}$ is an unbiased estimate of θ .

|||| Proof

If $Y \in N(x\theta, \sigma^2\Sigma)$, where Σ is regular then the density for Y

$$\begin{aligned} f(y) &= \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sigma^n} \frac{1}{\sqrt{\det\Sigma}} \exp\left[-\frac{1}{2\sigma^2}(y - x\theta)' \Sigma^{-1} (y - x\theta)\right] \\ &= k \cdot \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2} \|y - x\theta\|^2\right]. \end{aligned}$$

We have the likelihood function

$$L(\theta) = k \cdot \frac{1}{\sigma^n} \exp\left[-\frac{1}{2\sigma^2} \|y - x\theta\|^2\right],$$

taking the logarithm on each side gives

$$\ln L(\theta) = k_1 - \frac{1}{2\sigma^2} \|y - x\theta\|^2.$$

It is now evident that maximisation of the likelihood function is equivalent to minimisation of the squared distance between any point in M and the observation i.e.

equivalent to minimisation of

$$\|\mathbf{y} - \mathbf{x}\boldsymbol{\theta}\|^2 = (\mathbf{y} - \mathbf{x}\boldsymbol{\theta})'\Sigma^{-1}(\mathbf{y} - \mathbf{x}\boldsymbol{\theta}).$$

From the result p. 382 the value of $\mathbf{x}\boldsymbol{\theta}$, giving the minimum is equal to the orthogonal projection (with respect to Σ^{-1}) of \mathbf{y} on M . From example A.48 p. 379 the optimal $\boldsymbol{\theta}$ is the solution to the equation

$$(\mathbf{x}'\Sigma^{-1}\mathbf{x})\boldsymbol{\theta} = \mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

If $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ has full rank k , i.e. if \mathbf{x} has rank k (cf. p. 365) we therefore have

$$\boldsymbol{\theta}_{\text{opt.}} = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

We have now shown the first half of the theorem.

From theorem 1.2 we find that

$$E(\hat{\boldsymbol{\theta}}) = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{x}\boldsymbol{\theta} = \boldsymbol{\theta},$$

And from theorem 1.6 we find

$$\begin{aligned} D(\hat{\boldsymbol{\theta}}) &= (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}(\sigma^2\Sigma)\Sigma^{-1}\mathbf{x}(\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1} \\ &= \sigma^2(\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}, \end{aligned}$$

■

The situation is illustrated in the following figure 2.2.

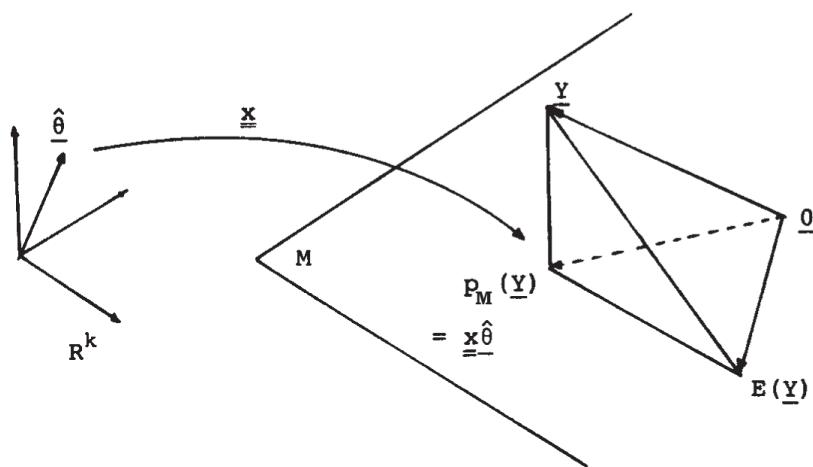


Figure 2.2: Geometric sketch of the problem of estimation in the general linear model.

|||| Remark 2.4

We note that θ is estimated by minimising the squared distance onto M . $\hat{\theta}$ is therefore also a *least squares estimate* of θ . If we do not have the distributional assumption we will often be able to use the estimator $\hat{\theta}$ in theorem 2.3 as an estimate of θ . It can be shown that the least squares estimator $\hat{\theta}$ has the least generalised variance among all the estimators that are linear functions of the observations (the so-called *Gauss-Markov theorem*) cf. [5]. We also say that the least squares estimators are *BLUE - Best Linear Unbiased Estimators*.

Since σ^2 is often unknown we will now find estimators for it. We have

|||| Theorem 2.5

Let the situation be as above. The maximum likelihood estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n} \|Y - \mathbf{x}\hat{\theta}\|^2 = \frac{1}{n} (Y - \mathbf{x}\hat{\theta})' \Sigma^{-1} (Y - \mathbf{x}\hat{\theta}).$$

The unbiased estimator of σ^2 is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - rkx} \|Y - \mathbf{x}\hat{\theta}\|^2 \\ &= \frac{1}{n - rkx} (Y - \mathbf{x}\hat{\theta})' \Sigma^{-1} (Y - \mathbf{x}\hat{\theta})\end{aligned}$$

where $\mathbf{x}\hat{\theta}$ is the maximum likelihood estimator of $E(Y)$. The following holds

$$\hat{\sigma}^2 \in \sigma^2 \chi^2(n - rkx) / (n - rkx)$$

and $\hat{\sigma}^2$ is independent of the maximum likelihood estimator of the expected value and is therefore independent of $\hat{\theta}$.

|||| Proof

The likelihood function is

$$L(\theta, \sigma^2) = k \cdot \frac{1}{\sigma^n} \exp\left[-\frac{1}{2} \frac{1}{\sigma^2} \|y - \mathbf{x}\theta\|^2\right],$$

and

$$\ln L(\theta, \sigma^2) = k_1 - \frac{n}{2} \ln \sigma^2 - \frac{1}{2} \frac{1}{\sigma^2} \|y - \mathbf{x}\theta\|^2.$$

now

$$\begin{aligned}\frac{\partial}{\partial \sigma^2} \ln L &= -\frac{n}{2} \frac{1}{\sigma^2} + \frac{1}{2\sigma^4} \|\mathbf{y} - \mathbf{x}\boldsymbol{\theta}\|^2 \\ &= -\frac{n}{2} \frac{1}{\sigma^4} (\sigma^2 - \frac{1}{n} \|\mathbf{y} - \mathbf{x}\boldsymbol{\theta}\|^2).\end{aligned}$$

After differentiating with respect to $\boldsymbol{\theta}$ we get the ordinary system of normal equations. We therefore find that the maximum likelihood estimates to $(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$ for $(\boldsymbol{\theta}, \sigma^2)$ are solutions for

$$\begin{aligned}\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} \hat{\boldsymbol{\theta}} &= \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{Y} \\ \hat{\sigma}^2 &= \frac{1}{n} \|\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}}\|^2 = \frac{1}{n} (\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}}) \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}}).\end{aligned}$$

If we consider the partitioning of R^n as the direct sum of M and M^\perp , where M^\perp is the orthogonal component (with respect to $\boldsymbol{\Sigma}^{-1}$) of M , we get that

$$\mathbf{P}_M(\mathbf{Y} - \mathbf{x} \boldsymbol{\theta}) = \mathbf{x} \hat{\boldsymbol{\theta}} - \mathbf{x} \boldsymbol{\theta}$$

and

$$\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}}$$

are stochastically independent and that

$$\begin{aligned}\|\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}}\|^2 &\in \sigma^2 \chi^2(\dim M^\perp) \\ &= \sigma^2 \chi^2(n - \text{rk} \mathbf{x}).\end{aligned}$$

From this we especially get

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n} (n - \text{rk} \mathbf{x}) \sigma^2,$$

i.e. the likelihood estimator of σ^2 is not unbiased. If we want an unbiased estimate we can obviously use

$$\frac{1}{n - \text{rk} \mathbf{x}} \|\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}}\|^2.$$

Most often we will be using the unbiased estimate of σ^2 , and we will therefore use the notation $\hat{\sigma}^2$ for this.

■

||| Remark 2.6

If Σ is the identity matrix then $\|\mathbf{y}\|^2 = \sum y_i^2$. So in this case we have

$$\hat{\sigma}^2 = \frac{1}{n - \text{rk}\mathbf{x}} \sum_{i=1}^n (\mathbf{Y}_i - \hat{\mathbf{e}}(\mathbf{Y}_i))^2,$$

where $\hat{\mathbf{e}}(\mathbf{Y}_i) = (\mathbf{x}\hat{\boldsymbol{\theta}})_i$.

||| Definition 2.7

The deviation

$$R_i = \mathbf{Y}_i - \hat{\mathbf{e}}(\mathbf{Y}_i) = \mathbf{Y}_i - (\mathbf{x}\hat{\boldsymbol{\theta}})_i$$

between the i 'th observation and its estimated value $\hat{\mathbf{e}}(\mathbf{Y}_i) = (\mathbf{x}\hat{\boldsymbol{\theta}})_i$ is called the i 'th *residual*. The squared distance between the observation and the estimated model is

$$\text{SSR} = \text{SS}_{\text{res}} = \|\mathbf{Y} - \mathbf{x}\boldsymbol{\theta}\|^2 = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})' \Sigma^{-1} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}).$$

In the case $\Sigma = \mathbf{I}$ we see that SSR is the sum of the squared residuals, and also in the general case (if misunderstandings don't occur) we will denote this as the *residual sum of squares*.

Before we will go on we will give a small example for the purpose of illustration.

||| Example 2.8

In the production of a certain synthetic product two raw materials A and B are mainly used. The quality of the end product can be described by a stochastic variable which is normally distributed with mean value μ and variance σ^2 . The mean-value is known to depend linearly on the added amount of A and B respectively i.e.

$$\mu = x_A \theta_A + x_B \theta_B,$$

where x_A is the added amount of A and x_B is the corresponding added amount of B. σ^2 is assumed to be independent of the added amount of raw-materials. For the determination of θ_A and θ_B three experiments were performed after the following plan.

Experiment	Content of A	Content of B
1	100%	0%
2	0%	100%
3	50%	50%

The single experiments are assumed to be stochastically independent. The simultaneous distribution of the experimental results Y_1, Y_2, Y_3 is then a three dimensional normal distribution with mean value

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_A \\ \theta_B \end{bmatrix} = \mathbf{x}\boldsymbol{\theta},$$

and variance-covariance matrix $\sigma^2\mathbf{I}$.

We have

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} \frac{5}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{5}{4} \end{bmatrix} \Rightarrow (\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix},$$

and

$$\mathbf{x}'\mathbf{y} = \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix},$$

giving

$$\begin{bmatrix} \hat{\theta}_A \\ \hat{\theta}_B \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix} \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix} = \begin{bmatrix} \frac{5}{6}y_1 - \frac{1}{6}y_2 + \frac{1}{3}y_3 \\ -\frac{1}{6}y_1 + \frac{5}{6}y_2 + \frac{1}{3}y_3 \end{bmatrix}.$$

In this case we observed

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 90 \\ 30 \\ 75 \end{bmatrix},$$

so that

$$\begin{bmatrix} \hat{\theta}_A \\ \hat{\theta}_B \end{bmatrix} = \begin{bmatrix} 95 \\ 35 \end{bmatrix}.$$

From this we easily find

$$\hat{\mathbf{e}}(\mathbf{Y}) = \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 95 \\ 35 \\ 65 \end{bmatrix},$$

and

$$\mathbf{Y} - \hat{\mathbf{e}}(\mathbf{Y}) = \mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} -5 \\ -5 \\ 10 \end{bmatrix}.$$

This gives the residual sum of squares

$$(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = 25 + 25 + 100 = 150,$$

Alternatively we may compute

$$\begin{aligned} (\mathbf{x}\hat{\boldsymbol{\theta}})'(\mathbf{x}\hat{\boldsymbol{\theta}}) &= 14475 \\ \mathbf{y}'\mathbf{y} &= 14625 \end{aligned}$$

and obtain the sum of squared residuals as the difference between those, i.e. $14625 - 14475 = 150$. In any case we obtain that an unbiased estimate of σ^2 is

$$\frac{1}{3-2}150 = 150$$

2.1.3 The case of $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ singular

If $\text{rk}(\mathbf{x}) = p < k$ then $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ is singular and we cannot find a unique solution to the equation.

$$(\mathbf{x}'\Sigma^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} = \mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

However, if we have a pseudoinverse for $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ then we can write

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-}\mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

However, sometimes it is possible to use a little trick in the determination of the pseudo inverse. The reason for the singularity is that we have too many parameters. It would therefore be reasonable to restrict $\boldsymbol{\theta}$ to only vary freely in a (side-)subspace of R^k . See fig. 2.3. One of those could e.g. be determined by $\boldsymbol{\theta}$ satisfying the linear equations (restrictions)

$$\mathbf{b}\boldsymbol{\theta} = \mathbf{c}$$

or

$$\begin{bmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}.$$

If there exist $\boldsymbol{\theta}$'s that satisfy this equation system then they span a subspace of dimension $k - \text{rk}(\mathbf{b})$.

Since

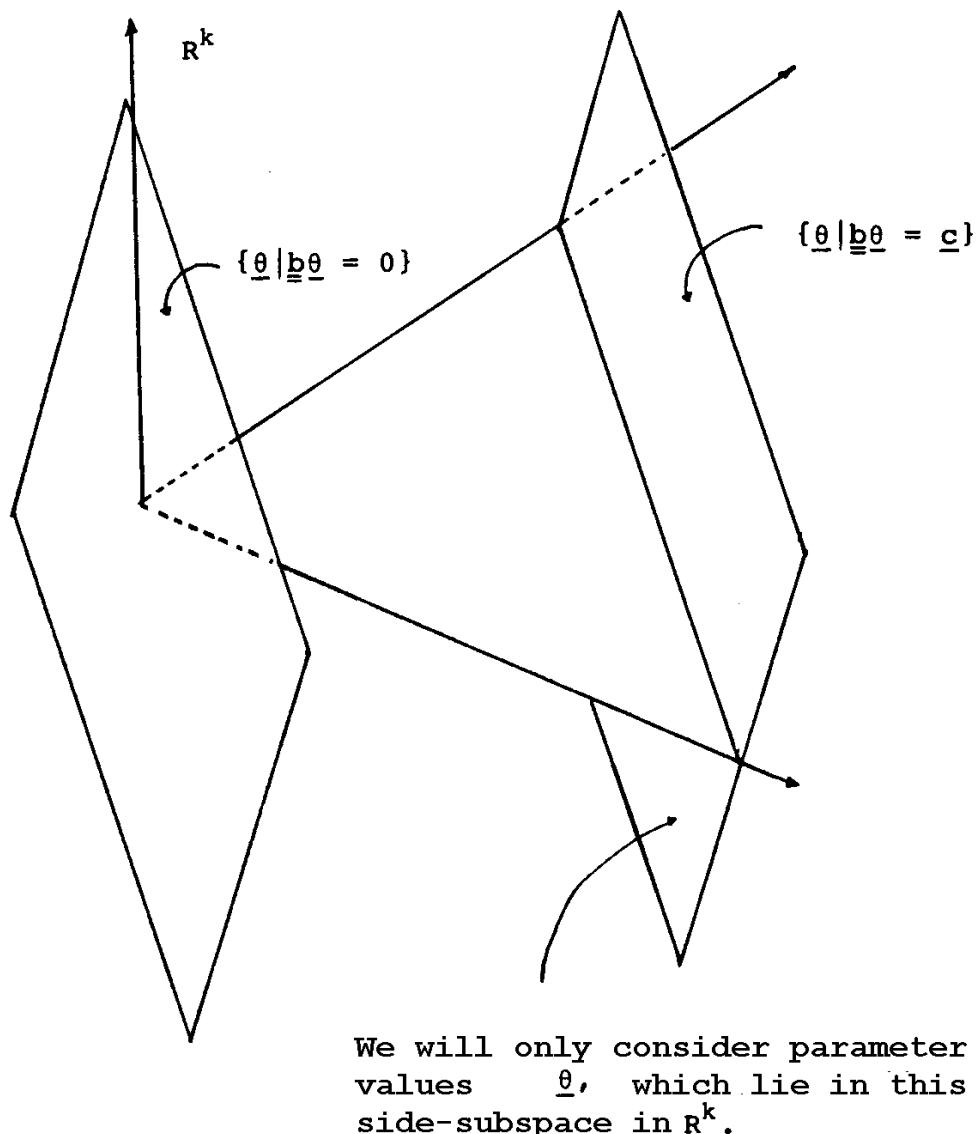


Figure 2.3:

$\text{rk}(\mathbf{x}) = p$, and we have $k \theta$ -components it would be reasonable to remove $k - p$ of these i.e. impose the restriction $k - \text{rk}(\mathbf{b}) = p$ or $k = p + \text{rk}(\mathbf{b})$.

Now if

$$\text{rk} \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix} = \text{rk} \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \\ b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mk} \end{bmatrix} = k,$$

we can consider the model

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix} \boldsymbol{\theta} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \mathbf{0} \end{bmatrix}.$$

We let

$$\mathbf{D} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0}_{n,m} \\ \mathbf{0}_{m,n} & \mathbf{I}_{m,m} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

where the short notation should not cause confusion.

If we in the usual way compute

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \{[\mathbf{x}' \mathbf{b}'] \mathbf{D} \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}\}^{-1} \{[\mathbf{x}' \mathbf{b}'] \mathbf{D} \begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix}\} \\ &= \{\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} + \mathbf{b}' \mathbf{b}\}^{-1} \{\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{y} + \mathbf{b}' \mathbf{c}\}, \end{aligned}$$

then we have a quantity which minimises

$$\begin{aligned} g(\boldsymbol{\theta}) &= \left\{ \begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix} - \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix} \boldsymbol{\theta} \right\}' \mathbf{D} \left\{ \begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix} - \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix} \boldsymbol{\theta} \right\} \\ &= \begin{bmatrix} \mathbf{y} - \mathbf{x} \boldsymbol{\theta} \\ \mathbf{0} \end{bmatrix}' \begin{bmatrix} \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{y} - \mathbf{x} \boldsymbol{\theta} \\ \mathbf{0} \end{bmatrix} \\ &= (\mathbf{y} - \mathbf{x} \boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \mathbf{x} \boldsymbol{\theta}) \\ &= \|\mathbf{y} - \mathbf{x} \boldsymbol{\theta}\|^2. \end{aligned}$$

Since this is exactly the same quantity we must minimize in order to find the ML-estimates, we find that

$$\hat{\boldsymbol{\theta}} = \{\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} + \mathbf{b}' \mathbf{b}\}^{-1} \{\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{y} + \mathbf{b}' \mathbf{c}\}$$

really is the maximum likelihood estimator for $\boldsymbol{\theta}$. The only requirement is that we must find a matrix \mathbf{b} so $\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}$ has full rank and this corresponds to restricting $\boldsymbol{\theta}$'s region of variation.

The variance-covariance matrix of $\hat{\theta}$ becomes

$$D(\hat{\theta}) = \sigma^2 \{ \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} + \mathbf{b}' \mathbf{b} \}^{-1} \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} \{ \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} + \mathbf{b}' \mathbf{b} \}^{-1}.$$

This expression is found immediately by using theorem 1.6.

As before the unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n - rk\mathbf{x}} \| \mathbf{Y} - \mathbf{x}\hat{\theta} \|^2 = \frac{1}{n - rk\mathbf{x}} (\mathbf{Y} - \mathbf{x}\hat{\theta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\theta})$$

Here we have $n - rk\mathbf{x} = n - k + rkb$.

First we give a little theoretical

||| Example 2.9

Consider a very simple one-sided analysis of variance with two groups with two observations in each group. We could imagine that we were examining the effect of a catalyst on the results of some process. We therefore conduct four experiments, two with the catalyst at level A and two with the catalyser at level B. We therefore have the following observations

level A: Y_{11}, Y_{12}

level B: Y_{21}, Y_{22}

If we assume that the observations are stochastically independent and have mean values

$$E(Y_{11}) = E(Y_{12}) = \theta_1$$

$$E(Y_{21}) = E(Y_{22}) = \theta_2,$$

then we can express the model as

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \varepsilon = \mathbf{x}\boldsymbol{\theta} + \varepsilon.$$

We easily find that

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

and

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix},$$

which are the usual estimators. If we instead use the (commonly used) parametrisation

$$E(Y_{11}) = E(Y_{12}) = \mu + \alpha_1$$

$$E(Y_{21}) = E(Y_{22}) = \mu + \alpha_2$$

i.e. we express the effect of a catalyst as a level plus the specific effect of that catalyst. Then we have

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \varepsilon = \mathbf{x}\boldsymbol{\alpha} + \varepsilon.$$

It is easily seen that \mathbf{x} has rank 2 (the sum of the last two columns equals the first). We will therefore try to introduce a linear restriction between the parameters. We will try with

$$\alpha_1 + \alpha_2 = 0 \quad \text{i.e. : } (0 \ 1 \ 1) \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = 0.$$

We can now formally introduce the model

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix},$$

or

$$\begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ 0 \ 1 \ 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}.$$

We now have that

$$\begin{bmatrix} \mathbf{x} \\ 0 \ 1 \ 1 \end{bmatrix}' \begin{bmatrix} \mathbf{x} \\ 0 \ 1 \ 1 \end{bmatrix} = \mathbf{x}'\mathbf{x} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

The inverse of this matrix is

$$\begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}.$$

Now, since

$$\begin{bmatrix} \mathbf{x} \\ 0 \ 1 \ 1 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} \sum y_{ij} \\ y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix},$$

we have

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \sum y_{ij} \\ y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{y}_1 - \bar{y} \\ \bar{y}_2 - \bar{y} \end{bmatrix},$$

i.e. exactly the same estimators we are used to from a balanced one-sided analysis of variance (note: We know in beforehand that we will get these estimators cf. the results earlier in this section).

We will now give a more practical example of the estimation of parameters in the case where $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ is singular.

||| Example 2.10

In the production of enzymes one can use two principally different types of bacteria. Via its metabolism one type of bacteria liberates acid during the production (acid producer). The other produces neutral metabolic products. In order to regulate the pH-value in the substrate on which the bacteria are produced, one can add a so-called pH-buffer. It is known, that the pH-buffer itself does not have any effect on the production of the enzyme, rather it works through an interaction with the acid content and the metabolic products of the bacteria.

For a "neutral" type of bacteria which lives on a substrate without pH-buffer the mean production of enzyme (normal production) is known. In order to estimate the above mentioned interactions one has measured the difference between the normal production and the actual production of enzyme in 7 experiments as shown below.

		pH-buffer	
		added	not added
bacteria culture	acid producer	0,-2	-19,-15
	neutral	-6, 0,-2	

Table 2.10: Differences between nominal yield and actual yield under different experimental circumstances.

First we will formulate a mathematical model that can describe the above mentioned experiment.

We have observations

$$\begin{aligned} y_{11\nu}, \quad \nu &= 1, 2 \\ y_{12\nu}, \quad \nu &= 1, 2 \\ y_{21\nu}, \quad \nu &= 1, 2, 3. \end{aligned}$$

These are assumed to have the mean values

$$\begin{aligned} E(y_{11\nu}) &= \mu_1 + \theta_{11} \\ E(y_{12\nu}) &= \mu_1 + \theta_{12} \\ E(y_{21\nu}) &= \theta_{21}, \end{aligned}$$

where μ_1 is the effect of using acid producing bacteria and θ_{ij} is the interaction between pH-buffer and bacteria culture.

Furthermore we assume that the observations are stochastically independent and we have the same but unknown variance σ^2 .

We can now formulate the model as a general linear model. We have

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{213} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \end{bmatrix} + \varepsilon,$$

where the error $\varepsilon \in N_7(\mathbf{0}, \sigma^2 \mathbf{I})$.

We find

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

and

$$\mathbf{x}'\mathbf{y} = \begin{bmatrix} y_{1..} \\ y_{11.} \\ y_{12.} \\ y_{21.} \end{bmatrix},$$

where a dot as an index-value indicates that we have summed over the corresponding index.

Since $\mathbf{x}'\mathbf{x}$ only has the rank 3, we are unable to invert it. Instead we can find a pseudo-inverse. We use the theorem A.17 p. 357 and get

$$(\mathbf{x}'\mathbf{x})^{-} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix},$$

so the estimates from the parameters become - with this special choice of pseudo-inverse -

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'\mathbf{y} = \begin{bmatrix} 0 \\ \bar{y}_{11.} \\ \bar{y}_{12.} \\ \bar{y}_{21.} \end{bmatrix},$$

where e.g.

$$\bar{y}_{21.} = \frac{1}{3} \sum_{v=1}^3 y_{21v}.$$

Now, since

$$\mathbf{I} - (\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

we have

$$(\mathbf{I} - (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{x})\mathbf{z} = \begin{bmatrix} z_1 \\ -z_1 \\ -z_1 \\ 0 \end{bmatrix}$$

From theorem A.15 the complete solution to the normal equations is therefore all vectors of the form

$$\hat{\boldsymbol{\theta}} + \begin{bmatrix} t \\ -t \\ -t \\ 0 \end{bmatrix} = \begin{bmatrix} t \\ \bar{y}_{11} - t \\ \bar{y}_{12} - t \\ \bar{y}_{21} \end{bmatrix}, \quad t \in R.$$

An arbitrary maximum likelihood estimator for $\boldsymbol{\theta}$ is then of this form.

The observed value of $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}}_{\text{obs}} = \begin{bmatrix} 0 \\ -1 \\ -17 \\ -2\frac{2}{3} \end{bmatrix}.$$

It is obvious that this estimator is not very satisfactory since e.g. $\hat{\mu}_1$ always will be 0. In order to get estimators which correspond to our expectations about physical reality we must impose some constraints on the parameters. It seems reasonable to demand that

$$\theta_{11} + \theta_{12} = 0,$$

i.e.

$$(0 \ 1 \ 1 \ 0) \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \end{bmatrix} = 0,$$

or

$$\mathbf{b}'\boldsymbol{\theta} = 0.$$

It is obvious that

$$\text{rk}(\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}) = 4,$$

so we can use the result from p. 85. We find

$$\begin{aligned} \mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b} &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 3 & 1 & 0 \\ 2 & 1 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}. \end{aligned}$$

Since

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix},$$

we find

$$(\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b})^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

We now get

$$\hat{\theta} = (\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b})^{-1}\mathbf{x}'\mathbf{y} = \begin{bmatrix} \bar{y}_{1..} \\ \bar{y}_{11..} - \bar{y}_{1..} \\ \bar{y}_{12..} - \bar{y}_{1..} \\ \bar{y}_{21..} \end{bmatrix}.$$

The observed value is

$$\begin{bmatrix} -9 \\ 8 \\ -8 \\ -2\frac{2}{3} \end{bmatrix} \left(= \begin{bmatrix} \text{acid producing effect} \\ \text{buffer \& acid interaction} \\ (-\text{buffer}) \& \text{acid interaction} \\ \text{buffer \& neutral interaction} \end{bmatrix} \right).$$

We now find the variance-covariance matrix for $\hat{\theta}$. We have

$$\begin{aligned} D(\hat{\theta}) &= \sigma^2(\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b})^{-1}\mathbf{x}'\mathbf{x}(\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b})^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}, \end{aligned}$$

i.e. the estimators are not independent.

In order to estimate σ^2 we find the vector of residuals. Since

$$\mathbf{x}\hat{\theta} = \begin{bmatrix} \hat{\mu}_1 + \hat{\theta}_{11} \\ \hat{\mu}_1 + \hat{\theta}_{11} \\ \hat{\mu}_1 + \hat{\theta}_{12} \\ \hat{\mu}_1 + \hat{\theta}_{12} \\ \hat{\theta}_{21} \\ \hat{\theta}_{21} \\ \hat{\theta}_{21} \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -17 \\ -17 \\ -2\frac{2}{3} \\ -2\frac{2}{3} \\ -2\frac{2}{3} \end{bmatrix},$$

the vector of residuals is

$$\mathbf{y} - \mathbf{x}\hat{\theta} = \begin{bmatrix} 1 \\ -1 \\ -2 \\ 2 \\ -3\frac{1}{3} \\ 2\frac{2}{3} \\ \frac{2}{3} \end{bmatrix}.$$

We then find

$$\|\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\theta}}\|^2 = (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\theta}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = 1^2 + \cdots + \left(\frac{2}{3}\right)^2 = 28\frac{2}{3}.$$

An unbiased estimate of σ^2 is therefore

$$s^2 = \frac{1}{7-3} \cdot 28\frac{2}{3} = 7\frac{1}{6}.$$

2.1.4 Constrained estimation

We now consider a problem that resembles the situation in the previous sections. More specifically we want to estimate parameters that satisfy a linear constraint

$$\mathbf{H}'\boldsymbol{\theta} = \boldsymbol{\xi}.$$

This is e.g. the case when estimating angles in a triangle. They obviously satisfy

$$(1 \ 1 \ 1) \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 180^\circ,$$

and therefore we must require this for the estimates as well.

The main result on estimation of $\boldsymbol{\theta}$ is expressed in

||| Theorem 2.11

Let $E(\mathbf{Y}) = \mathbf{x}\boldsymbol{\theta}$ where \mathbf{Y} is an n -dimensional random variable, \mathbf{x} a known $n \times k$ matrix and $\boldsymbol{\theta}$ a k -dimensional vector of unknown parameters satisfying the s linear constraints

$$\mathbf{H}'\boldsymbol{\theta} = \boldsymbol{\xi},$$

where \mathbf{H} is a known $k \times s$ matrix and $\boldsymbol{\xi}$ a known s -dimensional vector. Finally we suppose that $D(\mathbf{Y}) = \sigma^2 \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is known. The least squares estimator $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ under the constraint $H'\boldsymbol{\theta} = \boldsymbol{\xi}$ is a solution to the equations

$$\begin{bmatrix} \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} & \mathbf{H}' \\ \mathbf{H} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ \boldsymbol{\xi} \end{bmatrix}.$$

|||| Proof

We must determine a θ that minimizes

$$\min_{\mathbf{H}'\theta = \xi} (\mathbf{Y} - \mathbf{x}\theta)' \Sigma^{-1} (\mathbf{Y} - \mathbf{x}\theta).$$

we introduce the Lagrange multiplier λ and put

$$F(\theta, \lambda) = \frac{1}{2} (\mathbf{Y} - \mathbf{x}\theta)' \Sigma^{-1} (\mathbf{Y} - \mathbf{x}\theta) + \lambda' (\mathbf{H}'\theta - \xi).$$

Then

$$\begin{aligned} \frac{\partial F}{\partial \theta} &= -\mathbf{x}' \Sigma^{-1} \mathbf{y} + \mathbf{x}' \Sigma^{-1} \mathbf{x} \theta + \mathbf{H} \lambda \\ \frac{\partial F}{\partial \lambda} &= \mathbf{H}' \theta - \xi. \end{aligned}$$

Those two derivatives are 0 in any extremum for $(\mathbf{Y} - \mathbf{x}\theta)' \Sigma^{-1} (\mathbf{Y} - \mathbf{x}\theta)$ under the constraint $\mathbf{H}'\theta = \xi$. Using this, the result in the theorem follows immediately.

■

Next we consider the problem of estimating σ^2 in

|||| Theorem 2.12

Letting

$$\begin{bmatrix} \mathbf{x}' \Sigma^{-1} \mathbf{x} & \mathbf{H} \\ \mathbf{H}' & \mathbf{0} \end{bmatrix}^- = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & \mathbf{C}_4 \end{bmatrix}.$$

be a pseudoinverse to the coefficient matrix in Theorem 2.11. Then

$$\mathbf{D}(\tilde{\theta}) = \sigma^2 \mathbf{C}_1,$$

and an unbiased estimation of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{f} (\mathbf{Y}' \Sigma^{-1} \mathbf{Y} - \tilde{\theta}' \mathbf{x}' \Sigma^{-1} \mathbf{Y} - \xi' \tilde{\lambda}),$$

where $(\tilde{\theta}', \tilde{\lambda}')'$ is a solution to the equations in Theorem 2.11, and

$$f = n - \text{rk}(\mathbf{x}', \mathbf{H}) + \text{rk}(\mathbf{H}).$$

|||| Proof

By introducing the pseudoinverse we get

$$\begin{aligned}\tilde{\theta} &= \mathbf{C}_1 \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{Y} + \mathbf{C}_2 \boldsymbol{\xi} \\ \tilde{\lambda} &= \mathbf{C}_3 \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{Y} + \mathbf{C}_4 \boldsymbol{\xi}.\end{aligned}$$

From this we immediately obtain

$$\begin{aligned}D(\tilde{\theta}) &= \sigma^2 \mathbf{C}_1 \mathbf{x}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \mathbf{x} \mathbf{C}_1' \\ &= \sigma^2 \mathbf{C}_1 \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} \mathbf{C}_1' \\ &= \sigma^2 \mathbf{C}_1\end{aligned}$$

The last equality sign follows by using properties of pseudoinverse matrices.

By plugging in it is seen that

$$(\mathbf{Y} - \mathbf{x} \tilde{\theta})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x} \tilde{\theta}) = (\mathbf{Y}' \boldsymbol{\Sigma}^{-1} \mathbf{Y} - \tilde{\theta}' \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{Y} - \boldsymbol{\xi}' \tilde{\lambda}),$$

so now it just remains to be shown that the degrees of freedom are as postulated in the theorem. The solution to

$$\mathbf{H}' \theta = \boldsymbol{\xi},$$

can be written as

$$\theta = \theta_0 + \mathbf{B} \beta,$$

where θ_0 is a particular solution and \mathbf{B} is a $(k \times s)$ matrix ($\text{rk}(\mathbf{H}) = k - s$) satisfying

$$\mathbf{H}' \mathbf{B} = \mathbf{0}.$$

Finally β is a s -dimensional vector of "free", new parameters. If we consider

$$\mathbf{Z} = \mathbf{Y} - \mathbf{x} \theta_0,$$

we get

$$\begin{aligned}E(\mathbf{Z}) &= \mathbf{x} \theta - \mathbf{x} \theta_0 = \mathbf{x}(\theta - \theta_0) \\ &= \mathbf{x}(\theta - \theta_0) \\ &= \mathbf{x} \mathbf{B} \beta.\end{aligned}$$

We may now consider the model

$$\mathbf{Z} = \mathbf{x} \mathbf{B} \beta + \varepsilon,$$

where ε is the error vector and solve this. By doing this we obtain the earlier stated estimates.

Letting

$$\hat{\beta} = (\mathbf{B}' \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} \mathbf{B})^{-1} \mathbf{B}' \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{Y},$$

we obtain

$$\tilde{\theta} = \theta_0 + \mathbf{B} \hat{\beta},$$

and consequently

$$\begin{aligned} \mathbf{Y} - \mathbf{x}\tilde{\theta} &= \mathbf{Z} + \mathbf{x}\theta_0 - \mathbf{x}\theta_0 - \mathbf{x}\mathbf{B}\hat{\beta} \\ &= \mathbf{Z} - \mathbf{x}\mathbf{B}\hat{\beta}. \end{aligned}$$

From the general theory it follows that the degrees of freedom are $n - \text{rk}(\mathbf{x}\mathbf{B})$. We have

$$\begin{aligned} \text{rk}(\mathbf{x}\mathbf{B}) &= \dim\{\mathbf{x}\mathbf{B}\beta | \beta \in R^s\} \\ &= \dim\{\mathbf{x}\gamma | \mathbf{H}'\gamma = 0, \gamma \in R^m\} \\ &= \text{rk}\left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}' \end{array}\right) - \text{rk}(\mathbf{H}). \end{aligned}$$

The last equality sign follows from the relation

$$\dim S_1^* + \dim S_2^* = \text{rk}\left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}' \end{array}\right),$$

where

$$\begin{aligned} S_1^* &= \left\{ \left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}' \end{array} \right) \gamma \mid \gamma \in N(H) \right\} \\ S_2^* &= \left\{ \left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}' \end{array} \right) \gamma \mid \gamma \in N(H)^\perp \right\} \end{aligned}$$

remembering that $\dim S_2^* = \text{rk}\mathbf{H}$.

■

We now present an illustrative example.

||| Example 2.13

Suppose that we have 3×2 independent measurements of the angles in a triangle (e.g. measured in the field), and that they were

$$\begin{aligned} v_1 &= 52^\circ, 54^\circ \\ v_2 &= 74^\circ, 74^\circ \\ v_3 &= 48^\circ, 46^\circ. \end{aligned}$$

Furthermore we suppose that the uncertainty on these values are the same and may be expressed by a variance σ^2 .

We state this as a linear model with constraints, i.e.

$$\begin{bmatrix} v_{11} \\ v_{12} \\ v_{21} \\ v_{22} \\ v_{31} \\ v_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$$(1,1,1) \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 180,$$

$$D(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}.$$

We get

$$\left[\begin{array}{cc|c} \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} & \mathbf{H} \\ \mathbf{H}' & \mathbf{0} \end{array} \right] = \left[\begin{array}{ccc|c} 2 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ \hline 1 & 1 & 1 & 0 \end{array} \right].$$

A (pseudo)inverse of this matrix is

$$\left[\begin{array}{cc} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & \mathbf{C}_4 \end{array} \right] = \frac{1}{6} \left[\begin{array}{ccc|c} 2 & -1 & -1 & 2 \\ -1 & 2 & -1 & 2 \\ -1 & -1 & 2 & 2 \\ \hline 2 & 2 & 2 & -4 \end{array} \right].$$

Therefore we get

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= \frac{1}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 106 \\ 148 \\ 94 \end{bmatrix} + \frac{1}{6} \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} 180 \\ &= \begin{bmatrix} -5 \\ 16 \\ -11 \end{bmatrix} + \begin{bmatrix} 60 \\ 60 \\ 60 \end{bmatrix} \\ &= \begin{bmatrix} 55 \\ 76 \\ 49 \end{bmatrix}. \end{aligned}$$

We observe that the sum of the estimates is 180.

The dispersion matrix is

$$D(\tilde{\boldsymbol{\theta}}) = \sigma^2 \mathbf{C}_1 = \frac{\sigma^2}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

The estimate of σ^2 is

$$\sigma^2 = \frac{1}{6-3+1} (20992 - 21684 - (-720)) = 7 = 2.6^2,$$

since

$$\tilde{\lambda} = \frac{1}{6}[2, 2, 2] \begin{bmatrix} 106 \\ 148 \\ 94 \end{bmatrix} + \left(-\frac{4}{6}\right)180 = 116 - 120 = -4$$

||| Remark 2.14

As indicated in the example this type of model is particularly relevant in *geodesy* and *surveying*.

2.1.5 Confidence-intervals for estimated values. Prediction-intervals

We consider the usual model ($n > k$)

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where

$$\varepsilon \in N(\mathbf{0}, \sigma^2 \Sigma).$$

Here we will denote the Y 's as dependent variables and the x 's as the independent variables.

As usual σ^2 is (assumed) unknown and Σ is (assumed) known. We have the estimator

$$\hat{\theta} = (\mathbf{x}' \Sigma^{-1} \mathbf{x})^{-1} \mathbf{x}' \Sigma^{-1} \mathbf{Y}$$

for θ , and σ^2 is estimated using

$$\begin{aligned} \hat{\sigma}^2 &= s^2 = \frac{1}{n-k} \|\mathbf{Y} - \mathbf{x}\hat{\theta}\|^2 \\ &= \frac{1}{n-k} (\mathbf{Y} - \mathbf{x}\hat{\theta})' \Sigma^{-1} (\mathbf{Y} - \mathbf{x}\hat{\theta}). \end{aligned}$$

If we wish to estimate the expected value of the new observation Y of the dependent variable corresponding to the values of the independent variables:

$$(z_1, \dots, z_k) = \mathbf{z}'$$

i.e. a new row in the \mathbf{x} -matrix, it is obvious that we will use

$$U = (z_1, \dots, z_k) \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_k \end{bmatrix} = \mathbf{z}' \hat{\boldsymbol{\theta}}$$

as our predictor.

We have that $E(U) = E(Y)$ and that

$$\begin{aligned} V(U) &= \mathbf{z}' D(\hat{\boldsymbol{\theta}}) \mathbf{z} \\ &= \sigma^2 \mathbf{z}' (\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \mathbf{z} \\ &= \sigma^2 c, \end{aligned}$$

where

$$c = (z_1, \dots, z_k) (\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix}.$$

We therefore immediately have

$$\frac{U - E(Y)}{\sigma \sqrt{c}} \in N(0, 1),$$

and therefore also

$$\frac{U - E(Y)}{S \sqrt{c}} \in t(n - k).$$

We are now able to formulate and prove

|||| Theorem 2.15

Let the situation be as above. Then the $(1 - \alpha)$ -confidence interval for the expected value of a new observation Y will be

$$[u - t(n - k)_{1-\frac{\alpha}{2}} s \sqrt{c}, \quad u + t(n - k)_{1-\frac{\alpha}{2}} s \sqrt{c}].$$

|||| Proof

From the above considerations we have

$$1 - \alpha = P\{U - t(n - k)_{1-\frac{\alpha}{2}} s \sqrt{c} \leq E(Y) \leq U + t(n - k)_{1-\frac{\alpha}{2}} s \sqrt{c}\},$$

and therefore we immediately have the theorem. ■

Often one is more interested in a confidence interval for the new (or future) observations than for the expected value of the observations. We now consider the more general problem of determining the confidence interval for the average \bar{Y}_q of q (coming) observations taken at (z_1, \dots, z_k) . If $Y_{iq} \in N(E(Y), c_1\sigma^2)$, then we have that

$$\bar{Y}_q \in N(E(Y), \frac{c_1}{q}\sigma^2).$$

If we now assume that the new (or future) observations are independent of those we already have then

$$U - \bar{Y}_q \in N(0, \sigma^2(c + \frac{c_1}{q})),$$

i.e.

$$\frac{U - \bar{Y}_q}{S\sqrt{c + \frac{c_1}{q}}} \in t(n - k).$$

From this we can as before derive

|||| Theorem 2.16

Let us assume that q new observations taken at (z_1, \dots, z_k) each have a variance $c_1\sigma^2$. Furthermore, they are independent of each other and independent of the earlier observations. In that case a $(1 - \alpha)$ prediction interval for the average of the q observations equals the interval

$$[u - t(n - k)_{1 - \frac{\alpha}{2}} s \sqrt{c + \frac{c_1}{q}}, u + t(n - k)_{1 - \frac{\alpha}{2}} s \sqrt{c + \frac{c_1}{q}}].$$

|||| Remark 2.17

The above mentioned interval is a confidence interval for an observation and not for a parameter as we are used to. One therefore speaks of a *prediction interval* in order to distinguish between the two situations.

|||| **Remark 2.18**

We see that the correspondence to the interval for \bar{Y}_q instead of the interval for $E(\bar{Y}_q) = E(Y)$ just consists of the expression under the square root sign being larger by an amount equal to $\frac{c_1}{q}$ which is the variance of $\frac{\bar{Y}_q}{\sigma}$.

|||| **Example 2.19**

We consider the following corresponding observations of an independent variable x and a dependent variable y :

x	0	1	2	3	4	5	6
y	0.4	0.3	1.5	1.3	1.9	4.2	8

We assume that the y 's originate from independent stochastic variables Y_1, \dots, Y_7 which are normally distributed with mean values

$$E(Y|x) = \beta x^2$$

and variances

$$V(Y|0) = \sigma^2, \quad V(Y|x) = x^2 \sigma^2, \quad x > 0.$$

We would now like to find a confidence interval for a new (or future) observation corresponding to $x = 10$. This observation is called Y , and we have

$$\begin{aligned} E(Y) &= 100\beta \\ V(Y) &= 100\sigma^2. \end{aligned}$$

We now reformulate the problem in matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_7 \end{bmatrix} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$D(\boldsymbol{\varepsilon}) = \sigma^2 \begin{bmatrix} 1 & & & & & 0 \\ & \ddots & \ddots & \ddots & & \\ & & 4 & & & \\ \vdots & & & 9 & & \ddots \\ & & & & 16 & \ddots \\ & & & & & 25 \\ 0 & & \ddots & \ddots & & 36 \end{bmatrix} = \sigma^2 \boldsymbol{\Sigma}.$$

We have that

$$\begin{aligned}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} &= (0, 1, 4, 9, 16, 25, 36) \text{diag}(1, 1, \frac{1}{4}, \dots, \frac{1}{36}) \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 36 \end{bmatrix} \\ &= 91. \\ \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{y} &= 0.3 + 1.5 + 1.3 + 1.9 + 4.2 + 8.0 = 17.2.\end{aligned}$$

so

$$\hat{\beta} = \frac{17.2}{91} = 0.1890,$$

and

$$P_M(\mathbf{y}) = \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \cdot 0.1890 = \begin{bmatrix} 0 \\ 0.1890 \\ 0.7560 \\ 1.7010 \\ 3.0240 \\ 4.7250 \\ 6.8040 \end{bmatrix}.$$

The residuals are

$$\mathbf{y} - P_M(\mathbf{y}) = \begin{bmatrix} 0.4000 \\ 0.1110 \\ 0.7440 \\ -0.4010 \\ -1.1240 \\ -0.5250 \\ 1.1960 \end{bmatrix},$$

so

$$\begin{aligned}\|\mathbf{y} - P_M(\mathbf{y})\|^2 &= (0.4000 \cdots 1.1960) \begin{bmatrix} \frac{1}{1} & & & \\ & \frac{1}{1} & & \\ & & \ddots & \\ & & & \frac{1}{36} \end{bmatrix} \begin{bmatrix} 0.4000 \\ \vdots \\ 1.1960 \end{bmatrix} \\ &= 0.45829\end{aligned}$$

i.e.

$$\hat{\sigma}^2 = s^2 = \frac{1}{7-1} 0.45829 = 0.07638 = 0.27637^2.$$

The constants c and c_1 are equal to

$$\begin{aligned}c &= 100 \cdot \frac{1}{91} \cdot 100 = 109.89 \\ c_1 &= 10^2 = 100.\end{aligned}$$

The prediction for $x = 10$ is

$$z = 100\hat{\beta} = 18.90$$

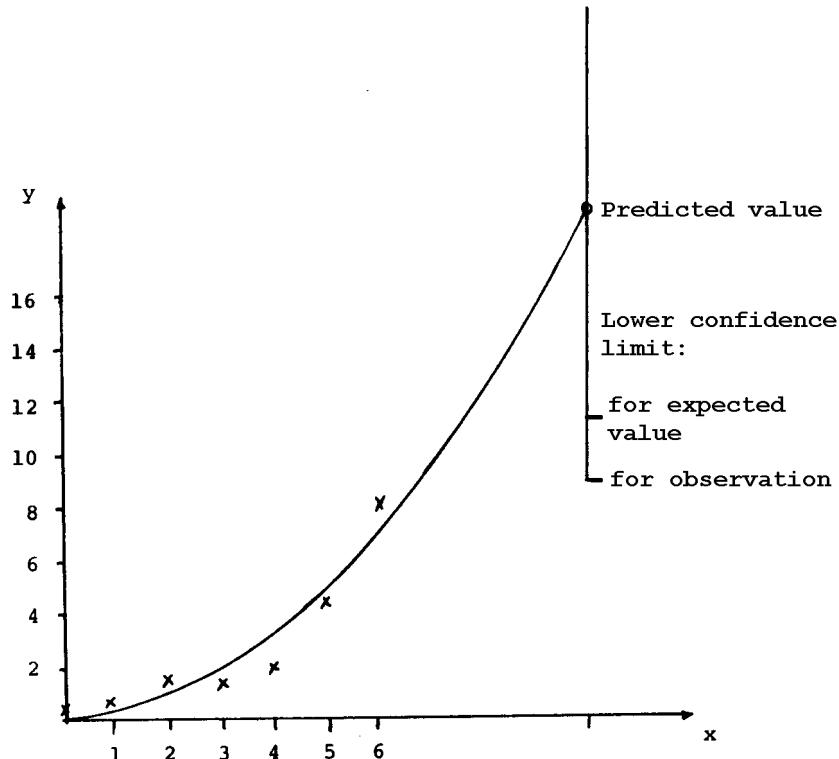
The confidence interval for the expected value at $x = 10$ is therefore given by

$$\begin{aligned} & 18.90 \pm t(6)_{0.975} 0.2764 \sqrt{109.89} \\ &= 18.90 \pm 2.447 \cdot 0.2764 \sqrt{109.89} \\ &= 18.90 \pm 7.09. \end{aligned}$$

The corresponding prediction interval for the next observation is

$$\begin{aligned} & 18.90 \pm t(6)_{0.975} \cdot 0.2764 \sqrt{109.89 + 100} \\ &= 18.90 \pm 9.80, \end{aligned}$$

i.e. a somewhat broader interval than for the expected value. The explanation is simply that we have a variance of $10^2\sigma^2 = 100\sigma^2$ in $x=10$. We depict the observations and estimated polynomial in the following graph. Further the two confidence intervals are given.



2.2 Tests in the general linear model

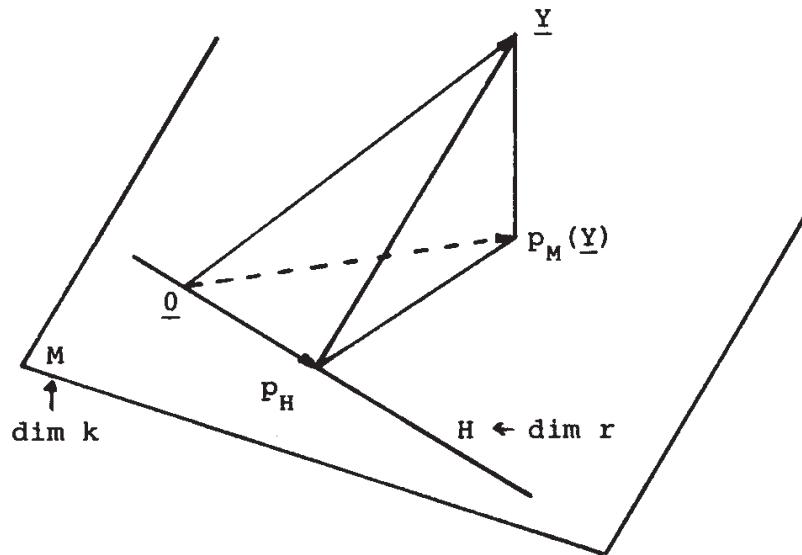
In this section we will check if the mean vector can be assumed to lie in a true sub-space of the model space and also check if the mean vector successively can be assumed to lie in sub-spaces of smaller and smaller dimensions, i.e. we want to successively test whether we can use fewer and fewer parameters to describe the data.

2.2.1 Test for a lower dimension of model space

Let $\mathbf{Y} \in N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is regular and known. We assume that $\boldsymbol{\mu} \in M$, is a k -dimensional sub-space and we will test the hypothesis

$$H_0 : \boldsymbol{\mu} \in H \quad \text{against} \quad H_1 : \boldsymbol{\mu} \in M \setminus H,$$

where H is an r -dimensional sub-space of M . In the following we will consider the norm given by $\boldsymbol{\Sigma}^{-1}$. The maximum likelihood estimator for $\boldsymbol{\mu}$ is then the projection $p_M(\mathbf{Y})$ onto M and if H_0 is true then the maximum likelihood estimator $p_H(\mathbf{Y})$, is \mathbf{Y} 's projection onto H . The ML estimator for σ^2 in the two cases are respectively $\frac{1}{n} \|\mathbf{y} - p_M(\mathbf{y})\|^2$ and $\frac{1}{n} \|\mathbf{y} - p_H(\mathbf{y})\|^2$.



The likelihood function is

$$\begin{aligned} L(\boldsymbol{\mu}, \sigma^2) &= \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sigma^n} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right) \\ &= k \cdot \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \boldsymbol{\mu}\|^2\right). \end{aligned}$$

With this notation we have

|||| **Theorem 2.20**

Let the situation be as above. Then the likelihood ratio test at level α of testing

$$H_0 : \boldsymbol{\mu} \in H \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \in M \setminus H,$$

is equivalent to the test given by the critical region

$$C_\alpha = \{(y_1, \dots, y_n) \mid \frac{\|p_M(\mathbf{y}) - p_H(\mathbf{y})\|^2 / (k-r)}{\|\mathbf{y} - p_M(\mathbf{y})\|^2 / (n-k)} > F(k-r, n-k)_{1-\alpha}\}.$$

|||| **Proof**

The likelihood ratio test statistic is

$$\begin{aligned} Q &= \frac{\sup_{H_0} L(\boldsymbol{\mu}, \sigma^2)}{\sup L(\boldsymbol{\mu}, \sigma^2)} = \frac{L(p_H(\mathbf{y}), \hat{\sigma}^2)}{L(p_M(\mathbf{y}), \hat{\sigma}^2)} \\ &= \left[\frac{\|\mathbf{y} - p_M(\mathbf{y})\|^2}{\|\mathbf{y} - p_H(\mathbf{y})\|^2} \right]^{\frac{n}{2}} \frac{\exp(-\frac{n}{2})}{\exp(-\frac{n}{2})} = \left[\frac{\|\mathbf{y} - p_M(\mathbf{y})\|^2}{\|\mathbf{y} - p_H(\mathbf{y})\|^2} \right]^{\frac{n}{2}}. \end{aligned}$$

From this we see

$$Q < q \iff \frac{\|\mathbf{y} - p_M(\mathbf{y})\|^2}{\|\mathbf{y} - p_H(\mathbf{y})\|^2} < k_1.$$

Since we reject the hypothesis for small values of Q we see that we reject when the length of the leg (cathetus) $\mathbf{Y} - p_M(\mathbf{Y})$ is much less than the length of the hypotenuse. From Pythagoras we have that

$$\|\mathbf{y} - p_H(\mathbf{y})\|^2 = \|\mathbf{y} - p_M(\mathbf{y})\|^2 + \|p_H(\mathbf{y}) - p_M(\mathbf{y})\|^2,$$

we see that we may just as well compare the two legs i.e. use

$$Q < q \iff \frac{\|p_M(\mathbf{y}) - p_H(\mathbf{y})\|^2 / (k-r)}{\|\mathbf{y} - p_M(\mathbf{y})\|^2 / (n-k)} > c. \quad (2-1)$$

Under H_0 both the numerator and denominator are $\sigma^2 \chi^2(f)/f$ distributed with respectively $k-r$ and $n-k$ degrees of freedom and they are furthermore independent (follows from the partition theorem). The ratio will therefore be F-distributed under H_0 , and the theorem follows from this. The reason why we in (2-1) have divided the respective norms with the dimension of the relevant sub-space is of course that we want the test statistic to be F-distributed under H_0 , and not just proportional to an F-distribution.

■

One usually collects the calculations in an analysis of variance table.

Variation	SS	Degrees of freedom = dimension
Of model from hypothesis	$\ p_M(\mathbf{Y}) - p_H(\mathbf{Y})\ ^2$	$k - r$
Of observations from model	$\ \mathbf{Y} - p_M(\mathbf{Y})\ ^2$	$n - k$
Of observations from hypothesis	$\ \mathbf{Y} - p_H(\mathbf{Y})\ ^2$	$n - r$

||| Remark 2.21

Often one will be in the situation that the sub-spaces M and H are parameterised, i.e.

$$\begin{aligned}\boldsymbol{\mu} \in M &\Leftrightarrow \exists \boldsymbol{\theta} \in R^k (\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\theta}) \\ \boldsymbol{\mu} \in H &\Leftrightarrow \exists \boldsymbol{\gamma} \in R^r (\boldsymbol{\mu} = \mathbf{x}_0\boldsymbol{\gamma}),\end{aligned}$$

where \mathbf{x} and \mathbf{x}_0 are $n \times k$ respectively $n \times r$ (with $r \leq k$) matrices. We then have that $p_M(\mathbf{y}) = \mathbf{x}\hat{\boldsymbol{\theta}}$ and $p_H(\mathbf{y}) = \mathbf{x}_0\hat{\boldsymbol{\gamma}}$ are computed by solving the equations

$$\begin{aligned}(\mathbf{x}'\Sigma^{-1}\mathbf{x})\hat{\boldsymbol{\theta}} &= \mathbf{x}'\Sigma^{-1}\mathbf{y} \\ (\mathbf{x}'_0\Sigma^{-1}\mathbf{x}_0)\hat{\boldsymbol{\gamma}} &= \mathbf{x}'_0\Sigma^{-1}\mathbf{y}\end{aligned}$$

with respect to $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$.

We now state a useful theorem used in residual analysis in the general linear model.

|||| **Theorem 2.22**

Let \mathbf{x} be an $n \times k$ matrix not necessarily of full rank. Then the socalled *hat matrix*

$$\mathbf{H} = \mathbf{x}(\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'$$

is independent of the choice of the generalised inverse $(\mathbf{x}'\mathbf{x})^{-}$. Furthermore it is idempotent and symmetric, and

$$\text{rk}(\mathbf{H}) = \text{tr}(\mathbf{H}) = \text{rk}(\mathbf{x})$$

|||| **Proof**

See Graybill (1976, p. 32). If $(\mathbf{x}'\mathbf{x})$ has full rank we of course use the inverse, and (the last part of) the theorem is seen immediately.

\mathbf{H} corresponds to projection on the column space of \mathbf{x} , and it is easily seen that the matrix

$$\mathbf{M} = \mathbf{I} - \mathbf{H} = \mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'$$

projects on the orthogonal complement, i.e.

$$\mathbf{MH} = \mathbf{0}$$

■

|||| **Remark 2.23**

Using Pythagoras' theorem we see that there are two other ways of computing

$$\|p_M(\mathbf{Y}) - p_H(\mathbf{Y})\|^2 = (\mathbf{x}\hat{\boldsymbol{\theta}} - \mathbf{x}_0\hat{\boldsymbol{\gamma}})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}\hat{\boldsymbol{\theta}} - \mathbf{x}_0\hat{\boldsymbol{\gamma}}) \quad (2-2)$$

besides the direct formula, namely as

$$\|p_M(\mathbf{Y})\|^2 - \|p_H(\mathbf{Y})\|^2 = (\mathbf{x}\hat{\boldsymbol{\theta}})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}\hat{\boldsymbol{\theta}}) - (\mathbf{x}_0\hat{\boldsymbol{\gamma}})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_0\hat{\boldsymbol{\gamma}}) \quad (2-3)$$

or as

$$\begin{aligned} & \| \mathbf{Y} - p_H(\mathbf{Y}) \|^2 - \| \mathbf{Y} - p_M(\mathbf{Y}) \|^2 \\ &= (\mathbf{Y} - \mathbf{x}_0\hat{\mathbf{Y}})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x}_0\hat{\mathbf{Y}}) - (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})' \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) \end{aligned} \quad (2-4)$$

For numerical reasons (2-3) is normally preferred, but if one has computed the residual sum of squares using (2-4) is straight forward.

|||| Remark 2.24

Output from statistical standard software will often be organised slightly different from what is presented above. We assume that $\Sigma = \mathbf{I}$ so that the norm we are considering is given by

$$\|\mathbf{z}\|^2 = \mathbf{z}'\mathbf{z} = \sum_{i=1}^n z_i^2$$

The output from e.g. SAS using the General Linear Model procedure GLM will then include a Analysis of Variance table (ANOVA table) as

Source of variation	Sum of Squares	Degrees of Freedom
Model	SS(Model)	$\text{rk}(\mathbf{x})$
Error	SSRes(Model)	$n - \text{rk}(\mathbf{x})$
Uncorrected Total	SSTot(Uncorrected)	n

Here

$$\text{SS(Model)} = \|p_M(\mathbf{Y})\|^2 = (\mathbf{x}\hat{\theta})'(\mathbf{x}\hat{\theta}) = \mathbf{Y}'\mathbf{H}\mathbf{Y}$$

$$\text{SSRes(Model)} = \|\mathbf{Y} - p_M(\mathbf{Y})\|^2 = (\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta}) = \mathbf{Y}'\mathbf{M}\mathbf{Y}$$

$$\text{SSTot(Uncorrected)} = \|\mathbf{Y}\|^2 = \mathbf{Y}'\mathbf{Y} = \sum_{i=1}^n Y_i^2$$

If we want to test a hypothesis then we may obtain the necessary sums of squares by applying the GLM procedure on the model and on the hypothesis and then compute the denominator in the test statistic (2-1) using one of the formulas (2-2), (2-3) or (2-4).

|||| Explanation 2.25

We now consider general linear models with an *intercept* α , i.e. models of the form

$$Y_i = \alpha + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon \quad i = 1, \dots, n$$

or

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

We still use the compact matrix terminology

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

where \mathbf{x} now is $n \times (k+1)$ and $\boldsymbol{\theta}$ is $(k+1) \times 1$. We also assume that $D(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$.

Many systems for statistical computing will automatically add a column of 1's to the design matrix unless one directly specifies that this should not be done. In the SAS procedure GLM a model statement

model $y = x1\ x2;$

will thus be interpreted as

$$y = \alpha + \beta_1 x1 + \beta_2 x2 + \boldsymbol{\varepsilon}.$$

If we want to avoid the intercept term α we must write

model $y = x1\ x2 /noint;$

In the intercept case the output from the SAS GLM procedure includes an ANOVA table

Source of variation	Sum of Squares	Degrees of Freedom
Model	SS(Model)	$\text{rk}(\mathbf{x}) - 1$
Error	SSRes(Model)	$n - \text{rk}(\mathbf{x})$
Corrected Total	SSTot(Corrected)	$n - 1$

Here

$$\text{SS(Model)} = (\mathbf{x}\hat{\boldsymbol{\theta}} - \bar{\mathbf{Y}}\mathbf{1})'(\mathbf{x}\hat{\boldsymbol{\theta}} - \bar{\mathbf{Y}}\mathbf{1}) = \mathbf{Y}'\mathbf{H}\mathbf{Y} - n\bar{Y}^2$$

$$\text{SSRes(Model)} = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = \mathbf{Y}'\mathbf{M}\mathbf{Y}$$

$$\text{SSTot(Corrected)} = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2 = \sum_{i=1}^h (Y_i - \bar{Y})^2$$

Also in this case we may test a hypothesis by applying the GLM procedure on the model and on the hypothesis and then compute the necessary sums of squares using formulas (2-2), (2-3) or (2-4).

If we in the above case compute

$$F = \frac{\text{SS}(\text{Model}) / (\text{rk}(\mathbf{x}) - 1)}{\text{SSRes}(\text{Model}) / (n - \text{rk}(\mathbf{x}))}$$

this will be the test statistic for the hypothesis that all parameters *except* the intercept are zero, i.e.

$$H_0 : \beta_1 = \dots = \beta_k = 0$$

against all alternatives. The critical values are given by

$$C = \{Y | F > F(\text{rk}(\mathbf{x}) - 1, n - \text{rk}(\mathbf{x}))_{1-\alpha}\}$$

when testing at significance level α .

Once again we consider the model from Example 2.8 (p. 81).

||| Example 2.26

We have the model

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \epsilon.$$

We observe data where $y' = (90, 30, 75)$. We wish to test the hypothesis

$$H_0 : \theta_2 = 0 \quad \text{versus} \quad H_1 : \theta_2 \neq 0.$$

We reformulate the hypothesis into

$$H_0 : E(\mathbf{Y}) = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \theta_1 = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \gamma.$$

The estimator for γ is

$$\hat{\gamma} = [(1 \ 0 \ \frac{1}{2}) \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix}]^{-1} [(1 \ 0 \ \frac{1}{2}) \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}] = \frac{4}{5}y_1 + \frac{2}{5}y_3.$$

The observed value is $\hat{\gamma} = 102$. From this we have

$$\mathbf{x}_0 \hat{\gamma} = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} 102 = \begin{bmatrix} 102 \\ 0 \\ 51 \end{bmatrix},$$

we see that

$$\mathbf{x}\hat{\theta} - \mathbf{x}_0\hat{\gamma} = \begin{bmatrix} -7 \\ 35 \\ 14 \end{bmatrix}$$

and thus

$$\|\mathbf{x}\hat{\theta} - \mathbf{x}_0\hat{\gamma}\|^2 = 49 + 1225 + 196 = 1470.$$

Now, since

$$\mathbf{y} - \mathbf{x}_0\hat{\gamma} = \begin{bmatrix} -12 \\ 30 \\ 24 \end{bmatrix},$$

and

$$\|\mathbf{y} - \mathbf{x}_0\hat{\gamma}\|^2 = (\mathbf{y} - \mathbf{x}_0\hat{\gamma})'(\mathbf{y} - \mathbf{x}_0\hat{\gamma}) = 1620.$$

and since we had (p. 82)

$$\|\mathbf{y} - \mathbf{x}\hat{\theta}\|^2 = (\mathbf{y} - \mathbf{x}\hat{\theta})'(\mathbf{y} - \mathbf{x}\hat{\theta}) = 150,$$

we get

$$\|\mathbf{x}\hat{\theta} - \mathbf{x}_0\hat{\gamma}\|^2 = 1620 - 150 = 1470.$$

We may also compute this quantity as

$$\begin{aligned} \|\mathbf{x}\hat{\theta}\|^2 - \|\mathbf{x}_0\hat{\gamma}\|^2 &= (\mathbf{x}\hat{\theta})'(\mathbf{x}\hat{\theta}) - (\mathbf{x}_0\hat{\gamma})'(\mathbf{x}_0\hat{\gamma}) \\ &= 14475 - 13005 \\ &= 1470. \end{aligned}$$

From this the test statistic becomes

$$\frac{\|\mathbf{x}\hat{\theta} - \mathbf{x}_0\hat{\gamma}\|^2 / (2 - 1)}{\|\mathbf{y} - \mathbf{x}\hat{\theta}\|^2 / (3 - 2)} = \frac{1470}{150} = 9.8 \sim F(1, 1)_{0.80},$$

and we accept the hypothesis at least for any $\alpha < 20\%$.

Explanation of the degrees of freedom:

$$\begin{aligned} \text{rk } \mathbf{x} &= \text{rk} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = 2 = k \\ \text{rk } \mathbf{x}_0 &= \text{rk} \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} = 1 = r \\ n &= 3. \end{aligned}$$

We will now look at the continuation of example 2.10 p. 88.

||| Example 2.27

From the formulation of the problem it seems reasonable to assume that the parameter $\theta_{21} = 0$. We will therefore test the hypothesis

$$H_0 : \theta_{21} = 0 \quad \text{against} \quad H_1 : \theta_{21} \neq 0.$$

The hypothesis-space H is therefore given by

$$E(\mathbf{Y}) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \end{bmatrix} = \begin{bmatrix} \mu_1 + \theta_{11} \\ \mu_1 + \theta_{11} \\ \mu_1 + \theta_{12} \\ \mu_1 + \theta_{12} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We now find

$$\mathbf{x}'_1 \mathbf{x}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix},$$

and

$$\mathbf{x}'_1 \mathbf{Y} = \begin{bmatrix} Y_{1..} \\ Y_{11..} \\ Y_{12..} \end{bmatrix}.$$

We see that $\mathbf{x}'_1 \mathbf{x}_1$ is singular, and we add the linear restriction

$$\mathbf{b}' \boldsymbol{\theta} = (0 \ 1 \ 1) \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \end{bmatrix} = \theta_{11} + \theta_{12} = 0.$$

Since

$$\mathbf{b}'\mathbf{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

we have

$$\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

This matrix is inverted on p. 87. We therefore find the estimator under H_0 as

$$\hat{\theta}_1 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} Y_{1..} \\ Y_{11..} \\ Y_{12..} \end{bmatrix} = \begin{bmatrix} \bar{Y}_{1..} \\ \bar{Y}_{11..} - \bar{Y}_1 \\ \bar{Y}_{12..} - \bar{Y}_1 \end{bmatrix}.$$

The observed value is $(-9, +8, -8)'$. The new residual vector is

$$\mathbf{y} - \mathbf{x}_1 \hat{\theta}_1 = (1, -1, -2, +2, -6, 0, -2)'.$$

The norm of this vector is 50, and the number of degrees of freedom is $7-2=5$. We therefore find that

$$\begin{aligned} \|p_M(\mathbf{y}) - p_H(\mathbf{y})\|^2 &= \|\mathbf{y} - p_H(\mathbf{y})\|^2 - \|\mathbf{y} - p_M(\mathbf{y})\|^2 \\ &= 50 - 28\frac{2}{3} = 21\frac{1}{3}. \end{aligned}$$

We now collect the calculations in the following analysis of variance table.

Variation	SS	f	S^2	Test
$M - H$	$21\frac{1}{3}$	$3 - 2 = 1$	$21\frac{1}{3}$	2.97
$O - M$	$28\frac{2}{3}$	$7 - 3 = 4$	$7\frac{1}{6}$	
$O - H$	50	$7 - 2 = 5$		

Since the observed value of the test statistic $2.97 < F(1,4)_{0.90}$ we will accept the hypothesis, and therefore assume that H_0 is true.

2.2.2 Successive testing in the general linear model.

In this section we will illustrate the test procedure one should follow, when one successively wants to investigate if the mean vector for ones observations lies in sub-spaces H_i with

$$H_0 \supseteq H_1 \supseteq H_2 \supseteq \cdots \supseteq H_m, \quad m \leq k.$$

We will pursue that in the following example.

We will start by considering the following numbers from the yield of penicillin fermentation using two different types of sugar namely: lactose and cane sugar, at the concentrations 2%, 4%, 6% and 8% (in g./100 ml.).

		Factor B: concentration			
		2%	4%	6%	8%
Factor A:	Lactose	0.606	0.660	0.984	0.908
	Cane sugar	0.761	0.933	1.072	0.979

The numbers are from [6] p. 314. The yield has been expressed by the logarithm of the weight of the mycelium after one week of growth.

We are now interested in investigating the two factors A's and B's influence on the yield. We assume that the observations are stochastic independent and normally distributed. They are called

$$L : Y_{11}, Y_{12}, Y_{13}, Y_{14}$$

and

$$R : Y_{21}, Y_{22}, Y_{23}, Y_{24}$$

further we will assume that

$$E(Y_{ij}) = \alpha'_i + \beta'_i x'_j + \gamma'_i x_j^2$$

where x'_j gives the j 'th sugar concentration. We will perform change in scale of the sugar concentration

$$\begin{array}{ll} 2\% & -3 \\ 4\% & -1 \\ 6\% & 1 \\ 8\% & 3, \end{array}$$

or more stringently define x by

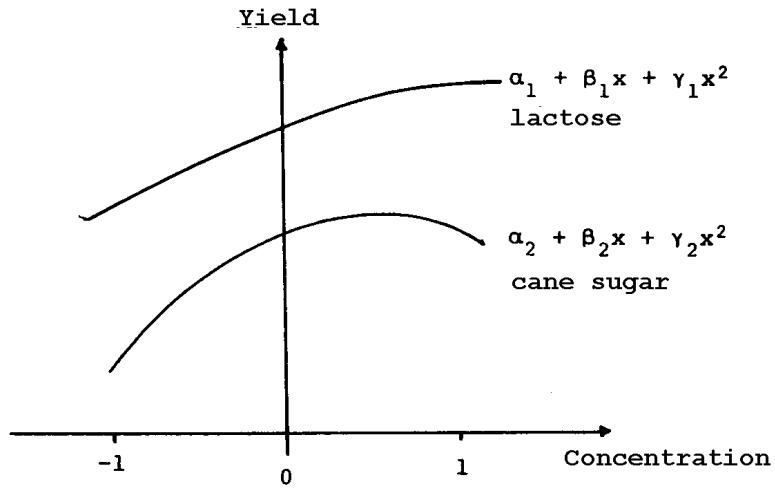
$$x_j = \frac{x'_j - 5\%}{1\%}.$$

We then get the following expression for the mean values

$$E(Y_{ij}) = \alpha_i + \beta_i x_j + \gamma_i x_j^2.$$

We are assuming that the yield within the given limits can be expressed as polynomials of second degree.

One could now e.g. successively investigate



- 1) if $\gamma_1 = \gamma_2 = 0$, i.e. if a description by affine functions is sufficient
- 2) if that is accepted then if $\beta_1 = \beta_2 = \beta$, i.e. if the marginal effect by increasing the concentration is the same for the two types of sugar
- 3) if that is accepted then if $\alpha_1 = \alpha_2 = \alpha$, i.e. if the two types of sugar are equal with respect to the yield and if this is accepted
- 4) then if $\beta = 0$, i.e. if the concentration has any influence at all

i) We first write the model in matrix form

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \\ \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix},$$

or

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}.$$

We find

$$\begin{aligned} \mathbf{x}'\mathbf{x} &= \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ -3 & -1 & 1 & 3 & 0 & 0 & 0 & 0 \\ 9 & 1 & 1 & 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -3 & -1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 9 & 1 & 1 & 9 \end{bmatrix} \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 & 9 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 0 & 20 & 0 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 & 0 & 0 & 0 \\ 20 & 0 & 164 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 20 & 0 \\ 0 & 0 & 0 & 0 & 20 & 0 & 0 \\ 0 & 0 & 0 & 20 & 0 & 164 & 0 \end{bmatrix}. \end{aligned}$$

Since

$$\begin{bmatrix} 4 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{41}{64} & 0 & -\frac{5}{64} \\ 0 & \frac{1}{20} & 0 \\ -\frac{5}{64} & 0 & \frac{1}{64} \end{bmatrix},$$

then

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} \frac{41}{64} & 0 & -\frac{5}{64} & 0 & 0 & 0 \\ 0 & \frac{1}{20} & 0 & 0 & 0 & 0 \\ -\frac{5}{64} & 0 & \frac{1}{64} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{41}{64} & 0 & -\frac{5}{64} \\ 0 & 0 & 0 & 0 & \frac{1}{20} & 0 \\ 0 & 0 & 0 & -\frac{5}{64} & 0 & \frac{1}{64} \end{bmatrix}.$$

From this we see that

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} -\frac{1}{16}y_{11} + \frac{9}{16}y_{12} + \frac{9}{16}y_{13} - \frac{1}{16}y_{14} \\ -\frac{3}{20}y_{11} - \frac{1}{20}y_{12} + \frac{1}{20}y_{13} + \frac{3}{20}y_{14} \\ \frac{1}{16}y_{11} - \frac{1}{16}y_{12} - \frac{1}{16}y_{13} + \frac{1}{16}y_{14} \\ -\frac{1}{16}y_{21} + \frac{9}{16}y_{22} + \frac{9}{16}y_{23} - \frac{1}{16}y_{24} \\ \frac{3}{20}y_{21} - \frac{1}{20}y_{22} - \frac{1}{20}y_{23} + \frac{3}{20}y_{24} \\ \frac{1}{16}y_{21} - \frac{1}{16}y_{22} - \frac{1}{16}y_{23} + \frac{1}{16}y_{24} \end{bmatrix} = \begin{bmatrix} 0.830 \\ 0.062 \\ -0.008 \\ 1.019 \\ 0.040 \\ -0.017 \end{bmatrix}.$$

The model corresponds to a 6-dimensional sub-space M in \mathbb{R}^8 ($\text{rk}\mathbf{x} = 6$), and since we are using the norm corresponding to \mathbf{I} we have that the

projection onto M is

$$p_M(\mathbf{y}) = \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} 0.830 \\ 0.062 \\ -0.008 \\ 1.019 \\ 0.040 \\ -0.017 \end{bmatrix} = \begin{bmatrix} 0.572 \\ 0.760 \\ 0.884 \\ 0.944 \\ 0.746 \\ 0.962 \\ 1.042 \\ 0.986 \end{bmatrix}.$$

We therefore have the residuals

$$\mathbf{y} - p_M(\mathbf{y}) = \begin{bmatrix} 0.034 \\ -0.100 \\ 0.100 \\ -0.036 \\ 0.015 \\ -0.029 \\ 0.030 \\ -0.007 \end{bmatrix}.$$

The squared length of this vector is

$$\|\mathbf{y} - p_M(\mathbf{y})\|^2 = 0.034^2 + \dots + (-0.007)^2 = 0.024467.$$

As an estimate of σ^2 we can therefore use

$$\hat{\sigma}^2 = \frac{1}{8-6} 0.024467 = 0.0122335.$$

ii) If the hypothesis $\boldsymbol{\mu} \in H_1$, i.e. $\gamma_1 = \gamma_2 = 0$, or

$$\mathbf{y} = \begin{bmatrix} 1 & -3 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{x}_1 \boldsymbol{\delta}_1 + \boldsymbol{\varepsilon}_1,$$

is true, then we get the estimates

$$\hat{\delta}_1 = (\mathbf{x}'_1 \mathbf{x}_1)^{-1} \mathbf{x}'_1 \mathbf{y} = \begin{bmatrix} \frac{1}{4}y_{11} + \frac{1}{4}y_{12} + \frac{1}{4}y_{13} + \frac{1}{4}y_{14} \\ -\frac{3}{20}y_{11} - \frac{1}{20}y_{12} + \frac{1}{20}y_{13} + \frac{3}{20}y_{14} \\ \frac{1}{4}y_{21} + \frac{1}{4}y_{22} + \frac{1}{4}y_{23} + \frac{1}{4}y_{24} \\ -\frac{3}{20}y_{21} - \frac{1}{20}y_{22} + \frac{1}{20}y_{23} + \frac{3}{20}y_{24} \end{bmatrix} = \begin{bmatrix} 0.790 \\ 0.062 \\ 0.936 \\ 0.040 \end{bmatrix}$$

The residuals are

$$\mathbf{y} - p_{H_1}(\mathbf{y}) = \mathbf{y} - \mathbf{x}_1 \hat{\boldsymbol{\delta}}_1 = \begin{bmatrix} 0.002 \\ -0.068 \\ 0.132 \\ -0.068 \\ -0.055 \\ 0.037 \\ 0.096 \\ -0.077 \end{bmatrix}.$$

The squared length of this vector is

$$\|\mathbf{y} - p_{H_1}(\mathbf{y})\|^2 = 0.002^2 + \dots + (-0.077)^2 = 0.046215.$$

iii) If $\boldsymbol{\mu} \in H_2$, d.v.s. $\beta_1 = \beta_2 = \beta$, the model becomes

$$\mathbf{y} = \begin{bmatrix} 1 & 0 & -3 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & 3 \\ 0 & 1 & -3 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \boldsymbol{\varepsilon}_2 = \mathbf{x}_2 \hat{\boldsymbol{\delta}}_2 + \boldsymbol{\varepsilon}_2.$$

The estimates become

$$\hat{\boldsymbol{\delta}}_2 = (\mathbf{x}'_2 \mathbf{x}_2)^{-1} \mathbf{x}'_2 \mathbf{y} = \begin{bmatrix} 0.790 \\ 0.936 \\ 0.051 \end{bmatrix},$$

and the residuals

$$\mathbf{y} - p_{H_2}(\mathbf{y}) = \begin{bmatrix} -0.031 \\ -0.079 \\ 0.143 \\ -0.035 \\ -0.022 \\ 0.048 \\ 0.085 \\ -0.110 \end{bmatrix}.$$

The squared norm of the residual vector is

$$\|\mathbf{y} - p_{H_2}(\mathbf{y})\|^2 = (-0.031)^2 + \dots + (-0.110)^2 = 0.050989.$$

iv) If $\mu \in H_3$, i.e. $\beta_1 = \beta_2 = \beta$ and $\alpha_1 = \alpha_2 = \alpha$, then the model is

$$\mathbf{y} = \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \\ 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \varepsilon_3 = \mathbf{x}_3 \boldsymbol{\delta}_3 + \varepsilon_3$$

We find

$$\hat{\boldsymbol{\delta}}_3 = (\mathbf{x}'_3 \mathbf{x}_3)^{-1} \mathbf{x}'_3 \mathbf{y} = \begin{bmatrix} 0.863 \\ 0.051 \end{bmatrix},$$

and

$$\mathbf{y} - p_{H_3}(\mathbf{y}) = \begin{bmatrix} -0.104 \\ -0.152 \\ 0.070 \\ -0.108 \\ 0.051 \\ 0.121 \\ 0.158 \\ -0.037 \end{bmatrix},$$

giving

$$\|\mathbf{y} - p_{H_3}(\mathbf{y})\|^2 = 0.094059.$$

v) Finally we consider the case $\mu \in H_4$, i.e. $\beta = 0$, or

$$\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \alpha = \mathbf{x}_4 \boldsymbol{\delta}_4 + \varepsilon_4.$$

We find

$$\hat{\boldsymbol{\delta}}_4 = \hat{\alpha} = (\mathbf{x}'_4 \mathbf{x}_4)^{-1} \mathbf{x}'_4 \mathbf{y}' = 0.863,$$

giving

$$\mathbf{y} - p_{H_4}(\mathbf{y}) = \begin{bmatrix} -0.250 \\ -0.203 \\ 0.121 \\ 0.045 \\ -0.102 \\ 0.070 \\ 0.209 \\ 0.116 \end{bmatrix},$$

and

$$\|\mathbf{y} - p_{H_4}(\mathbf{y})\|^2 = 0.196365.$$

Since we let $\text{rk}(\mathbf{x}_i) = r_i$ and $\text{rk}(\mathbf{x}) = k$ we can summarise the testing procedure in an analysis of variance table such as

Variation	SS	Degrees of freedom = dimension
$H_4 - H_3$	$\ p_{H_4}(\mathbf{y}) - p_{H_3}(\mathbf{y})\ ^2$	$r_3 - r_4 = 2 - 1 = 1$
$H_3 - H_2$	$\ p_{H_3}(\mathbf{y}) - p_{H_2}(\mathbf{y})\ ^2$	$r_2 - r_3 = 3 - 2 = 1$
$H_2 - H_1$	$\ p_{H_2}(\mathbf{y}) - p_{H_1}(\mathbf{y})\ ^2$	$r_1 - r_2 = 4 - 3 = 1$
$H_1 - M$	$\ p_{H_1}(\mathbf{y}) - p_M(\mathbf{y})\ ^2$	$k - r_1 = 6 - 4 = 2$
$M - \text{obs.}$	$\ p_M(\mathbf{y}) - \mathbf{y}\ ^2$	$n - k = 8 - 6 = 2$
$H_4 - \text{obs.}$	$\ p_{H_4}(\mathbf{y}) - \mathbf{y}\ ^2$	$n - r_4 = 8 - 1 = 7$

This table is a simple extension of the table on p. 105. We can use the partition theorem and get, under the different hypotheses, that the sum of squares are independent and distributed as $\sigma^2 \chi^2$ with the respective degrees of freedom.

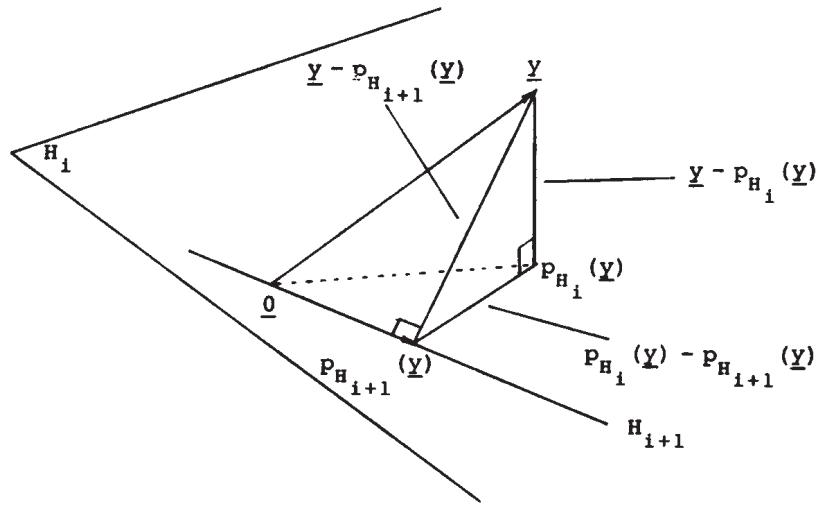
If a hypothesis H_i is accepted then the test statistic for the test of H_{i+1} becomes

$$\frac{\|p_{H_i}(\mathbf{y}) - p_{H_{i+1}}(\mathbf{y})\|^2 / (r_i - r_{i+1})}{\|p_{H_i}(\mathbf{y}) - \mathbf{y}\|^2 / (n - r_i)}.$$

Under the hypothesis this measure is $F(r_i - r_{i+1}, n - r_i)$ distributed (according to the partition theorem) and - still following the theory from the previous section - we reject for large values of Z i.e. for

$$Z > F(r_i - r_{i+1}, n - r_i)_{1-\alpha}.$$

Before we start testing it would be appropriate to give some computational formulas. We consider the transition from H_i to $H_{i+1} \subset H_i$.



Using Pythagoras' theorem we now see - c.f. Remark 2.23 - that there are two alternative ways of computation for

$$z = \|p_{H_{i+1}}(\mathbf{y}) - p_{H_i}(\mathbf{y})\|^2,$$

they are

$$z = \|p_{H_i}(\mathbf{y})\|^2 - \|p_{H_{i+1}}(\mathbf{y})\|^2 \quad (2-5)$$

and

$$z = \|\mathbf{y} - p_{H_{i+1}}(\mathbf{y})\|^2 - \|\mathbf{y} - p_{H_i}(\mathbf{y})\|^2. \quad (2-6)$$

Of these the first must be preferred from numerical reasons but if one has computed the residuals sum of squares anyhow it seems to be easier to use (2-6).

The analysis of variance table in our case becomes

Variation	SS	f	Test statistic
$H_4 - H_3$	0.102306	1	$\frac{0.102306/1}{0.094059/6} = 5.44$
$H_3 - H_2$	0.043070	1	$\frac{0.043070/1}{0.050981/5} = 4.22$
$H_2 - H_1$	0.004774	1	$\frac{0.004774/1}{0.046215/4} = 0.41$
$H_1 - M$	0.021748	2	$\frac{0.021748/2}{0.024467/2} = 0.89$
$M - \text{obs}$	0.024467	2	
$\text{Obs} - H_4$	0.196365	7	

Since

$$4.22 \simeq F(1, 5)_{0.91},$$

and

$$5.44 \simeq F(1, 6)_{0.94},$$

we will not by testing at say, level $\alpha = 5\%$ - reject any of the hypothesis H_1, H_2, H_3 or H_4 .



We will of course not test e.g. H_2 , if we had rejected H_1 , since H_2 is a sub-hypothesis of H_1 .

The conclusion is therefore that we (until new investigations reject this) will continue to work with the model that the yield Y by penicillin fermentation is independent of type of sugar and the concentration ($2\% \leq \text{concentration} \leq 8\%$) at which the fermentation takes place. We have with

$$E(Y) = \alpha \quad \text{and} \quad V(Y) = \sigma^2,$$

that

$$\hat{\alpha} = 0.863,$$

and

$$\hat{\sigma}^2 = \frac{0.196365}{7} = 0.028052 \simeq 0.17^2.$$

Finally

$$V(\hat{\alpha}) = \frac{\sigma^2}{8} \simeq \frac{\hat{\sigma}^2}{8} = 0.0035 \simeq 0.059^2.$$

|||| Chapter 3

Regression analysis

In this chapter we will give an overview on regression analysis. Most of it is a special case of the general linear model but since a number of uses are often concerned with regression situations we will try to describe the results in this language.

There is a small section on orthogonal regression (not to be confused with regression by orthogonal polynomials). From a statistical point of view this is more related to the section on principle components and factor analysis, and considering ways of computation we also refer to that chapter. However, from a curve-fitting point of view we have found it sensible to mention the concept in the present chapter too.

3.1 Linear regression analysis

In this section linear regression analysis will be analysed by means of the theory for the general linear model. We start with

3.1.1 Notation and model.

In the ordinary regression analysis we simply work with a general linear model with an intercept, i.e. we work with the model

$$E(Y) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k,$$

where the x 's are known variables and the β 's (and α) are unknown parameters. If we have given n observations of Y we could more precisely write the model

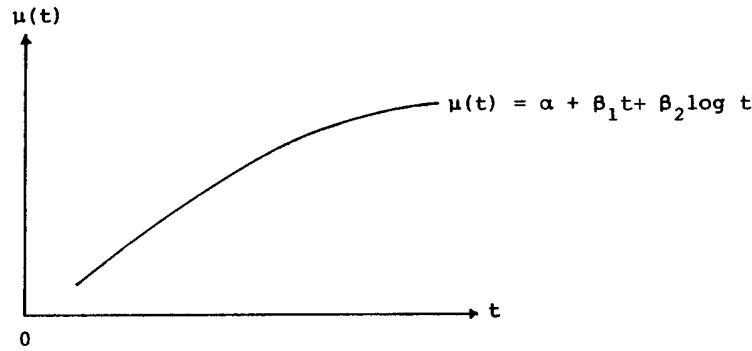


Figure 3.1:

as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

or

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

We assume as usual that

$$\mathbf{D}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is known and σ^2 is (usually) unknown.

The estimators are found in the usual way by solving the normal equations

$$\mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{x} \boldsymbol{\beta} = \mathbf{x}' \boldsymbol{\Sigma}^{-1} \mathbf{Y},$$

or if $\boldsymbol{\Sigma} = \mathbf{I}$

$$\mathbf{x}' \mathbf{x} \hat{\boldsymbol{\beta}} = \mathbf{x}' \mathbf{Y}.$$

In the first case we talk of a *weighted regression analysis*.

Before we go on it is probably appropriate once again to stress what is meant by the word linear in the term linear regression analysis.

As in the ordinary general linear model the meaning is that we have *linearity in the parameters*. We can easily do regression by e.g. time and the logarithm of the time. The model will then just be

$$\mathbb{E}(Y) = \alpha + \beta_1 t + \beta_2 \ln t,$$

cf. example 2.2 or fig. 3.1. With n observations this model in matrix form becomes

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \ln t_1 \\ \vdots & \vdots & \vdots \\ 1 & t_n & \ln t_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Another banality that could be useful to stress is that one can force the regression surface through 0 by deleting the α and first column in the \mathbf{x} -matrix i.e. use the model

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & & \vdots \\ x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

It can be useful to note that you can use the following trick if you wish the regression surface to go through 0. We assume that $\Sigma = \mathbf{I}$.

We consider the observations Y_1, \dots, Y_n and the corresponding values of the independent variables $1, x_{i1}, \dots, x_{ik}, i = 1, \dots, n$. If we add $-Y_1, \dots, -Y_n$ and $1, -x_{i1}, \dots, -x_{ik}, i = 1, \dots, n$ and write down the usual model we get

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \\ -Y_1 \\ \vdots \\ -Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \\ 1 & -x_{11} & \cdots & -x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & -x_{1n} & \cdots & -x_{kn} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \boldsymbol{\varepsilon},$$

or more compactly

$$\begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 \\ \mathbf{1} & -\mathbf{x}_1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \boldsymbol{\varepsilon},$$

The normal equations become

$$\begin{bmatrix} \mathbf{1}' & \mathbf{1}' \\ \mathbf{x}_1' & -\mathbf{x}_1' \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 \\ \mathbf{1} & -\mathbf{x}_1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{1}' & \mathbf{1}' \\ \mathbf{x}_1' & -\mathbf{x}_1' \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix},$$

or

$$\begin{bmatrix} 2n & 0 \\ 0 & 2\mathbf{x}_1' \mathbf{x}_1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 2\mathbf{x}_1' \mathbf{Y} \end{bmatrix}.$$

If we write out the equations we get

$$\begin{aligned} 2n\alpha &= 0 \\ 2\mathbf{x}_1' \mathbf{x}_1 \beta &= 2\mathbf{x}_1' \mathbf{Y}, \end{aligned}$$

or

$$\begin{aligned}\alpha &= 0 \\ \mathbf{x}'_1 \boldsymbol{\beta} &= \mathbf{x}'_1 \mathbf{Y}.\end{aligned}$$

In other words in this way we have found the estimators of the coefficients to a regression surface which has been forced through $\mathbf{0}$.

The reason why the above is useful is that a number of standard programmes cannot force the surface through $\mathbf{0}$. Using the above mentioned trick the problem can be circumvented.

The output from such a programme should be interpreted cautiously since all the sums of squares are twice their correct size. E.g. the residual sums of squares will be computed as

$$\begin{aligned}&\left(\begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{x}_1 \hat{\boldsymbol{\beta}} \\ -\mathbf{x}_1 \hat{\boldsymbol{\beta}} \end{bmatrix} \right)' \left(\begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{x}_1 \hat{\boldsymbol{\beta}} \\ -\mathbf{x}_1 \hat{\boldsymbol{\beta}} \end{bmatrix} \right) \\ &= ([\mathbf{Y} - \mathbf{x}_1 \hat{\boldsymbol{\beta}}]', [-\mathbf{Y} + \mathbf{x}_1 \hat{\boldsymbol{\beta}}]') \begin{bmatrix} \mathbf{Y} - \mathbf{x}_1 \hat{\boldsymbol{\beta}} \\ -\mathbf{Y} + \mathbf{x}_1 \hat{\boldsymbol{\beta}} \end{bmatrix} \\ &= 2[\mathbf{Y} - \mathbf{x}_1 \hat{\boldsymbol{\beta}}]'[\mathbf{Y} - \mathbf{x}_1 \hat{\boldsymbol{\beta}}],\end{aligned}$$

i.e. twice the correct residual sum of squares. The mentioned degrees of freedom will not be correct either. We have to write up the ordinary linear model and find the correct degrees of freedom by considering the dimensions.

3.1.2 Correlation and regression.

In theorem 1.43 section 39 a result was stated, which can be used for a test if the multiple correlation coefficient between normally distributed variables is 0. We will now show that this result corresponds to a certain test in a regression model.

We will assume that we have the usual model p. 123 and we assume that $\boldsymbol{\Sigma} = \mathbf{I}$.

Without any problems we can use the theory from chapter 2 to test different hypothesis about the parameters $\alpha, \beta_1, \dots, \beta_k$.

By formal calculations we can estimate the multiple correlation coefficient between \mathbf{Y} and $\mathbf{x}_1, \dots, \mathbf{x}_k$ using expressions mentioned in section 1.3.2.

It can be shown that we get

$$R^2 = \frac{\|\mathbf{Y} - p_0(\mathbf{Y})\|^2 - \|\mathbf{Y} - p_M(\mathbf{Y})\|^2}{\|\mathbf{Y} - p_0(\mathbf{Y})\|^2},$$

where

$$p_0(\mathbf{Y}) = \begin{bmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix} \quad (= \mathbf{x}_0 \cdot \hat{\boldsymbol{\gamma}}),$$

and

$$p_M(\mathbf{Y}) = \mathbf{x} \hat{\boldsymbol{\beta}} = \hat{\mathbf{e}}(\mathbf{Y}).$$

These results are not very surprising. We remember that the multiple correlation coefficient could be found as the linear combination of \mathbf{X} which minimises the variance of $(\mathbf{Y} - \boldsymbol{\alpha}' \mathbf{X})$ and this corresponds exactly to writing the condition for least squares estimates.

If we as in Remark 2.25 let

$$\text{SSTot} = \text{SSTot(Corrected)} = \|\mathbf{Y} - p_0(\mathbf{Y})\|^2 = \sum_i (Y_i - \bar{Y})^2,$$

and

$$\text{SSRes} = \text{SSRes(Model)} = \|\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\beta}}\|^2 = \sum_i (Y_i - \hat{e}(Y_i))^2,$$

we can write

$$R^2 = \frac{\text{SSTot} - \text{SSRes}}{\text{SSTot}},$$

i.e. the squared multiple correlation coefficient can also be expressed as the part of the total variation in the Y 's which are explained using the independent variables.

SAS also compute an R-square adjusted for the degrees of freedom. The *adjusted R-square* is calculated as

$$R_{\text{adj}}^2 = 1 - \frac{(n - i)(1 - R^2)}{n - p}$$

where i is equal to 1 if there is an intercept and 0 otherwise. n is the number of observations used to fit the model, and p is the number of parameters in the model including a possible intercept.

A corresponding re-interpretation of the partial correlations is of course also possible.

Furthermore, we see that if we formally write the test on p. 41 for $\rho_{Y|x_1, \dots, x_k} = 0$

we get - assuming that $\text{rk } \mathbf{x} = k + 1$ -

$$\begin{aligned}\frac{R^2}{1-R^2} \frac{n-k-1}{k} &= \frac{\|\mathbf{Y} - p_0(\mathbf{Y})\|^2 - \|\mathbf{Y} - p_M(\mathbf{Y})\|^2}{\|\mathbf{Y} - p_M(\mathbf{Y})\|^2} \frac{n-k-1}{k} \\ &= \frac{\|p_M(\mathbf{Y}) - p_0(\mathbf{Y})\|^2/k}{\|\mathbf{Y} - P_M(\mathbf{Y})\|^2/(n-k-1)} \\ &= \frac{(\text{SSTot} - \text{SSRes})/k}{\text{SSRes}/(n-k-1)}\end{aligned}$$

From the normal theory (p. 104) this is exactly the test statistic for the hypothesis

$$\begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and the distribution of the test statistic is a $F(k, n - k - 1)$ -distribution - exactly the same as we found on p. 104.

For testing it is from the numerical point of view therefore of no importance if we choose to consider the x 's as observations of a k -dimensional normally distributed random variable or as fixed deterministic variables.

This issue can therefore be separated from the assumptions we will consider in the next section.

3.1.3 Analysis of assumptions.

If we for corresponding x -values

$$x_{i1}, \dots, x_{ik}$$

have more observations of Y , it would be possible to compute the usual tests for distributional type (histograms, quantile diagrams, χ^2 -tests, etc.) and for the homogeneity of variances (Bartlett's test and others). Finally we could also do run tests for randomness etc. etc.

However, the situation is often that we very seldom have (more than maybe a couple) of repetitions for different values of the independent variable. It is therefore not possible to do these types of checks of the assumptions. Instead we consider the residuals

$$e_i = Y_i - \hat{e}(Y_i) = Y_i - \hat{\alpha} - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_k x_{ik}.$$

If the model is valid these will be approximately independent and $N(0, \sigma^2)$ distributed.

Initially we shall present some results on the residuals in a regression model. We recall the definitions of the hat matrix \mathbf{H} and the matrix \mathbf{M} presented in section 2.2. In the full rank case they are

$$\begin{aligned}\mathbf{H} &= \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' \\ \mathbf{M} &= \mathbf{I} - \mathbf{H}\end{aligned}$$

and we see that the predicted values are

$$\hat{\mathbf{Y}} = \mathbf{x}\hat{\boldsymbol{\theta}} = \mathbf{HY}$$

and the vector of residuals

$$\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{MY}.$$

Using results from section 1.1.2 and 1.1.4 we obtain

$$\mathbf{D}(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{HH}' = \sigma^2 \mathbf{H}$$

and

$$\mathbf{D}(\mathbf{R}) = \sigma^2 \mathbf{MM}' = \sigma^2 \mathbf{M}.$$

Since \mathbf{H} and \mathbf{M} are not diagonal it follows that the predicted values are correlated as are the residuals. If we denote element (i, j) in the hat matrix \mathbf{H} by h_{ij} we see that the variance of the predicted value is

$$\mathbf{V}(\hat{Y}_i) = h_{ii}\sigma^2 = \left(\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\right)_{ii} \sigma^2$$

and the residual variance becomes

$$\mathbf{V}(R_i) = (1 - h_{ii})\sigma^2 = \left(\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\right)_{ii} \sigma^2.$$

Furthermore it follows that *the residuals and the predicted values are uncorrelated* since

$$\text{Corr}(\hat{\mathbf{Y}}, \mathbf{R}) = \text{Corr}(\mathbf{HY}, \mathbf{MY}) = \sigma^2 \mathbf{HM} = \mathbf{0}.$$

And finally we find that the sum of the residuals is 0. This follows from

$$\mathbf{1}' \mathbf{R} = \mathbf{1}' \mathbf{M} \mathbf{Y} = 0$$

since $\mathbf{1}'$ is the first row in \mathbf{x}' and we have

$$\mathbf{x}' \mathbf{M} = \mathbf{x}' - \mathbf{x}' \mathbf{x} (\mathbf{x}' \mathbf{x})^{-1} \mathbf{x}' = \mathbf{0}.$$

If one depicts the residuals in different ways and thereby sees something which does not look (or could not be) observations of independently $N(0, \sigma^2)$ distributed random variables then we have an indication that there is something wrong with the model.

Most often we would probably start with a usual analysis of the distribution of the residuals i.e. do run-tests, draw histograms, quantile diagrams etc.

Afterwards we could depict the residuals against different quantities (time, independent variables, etc.). We show the following 4 sketches to illustrate often seen residual plots and give a short description of what the reason for plots of this kind could be. First we note that 1 always is acceptable (however, cf. p. 132).

i) Plot of residuals against time

- 2 The variance increases with time. Perform a weighted analysis.
- 3 Lack terms of the form $\beta \cdot \text{time}$
- 4 Lack terms of the form $\beta_1 \cdot \text{time} + \beta_2 \cdot \text{time}^2$

ii) Plot of residuals against $\hat{e}(Y_i)$

- 2 The variance increases with $E(Y_i)$. Perform a weighted analysis or transform the Y 's (e.g. with the logarithm or equivalent)
- 3 Lack constant term (the regression is possibly erroneously forced through 0). Error in the analysis.
- 4 Bad model. Try with a transformation of the Y 's.

iii) Plot against independent variable x_i

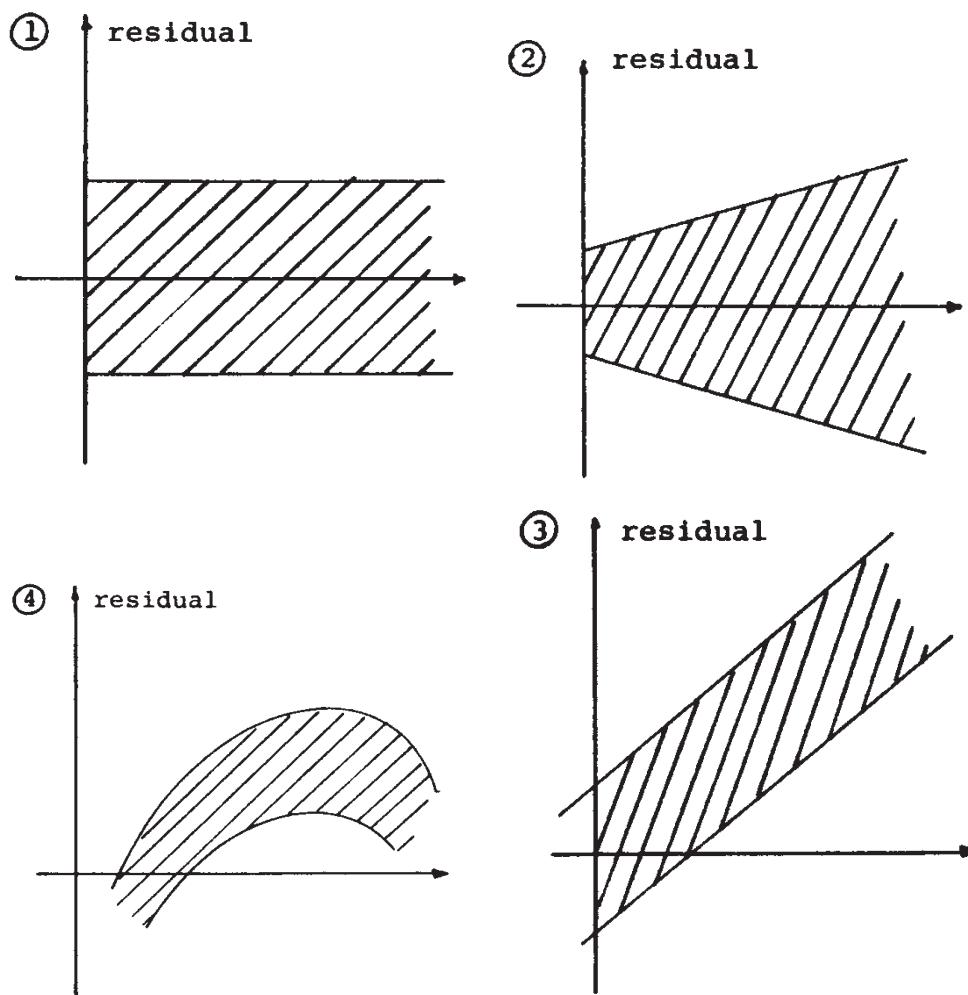


Figure 3.2: Residual plots.

- 2 The variance grows with x_i . Perform a weighted analysis or transform the Y's.
- 3 Error in the computations
- 4 Lacks quadratic term in x_i

The above is not meant to be an exhaustive description of how to analyse residual plots but may be considered as an indication of how such an analysis could be done.

|||| **Remark 3.1**

In practise we will often have our residual plot printed on printer listings. Then the plots might look as shown on p. 133. The 4 plots have been taken from [7] p. 14-15 in appendix C.

When interpreting these plots we should remember that there are not always an equal numbers of observations for each value of the independent variable.

This is e.g. the case in the plot which depicts the residual against variable 10.

There are 7 observations corresponding to $x_{10} \sim 0.2704\text{e}04$ and 35 observations corresponding to $x_{10} \sim 0.7126\text{e}03$. The range of variation for the residuals is approximately the same in the two cases. If the residuals corresponding to the 2 values of x_{10} had the same variance we would, however, expect the range of variation for the one with many observations to be the largest.

In other words if one has most observations around the centre of gravity for an independent variable a residual plot should rather be elliptical than of the form 1 to be satisfactory.

3.1.4 On “Influence Statistics”

When judging the quality of a regression analysis one often consider the following two possibilities:

- 1) Check if deviations from the model look random.
- 2) Check the effect of single observations on the parameter estimates etc.

Considerations regarding 1) are given in section 3.1.3 above. Here we will briefly consider 2).

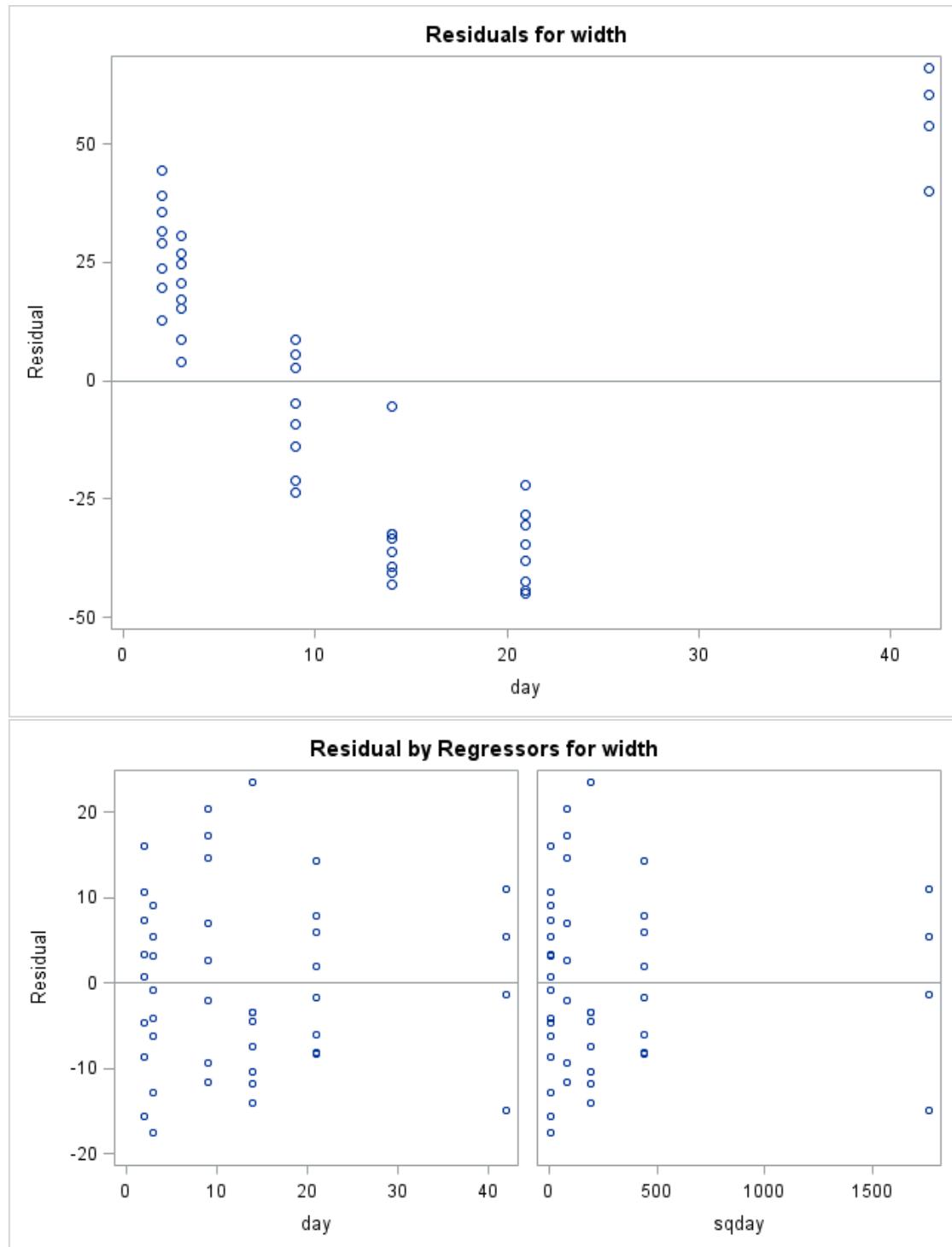


Figure 3.3: Top: Residuals in regression model describing the dependent variable "width" as a linear function of the independent variable "day". Bottom: Residuals after fitting the same data, but now as a linear function of "day" and "day squared".

The deletion formula

Re-calculation of parameter estimates when discarding a single observation can be done using the formula

$$(\mathbf{A} - \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 - \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}},$$

where the involved matrices are assumed to exist.

We let \mathbf{x}_i be the i 'th row in the design matrix \mathbf{x} . Letting $\mathbf{A} = \mathbf{x}'\mathbf{x}$ and $\mathbf{u} = \mathbf{v} = \mathbf{x}'_i$ we get

$$(\mathbf{x}'\mathbf{x} - \mathbf{x}'_i\mathbf{x}_i)^{-1} = (\mathbf{x}'\mathbf{x})^{-1} + \frac{(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'_i\mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}}{1 - h_{ii}}$$

If we denote the \mathbf{x} -matrix where the i 'th row is removed $\mathbf{x}(i)$ we have that

$$\mathbf{x}(i)' \mathbf{x}(i) = \mathbf{x}'\mathbf{x} - \mathbf{x}'_i\mathbf{x}_i.$$

The proof is omitted.

We can now state the relevant expressions.

Cook's D

A confidence region for the parameter θ is all the vectors θ^* , which satisfy

$$\frac{1}{p\hat{\sigma}^2}(\hat{\theta} - \theta^*)' \mathbf{x}'\mathbf{x}(\hat{\theta} - \theta^*) \leq F(p, n-p)_{1-\alpha}.$$

We use the left hand side as a measure of the distance between the parameter vector and $\hat{\theta}$. We let $\hat{\theta}(i)$ be the estimate, which corresponds to the deletion of the i 'th observation

$$\mathbf{y}(i) = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)'$$

and therefore have

$$\hat{\theta}(i) = [\mathbf{x}(i)' \mathbf{x}(i)]^{-1} \mathbf{x}(i)' \mathbf{y}(i).$$

Cook's D then equals

$$\frac{1}{p\hat{\sigma}^2}(\hat{\theta} - \hat{\theta}(i))' \mathbf{x}'\mathbf{x}(\hat{\theta} - \hat{\theta}(i)).$$

If Cook's D equals e.g. $F_{60\%}$ then this corresponds to the maximum likelihood estimate moving to the 60 % confidence-ellipsoid for θ . This is a relatively large change when just removing a single observation. There are several suggestions for cutoff values of Cook's D:

$$\begin{aligned} D &> 1 \\ D &> \frac{1}{n} \\ D &> \frac{1}{n-p} \\ D &> F(n, n-p)_{0.50} \end{aligned}$$

In the SAS-program REG one can find Cook's D together with other diagnostics statistics. Some are mentioned below.

RSTUDENT & STUDENT RESIDUAL

RSTUDENT is a so-called "studentised" residual, i.e.

$$\text{RSTUDENT}_i = \frac{r_i}{\hat{\sigma}(i)\sqrt{1-h_{ii}}},$$

where $\hat{\sigma}(i)^2$ is the estimate of variance corresponding to deletion of the i 'th observation.

SAS also computes a similar statistic, where the i 'th observation is not excluded

$$\text{STUDENT RESIDUAL}_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

Since both these types of residual are standardised a sensible rule of thumb is that they should lie within ± 2 or ± 3 .

COVRATIO

COVRATIO measures the change in the determinant of the dispersion matrix for the parameter estimate when excluding the i 'th observation. We find

$$\text{COVRATIO}_i = \frac{\det[\hat{\sigma}(i)^2(\mathbf{x}(i)'\mathbf{x}(i))^{-1}]}{\det[\hat{\sigma}^2(\mathbf{x}'\mathbf{x})^{-1}]}$$

This quantity "should" be close to 1. If it lies far from 1 then the i 'th observation has a too large influence. As a rule of thumb $|\text{COVRATIO}_i - 1| \leq 3p/n$

Leverage

The quantity h_{ii} introduced earlier is called the *Leverage*. It is

$$h_{ii} = \mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'_i$$

and it measures how far the i 'th vector of independent variables is from the mean of the remaining. Thus it is a measure of influence, since it 'forces' the regression surface to lie closer to this point. If we have p parameters in the model, points with a leverage $> 2p/n$ should be investigated.

DFFITS

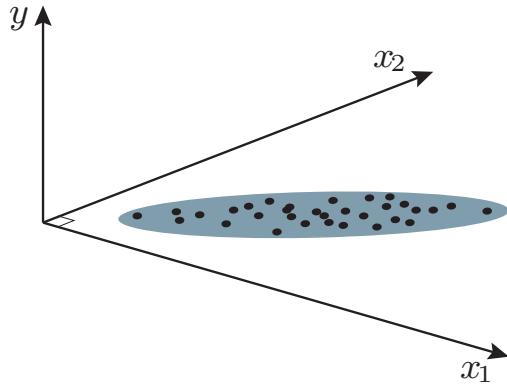
DFFITS is - like Cook's distance - a measure of the total change when deleting one single observation. As a rule of thumb they should lie within say ± 2 . A similar rule adjusted for number of observations says within $\pm 2\sqrt{p/(n-p)}$.

$$\begin{aligned} \text{DFFITS} &= \frac{\hat{y}_i - \hat{y}(i)_i}{\hat{\sigma}(i)\sqrt{h_{ii}}} \\ &= \frac{\mathbf{x}_i[\hat{\theta} - \hat{\theta}(i)]}{\hat{\sigma}(i)\sqrt{h_{ii}}}. \end{aligned}$$

DFBETAS

While DFFITS measures changes in the prediction of an observation corresponding to changes in all parameter estimates, then DFBETAS simply measures the change in each individual parameter estimate. As a rule of thumb they should lie within say ± 2 . A rule adjusted for number of observations says within $\pm 2/\sqrt{n}$.

$$\text{DFBETAS}_j = \frac{\hat{\theta}_j - \hat{\theta}(i)_j}{\hat{\sigma}(i)\sqrt{(\mathbf{x}'\mathbf{x})_{jj}^{-1}}}.$$



If we have a model

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and must estimate \$\beta_1\$ \$\beta_2\$ that can not be done in a reliable way since we cannot vary one \$x\$ with the other fixed.

Figure 3.4: All the \$(x_1, x_2)\$ are in the shaded (blue) area

Multicollinearity

If the independent (explanatory) variables in a multiple regression are highly correlated, we say that we have case with multicollinearity. This may cause that the individual parameter estimates are very uncertain, without necessarily ruining the descriptive and predictive power of the model, as long as the explanatory variables vary in the same range cf. figure 3.4. But predictions where we move out of the range may be highly unreliable.

Diagnostic checks for multicollinearity in SAS include methods based on measuring the correlation between one independent variable and all the others, and the other on analysing the eigenvalues of \$xx'\$.

||| Definition 3.2

We define the *tolerance* (TOL) and the *variance inflation* (VIF) as

$$\begin{aligned} TOL_i &= 1 - R^2(x_i | \text{all other } x\text{-variables}) \\ VIF_i &= \frac{1}{TOL_i} \end{aligned}$$

As a rule of thumb, \$TOL < 0.1\$ or equivalently \$VIF > 10\$ indicates a multicollinearity problem.

||| Definition 3.3

We define the *condition number* as the square root of the largest eigenvalue of $\mathbf{x}\mathbf{x}'$ divided by the smallest. The condition number should be below 15. If it is above 30, it is a matter of serious concern.

Call in SAS

All the mentioned statistics can be found using simple SAS statements e.g.

```
proc reg data = sundhed;
model ilt = maxpuls loebetid / r influence;
```

Model statements etc. are the same in REG as in GLM. The diagnostic tests come with the options / r influence.

3.2 Regression using orthogonal polynomials

When performing a regression analysis using polynomials one can often obtain rather large computational savings and numerical stability by introducing the so-called orthogonal polynomials. In the end this will give the same expression for estimates of the mean value as a function of the independent variable but with considerably smaller computational load.

3.2.1 Definition and formulation of the model.

We will assume that a polynomial regression model is given i.e. that

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \xi_0(t_1) & \xi_1(t_1) & \cdots & \xi_k(t_1) \\ \vdots & \vdots & & \vdots \\ \xi_0(t_n) & \xi_1(t_n) & \cdots & \xi_k(t_n) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Here ξ_i , $i = 0, 1, \dots, k$ are known polynomials of i 'th degree in t . We assume that

$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in N(\mathbf{0}, \sigma^2 \mathbf{I})$$

In the usual fashion we can in this model estimate and test hypotheses regarding the parameters $(\alpha, \beta_1, \dots, \beta_k)$.

As noted before it would be a great advantage to consider the so-called orthogonal polynomials ξ_i since the computational load will be reduced considerably. We introduce these polynomials in

||| Definition 3.4

By a set of orthogonal polynomials corresponding to the values t_1, \dots, t_n we mean polynomials ξ_0, ξ_1, \dots where ξ_i is of i 'th degree which satisfy

$$\sum_{j=1}^n \xi_i(t_j) = 0, \quad i = 1, 2, \dots, k \quad (3-1)$$

$$\sum_{j=1}^n \xi_\mu(t_j) \xi_\gamma(t_j) = 0, \quad \mu \neq \gamma. \quad (3-2)$$

||| Remark 3.5

It is seen that ξ_0 is a constant, so 3-1 is of course not used for ξ_0 . For notational reasons we let $\xi_i(t_j) = \xi_{ij}, \forall i, j$. Later we will return to the problem of actually determining orthogonal polynomials.

If we now assume that the polynomials in the model are orthogonal we find

using

$$\boldsymbol{\xi} = \begin{bmatrix} \xi_0 & \cdots & \xi_{k1} \\ \vdots & & \vdots \\ \xi_0 & \cdots & \xi_{kn} \end{bmatrix} = \begin{bmatrix} \xi_0(t_1) & \xi_1(t_1) & \cdots & \xi_k(t_1) \\ \vdots & \vdots & & \vdots \\ \xi_0(t_n) & \xi_1(t_n) & \cdots & \xi_k(t_n) \end{bmatrix},$$

that

$$\boldsymbol{\xi}' \boldsymbol{\xi} = \begin{bmatrix} n\xi_0^2 & 0 & \cdots & 0 \\ 0 & \sum \xi_{1j}^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sum \xi_{kj}^2 \end{bmatrix},$$

i.e. $\xi' \xi$ is a diagonal matrix. We therefore find

$$\hat{\beta} = (\xi' \xi)^{-1} \xi' Y = \begin{bmatrix} \bar{Y}/\xi_0 \\ \sum \xi_{1j} Y_j / \sum \xi_{1j}^2 \\ \vdots \\ \sum \xi_{kj} Y_j / \sum \xi_{kj}^2 \end{bmatrix}$$

and

$$D(\hat{\beta}) = \sigma^2 \begin{bmatrix} 1/n \xi_0^2 & 0 & \cdots & 0 \\ 0 & 1/\sum \xi_{1j}^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & 1/\sum \xi_{kj}^2 \end{bmatrix}.$$

We now have that the estimators for the parameters are uncorrelated and since we are working in a normal model they are therefore also stochastic independent.

We find that the residual sum of squares is

$$\begin{aligned} SS_{\text{res}} &= \|Y - \xi \hat{\beta}\|^2 \\ &= (Y - \xi \hat{\beta})' (Y - \xi \hat{\beta}) \\ &= Y' Y - \hat{\beta}' \xi' \xi \hat{\beta} \\ &= \sum Y_j^2 - \{\hat{\alpha}^2 n \xi_0^2 + \hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2\} \\ &= \sum (Y_j - \bar{Y})^2 - \{\hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2\}. \end{aligned}$$

From this we immediately have

||| Theorem 3.6

We have the following partitioning of the total variation

$$\sum (Y_j - \bar{Y})^2 = \hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2 + \sum \{Y_j - \bar{Y} - \hat{\beta}_1 \xi_1(t_j) - \cdots - \hat{\beta}_k \xi_k(t_j)\}^2,$$

or with an easily understood notation

$$SS_{\text{tot}} = SS_{1,\text{grad}} + \cdots + SS_{k,\text{grad}} + SS_{\text{res}},$$

i.e. the total sum of squares has been partitioned in terms corresponding to each polynomial plus the residual sum of squares. The degrees of freedom are $n - 1$ respectively $1, \dots, 1$ and $n - k - 1$.

||| Proof

Follows trivially from the above mentioned.

■

Using the partition theorem we furthermore have

||| Theorem 3.7

The sums of squares which have been stated in the previous theorem are stochastically independent with expected values

$$\begin{aligned} E(\text{SS}_{i,\text{deg}}) &= E\left(\hat{\beta}_i^2 \sum_j \xi_i(t_j)^2\right) \\ &= \sigma^2 + \beta_i^2 \sum_j \xi_i(t_j)^2, \quad i = 1, \dots, k. \end{aligned}$$

and

$$E(\text{SS}_{\text{res}}) = E\left[\sum_j (Y_j - \bar{Y} - \dots - \hat{\beta}_k \xi_k(t_j))^2\right] = (n - k - 1)\sigma^2.$$

Finally

$$\frac{1}{\sigma^2} \text{SS}_{\text{res}} \in \chi^2(n - k - 1),$$

and if $\beta_i = 0$ -

$$\frac{1}{\sigma^2} \text{SS}_{i,\text{deg}} \in \chi^2(1).$$

||| Proof

Obvious.

■

The theorems contain the necessary results to be able to establish tests for the hypotheses

$$H_{0i} : \beta_i = 0 \quad \text{against} \quad H_{1i} : \beta_i \neq 0.$$

We collect the results in a analysis of variance table

Variation	SS	f	$E(SS/f)$
Linear	$SS_{1.\text{deg}}$	1	$\sigma^2 + \beta_1^2 \sum_j \xi_1(t_j)^2$
Quadratic	$SS_{2.\text{deg}}$	1	$\sigma^2 + \beta_2^2 \sum_j \xi_2(t_j)^2$
Cubic	$SS_{3.\text{deg}}$	1	$\sigma^2 + \beta_3^2 \sum_j \xi_3(t_j)^2$
\vdots	\vdots	\vdots	\vdots
k' 'th order	$SS_{k.\text{deg}}$	1	$\sigma^2 + \beta_k^2 \sum_j \xi_k(t_j)^2$
Residual	SS_{res}	$n - k - 1$	σ^2
Total	SS_{tot}	$n - 1$	

||| **Remark 3.8**

The big advantage of using orthogonal polynomials in the regression analysis is that one without changing any of the previous computations can introduce polynomials of degree $(p + 1)$ and degree $(p + 2)$ etc. When establishing the order for the describing polynomial we will usually continue (estimation and) testing until 2 successive β_i 's = 0 since contributions which are caused by terms of even degree and terms of odd degree are different in nature. This is, however, a rule of thumb which should be used with caution. If we e.g. have an idea which is based on physical considerations that terms of 5th order are important, then we would not stop the analysis just because the 3rd and 4th degree coefficients do not differ significantly from 0.

3.2.2 Determination of orthogonal polynomials.

It is readily seen, that multiplication with a constant does not change the orthogonality conditions 3-1 and 3-2. We therefore choose to let

$$\xi_0(t) = \xi_0 = 1.$$

The polynomial of 1st degree is

$$\xi_1(t) = t + a,$$

since we can choose the coefficient for t as 1. From 3-1 we have

$$0 = \sum_{j=1}^n \xi_1(t_j) = \sum_{j=1}^n (t_j + a) = \sum_{j=1}^n t_j + na,$$

or

$$a = -\frac{1}{n} \sum_{j=1}^n t_j = -\bar{t},$$

i.e.

$$\xi_1(t) = t - \bar{t}.$$

We can then choose ξ_2 as a linear combination of 1, ξ_1 , ξ_1^2 , i.e.

$$\xi_2(t) = a_{02} + a_{12}(t - \bar{t}) + a_{22}(t - \bar{t})^2.$$

From 3-1 we have

$$\begin{aligned} 0 &= \sum_{j=1}^n \xi_2(t_j) = na_{02} + a_{12} \sum_j (t_j - \bar{t}) + a_{22} \sum_j (t_j - \bar{t})^2 \\ \frac{a_{02}}{a_{22}} &= -\frac{1}{n} \sum_j (t_j - \bar{t})^2. \end{aligned}$$

From 3-2 we have

$$\begin{aligned} 0 &= \sum_{j=1}^n \xi_1(t_j) \xi_2(t_j) \\ &= a_{02} \sum_j (t_j - \bar{t}) + a_{12} \sum_j (t_j - \bar{t})^2 + a_{22} \sum_j (t_j - \bar{t})^3 \\ &= a_{12} \sum_j (t_j - \bar{t})^2 + a_{22} \sum_j (t_j - \bar{t})^3. \end{aligned}$$

From this we get

$$\frac{a_{12}}{a_{22}} = -\frac{\sum_j (t_j - \bar{t})^3}{\sum_j (t_j - \bar{t})^2}.$$

ξ_3, ξ_4 etc. are found analogously.

The computations are especially simple if the t_j 's are equidistant. Then we let

$$u_j = \frac{t_j - (t_1 - w)}{w},$$

where $w = t_2 - t_1 = t_{i+1} - t_i$. We then have

$$u_i = i, \quad i = 1, \dots, n.$$

Corresponding to the values $1, \dots, n$ we then have the polynomials given by

$$\xi_0(t) = 1 \quad (3-3)$$

$$\xi_1(t) = t - \frac{n+1}{2} \quad (3-4)$$

$$\xi_{i+1}(t) = \xi_1(t)\xi_i(t) - \frac{i^2(n^2 - i^2)}{4(4i^2 - 1)}\xi_{i-1}(t). \quad (3-5)$$

In the table on p. 145 we have given some values of orthogonal polynomials ξ_1, \dots, ξ_k , $k \leq 5$, with $t = 1, \dots, n$ for $n = 1, \dots, 8$.

In order to avoid fractional numbers and large values we have chosen to give polynomials where the coefficient to the term of largest degree is a number λ which is also seen in the table. Furthermore we have stated the terms

$$D = \sum_{j=1}^n \xi_i(j)^2 = \sum_{j=1}^n \xi_{ij}^2.$$

n	3	4	5	6	7	8											
t	ξ_1	ξ_2	ξ_1	ξ_2	ξ_3	ξ_4	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	
1	-1	1	-3	1	-1	-2	2	-1	1	-5	5	-5	1	-1	-3	5	
2	0	-2	-1	-1	3	-1	-1	2	-4	-3	-1	7	-3	5	-2	0	
3	1	1	1	-1	-3	0	-2	0	6	-1	-4	4	3	-10	-1	-3	
4		3	1	1	1	-1	-2	-4	1	-4	-4	2	10	0	-4	0	
5			2	2	1	1	3	-1	-7	-3	-5	1	-3	-1	1	5	
6							5	5	5	1	1	2	0	-1	-7	-4	
7												3	5	1	3	1	
8																7	
D	2	6	20	4	20	10	14	10	70	84	180	28	252	28	84	6	
λ	1	3	2	1	$\frac{10}{3}$	1	$\frac{1}{3}$	1	$\frac{5}{6}$	$\frac{35}{12}$	2	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{7}{12}$	$\frac{21}{10}$	1	$\frac{1}{6}$
														$\frac{7}{12}$	$\frac{7}{20}$	2	
														1	$\frac{2}{3}$	$\frac{7}{12}$	
															$\frac{7}{10}$		

Table 3.2.2: Values of orthogonal polynomials.

We now give an illustrative

||| Example 3.9

In the following table corresponding values of reaction temperature and yield of a process (in a fixed time) have been given.

Temperature	Yield
200°F	0.75 oz.
210°F	1.00 oz.
220°F	1.35 oz.
230°F	1.80 oz.
240°F	2.60 oz.
250°F	3.60 oz.
260°F	5.45 oz.

We will try to describe the yield as a function of temperature using a polynomial. We will assume that the assumptions in order to perform a regression analysis are fulfilled. First we transform the temperatures $\tau_i, i = 1, \dots, 7$ by means of the following relation

$$t_i = \frac{\tau_i - (200 - 10)}{10} = \frac{\tau_i - 190}{10}$$

We then get the values $t_1, \dots, t_7 = 1, \dots, 7$.

We give the computations in the following table

t_j	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	y_j
1	-3	5	-1	3	-1	0.75
2	-2	0	1	-7	4	1.00
3	-1	-3	1	1	-5	1.35
4	0	-4	0	6	0	1.80
5	1	-3	-1	1	5	2.60
6	2	0	-1	-7	-4	3.60
7	3	5	1	3	1	5.45
$\sum \xi_{ij}^2$	28	84	6	154	84	$16.55 = \sum y_j$
$\sum \xi_{ij} y_j$	20.55	11.95	0.85	1.15	0.55	$56.0475 = \sum y_j^2$
λ	1	1	$\frac{1}{6}$	$\frac{7}{12}$	$\frac{7}{20}$	

$$\begin{aligned} \sum (y_i - \bar{y})^2 &= 56.0475 - \frac{16.55^2}{7} \\ &= 56.0475 - 39.1289 \\ &= 16.9186 \end{aligned}$$

$$\begin{aligned}
 \hat{\alpha} &= \frac{16.55}{7} = 2.36 \\
 \hat{\beta}_1 &= \frac{20.55}{28} = 0.7339 & SS_{1.\text{grad}} &= \frac{20.55^2}{28} = 15.0822 \\
 \hat{\beta}_2 &= \frac{11.95}{84} = 0.1423 & SS_{2.\text{grad}} &= \frac{11.95^2}{84} = 1.7000 \\
 \hat{\beta}_3 &= \frac{0.85}{6} = 0.1417 & SS_{3.\text{grad}} &= \frac{0.85^2}{6} = 0.1204 \\
 \hat{\beta}_4 &= \frac{1.15}{154} = 0.0075 & SS_{4.\text{grad}} &= \frac{1.15^2}{154} = 0.0086 \\
 \hat{\beta}_5 &= \frac{0.55}{84} = 0.0065 & SS_{5.\text{grad}} &= \frac{0.55^2}{84} = 0.0036
 \end{aligned}$$

We summarise the result in the following table.

We see that the terms of 1st, 2nd and 3rd degree are significant and the two following are not significant, so we will choose a polynomial of 3rd degree for the description.

Variation	SS	f	S^2	Test	F-percentile
Total	16.9186	6			
1. degree	15.0822	1	15.0822		
Residual 1	1.8364	5	0.3673	41.06	99.8%
2. degree	1.7000	1	1.7000		
Residual 2	0.1364	4	0.0341	49.85	99.7%
3. degree	0.1204	1	0.1204		
Residual 3	0.0160	3	0.0053	22.72	98.0%
4. degree	0.0086	1	0.0086		
Residual 4	0.0074	2	0.0037	2.32	75.0%
5. degree	0.0036	1	0.0036		
Residual 5	0.0038	1	0.0038	0.95	< 50.0%

From the recursion formulas 3-3, 3-4 and 3-5 we get - since $n = 7$

$$\begin{aligned}
 \xi_1(t) &= t - 4 \\
 \xi_2(t) &= (t - 4)^2 - \frac{48}{12} \\
 &= t^2 - 8t + 12 \\
 \xi_3(t) &= (t - 4)(t^2 - 8t + 12) - \frac{4 \cdot 45}{4 \cdot 15}(t - 4) \\
 &= t^3 - 12t^2 + 41t - 36.
 \end{aligned}$$

Since $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 1/6$ we get the following estimated polynomial

$$\begin{aligned}
 \hat{\mu}(t) &= 2.36 + 1 \cdot \hat{\beta}_1 \xi_1(t) + 1 \cdot \hat{\beta}_2 \xi_2(t) + \frac{1}{6} \hat{\beta}_3 \xi_3(t) \\
 &= 0.0236t^3 - 0.1409t^2 + 0.5631t + 0.2818.
 \end{aligned}$$

Since

$$t_i = \frac{\tau_i - 190}{10},$$

we can get an expression where the original temperatures are given by entering this relationship in the expression for $\hat{\mu}(t)$. We find

$$g(\tau) = 0.000024\tau^3 - 0.014861\tau^2 + 3.147610\tau - 223.15440.$$

The estimated polynomial is shown together with the original data in the following

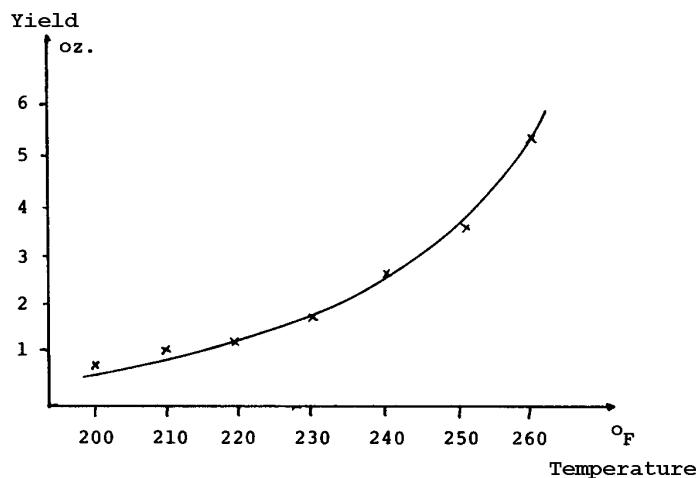


figure.

Figure 3.9: The correspondence between temperature and yield by the process given in example 3.9.

3.3 Choice of the "best" regression equation

In this section we will consider the problem of choosing a suitable (small) number of independent variables giving a reasonable description of our data.

3.3.1 The Problem.

If we are in the (unpleasant) situation of not being able to formulate a model based upon physical relationships for the phenomena we are studying, we will often simply register all the variables we think could have some effect on our observed values. If we then compute a regression by e.g. polynomials in these independent variables (from a Taylor-approximation point of view) we will very quickly have an enormous number of terms in our regression. If we start off with 10 basic-variables x_1, \dots, x_{10} , then an ordinary second order polynomial in these variables will contain 66 terms. If we include 3rd degree we have on the order of 150 terms. Expressions containing so many terms will (if it is at all possible to estimate all the parameters) be very tedious to work with. If we e.g. wish to determine optimal production conditions for a chemical process we could estimate the response surface and find the maximum for this. This will be extremely difficult if there are many variables involved. We would therefore

seek to find a considerably smaller number of terms which will give a reasonably good description of the variation in the material (cf. the section on ridge regression).

It is important, however, to note that an expression found by applying the methods discussed in the following should be used with caution. It will (probably) be an expression which describes the data at hand very well. Whether or not the method is adequate to predict future observations depends upon if the expression also describes the physical conditions well enough. One way of determining this is in the first instance only to base the estimation on half of the data and then compare the other half with the estimated model. If the degree of agreement is reasonable we have the indication that the model is not completely inadequate as a prediction model.

||| Example 3.10

We will use a single illustrative example for all the methods we will describe. In order for it to be possible to overlook (and maybe check) the individual calculations we have only taken a very small part of the original data material. We should therefore not evaluate the suitability of the methods by means of the example, but only use it as an illustration of the principals and the way of going about these. The data are some corresponding measurements of the quality Y of a food additive (measured using viscosity) and some production parameters x_1, x_2, x_3 (pressure, temperature and degree of neutralisation). In order to simplify the calculations the data are coded, i.e. the variables have had some constants subtracted and been divided by others. We have the following measurements

y	x_1	x_2	x_3
4.9	0	0	2
3.0	1	0	1
0.2	1	1	0
2.9	1	2	2
6.4	2	1	2

Experience shows that within a suitably small region of variation of the production parameters it is reasonable to assume that the quality shows a linear dependency on these. We will therefore use the following model

$$E(Y|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

or in matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix},$$

$$\varepsilon \in N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

In the numerical appendix (p. 159) all the 2^3 regression analysis with y as dependent variable and one of the more of the x 's as independent variables are shown. The following models are possible

$$\begin{aligned} M &: E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ H_{12} &: E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 \\ H_{13} &: E(Y) = \alpha + \beta_1 x_1 + \beta_3 x_3 \\ H_{23} &: E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3 \\ H_1 &: E(Y) = \alpha + \beta_1 x_1 \\ H_2 &: E(Y) = \alpha + \beta_2 x_2 \\ H_3 &: E(Y) = \alpha + \beta_3 x_3 \\ H_0 &: E(Y) = \alpha \end{aligned}$$

For each of these 8 models the estimators for α and the β 's are shown, we find the projection of the observation vector onto the sub-space corresponding to the model we determine the residual vector, the squared length of the residual vector (the residual sum of squares), the estimate of variance, and the (squared) multiple correlation coefficient. After that we show the analysis of variance tables for the possible sequences of successive testings of hypotheses: that the mean vector is a member of successively smaller (lower dimension) sub-spaces in sequences like

$$M \supseteq H_{12} \supseteq H_2 \supseteq H_0.$$

The above mentioned sequence of sub-spaces corresponds to successive testing of the hypothesis

$$\beta_3 = 0, \quad \beta_1 = 0, \quad \beta_2 = 0.$$

There are 6 (= 3!) possible tables of this type. Finally we show some partial correlation matrices. If we let $y = x_4$ the empirical variance-covariance matrix is (as usual) defined by the (i, j) 'th element being

$$S_{ij} = \frac{1}{n-1} \sum_{\mu} (x_{i\mu} - \bar{x}_i)(x_{j\mu} - \bar{x}_j).$$

The (i, j) 'th element in the correlation matrix is then

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}.$$

Using the formula on p. 32 in section 1 we then compute the partial correlations for given x_3 and for given x_2, x_3 .

We now have enough background material to mention some of the most popular ways of selecting single independent variables to describe the variation of the dependent variable.

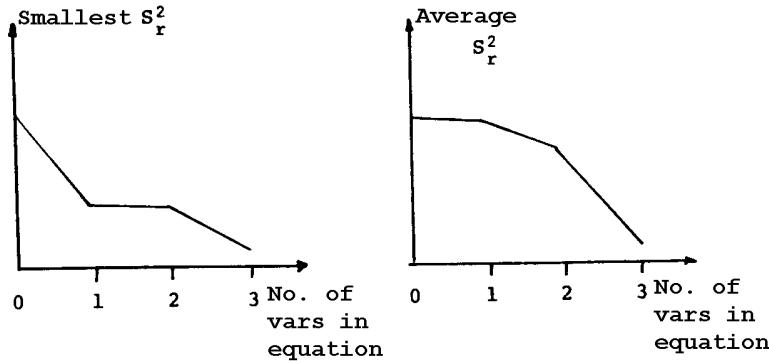


Figure 3.5:

3.3.2 Examination of all regressions.

This method can of course only be used if there are reasonably few variables. We summarise the result from the appendix in the following table

Model	Multiple R^2	Residual variance S_r^2	Average of S_r^2
$H_0 : E(Y) = \alpha$	0	5.47	5.47
$H_1 : E(Y) = \alpha + \beta_1 x_1$	5.1%	6.91	
$H_2 : E(Y) = \alpha + \beta_2 x_2$	3.8%	7.01	5.35
$H_3 : E(Y) = \alpha + \beta_3 x_3$	70.8%	2.13	
$H_{12} : E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$	15.3%	9.26	
$H_{13} : E(Y) = \alpha + \beta_1 x_1 + \beta_3 x_3$	76.0%	2.63	4.68
$H_{23} : E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$	80.4%	2.14	
$M : E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$	97.1%	0.634	0.634

Looking at the multiple correlation coefficient quickly indicates that we do not gain so much by going from one variable (x_3) up to 2 variables. The crucial jump happens when including all 3 variables. Considerations of this type lead us rather to just use x_3 i.e. the model $E(Y|x) = \alpha + \beta_3 x_3$. This decision is strengthened by looking at the residual variance S_r^2 . We then see that S_r^2 for the best equation in one variable is less than for the best equation in two variables which strongly indicates that we should just look at one variable (or use all three). If we besides looking at the smallest S_r^2 also look at the average values and depict them by number of included variables we have graphs like fig. 3.5.

This also indicates that the number of variables in an equation should be either 1 or 3 (there is no significant improvement by going from 1 to 2).

If we only look at the graph with the average values it is not obvious that we should include any independent variable at all. We could therefore test if $\beta_3 = 0$ in the model $H_3 \quad (\text{E}(y|x) = \alpha + \beta_3 x_3)$

$$\frac{\|p_{H_0}(\mathbf{y}) - p_{H_3}(\mathbf{y})\|^2/1}{\|\mathbf{y} - p_{H_3}\|^2/3} = \frac{21.868 - 6.38}{6.38/3} \simeq 7.28.$$

Therefore we will reject $\beta_3 = 0$ at all levels greater than 8%.

As a conclusion of these (rather loose) considerations we will use the model H_3 :

$$\text{E}(Y|x) = \alpha + \beta_3 x_3 \simeq 0.4 + 2.2x_3.$$

Here \simeq means estimated at). The estimate of the error (the variance) on the measurements is (estimated with 3 degrees of freedom):

$$s^2 = 2.13.$$

|||| Remark 3.11

It should be added here that the idea of looking at the averages of the residual variances does seem a bit dubious. It has been included merely because the method seems to enjoy widespread use - at least in some parts of the literature.

3.3.3 Backwards elimination.

This method is far more economical with respect to computational time than the previous one. Here we start with the full model M and then investigate which of the coefficients which has the smallest F-value for a test of the hypothesis that the coefficient might be 0.

This variable is then excluded and the procedure is repeated with the $k - 1$ remaining variables etc.

We can then stop the procedure when none of the remaining variables have an F-value less than the $1 - \alpha$ quantile in the relevant F-distribution.

We can illustrate the procedure using our example. We collect the data in the following table.

From the table can be seen that this procedure also will end with the model H_3 : $\text{E}(y) = \alpha + \beta_3 x_3$ when we use an α , greater than 8%.

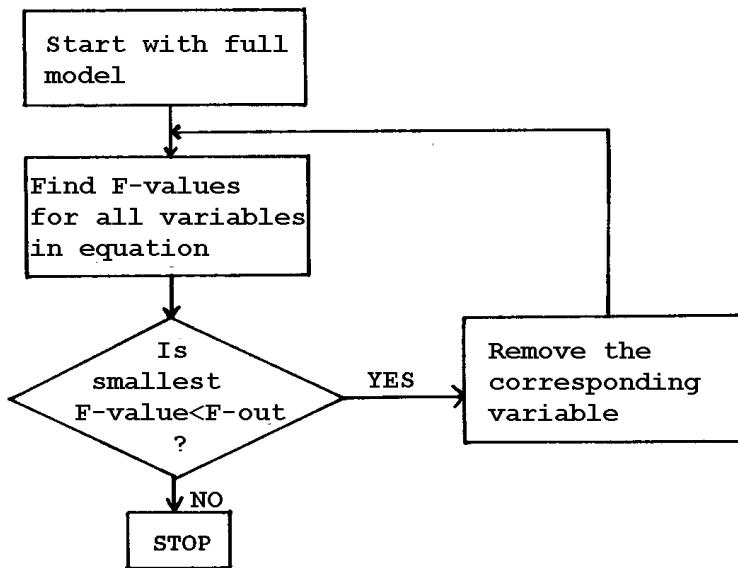


Figure 3.6: Flow diagram for Backwards-elimination procedure in stepwise regression analysis.

Step	F-value for test of $\beta_i = 0$	/ Quantile in F-distribution
Model : $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$		
1	$\beta_1 : \frac{3.44045/1}{0.845/1} = 4.07$	$= F(1, 1)_{0.71}$
	$\beta_2 : \frac{4.621/1}{0.634/1} = 7.29$	$= F(1, 1)_{0.72}$
	$\beta_3 : \frac{17.879/1}{0.634/1} = 28.20$	$= F(1, 1)_{0.86}$
Remove x_1 : Model is now : $E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$		
2	$\beta_2 : \frac{2.095/1}{4.285/2} = 0.98$	$= F(1, 2)_{0.55}$
	$\beta_3 : \frac{16.757/1}{4.285/2} = 7.82$	$= F(1, 2)_{0.88}$
Remove x_2 : Model is now : $E(Y) = \alpha + \beta_3 x_3$		
3	$\beta_3 : \frac{15.488/1}{6.38/3} = 7.28$	$= F(1, 3)_{0.92}$

The disadvantage with this method is, that we have to solve the full regression model which can be a problem if there many independent variables.

This problem is circumvented by using the following procedure.

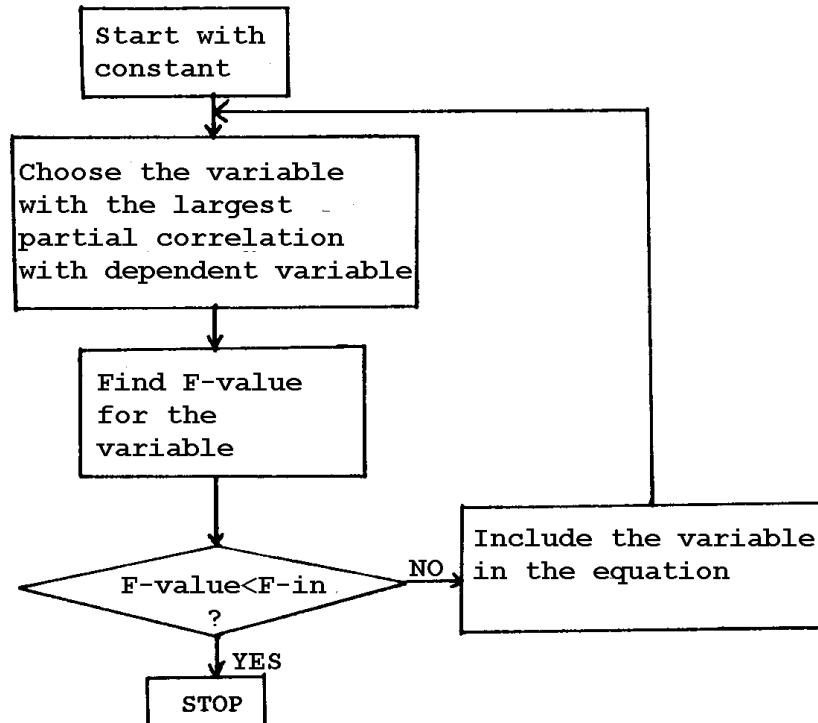


Figure 3.7: Flow diagram for Forward-selection procedure in stepwise regression analysis

3.3.4 Forward selection

In this procedure we start with the constant term in the equation only. Then we choose the independent variable which shows the greatest correlation with the dependent variable. We then perform an F-test to check if this coefficient is significantly different from 0. If so, then it is included in the model.

Among the independent variables not yet included we now choose the one that has the greatest (absolute) partial correlation coefficient with the dependent variable given the variables already in the equation. We perform an F-test to check if the new variable has contributed to the reduction of the residual variance, i.e. if the coefficient for it is different from 0. If so, continue as before if not stop the analysis.

In our example the steps will be the following

- 1) From the correlation matrix (p. 164) we see that x_3 has the greatest correlation coefficient with y , viz. 0.8416. We test if β_3 in the model $E(Y) = \alpha + \beta_3 x_3$ can be assumed to be 0 we have the test statistic (see p. 163).

$$\frac{15.488/1}{6.38/3} = 7.28 \simeq F(1, 3)_{0.92}.$$

If we use $\alpha = 10\%$ we continue (since we then reject $\beta_3 = 0$).

2) From the partial correlation matrix given x_3 (p. 164) we see that the variable which has the greatest partial correlation coefficient with the y 's (given that x_3 is in the equation) is x_2 ($\rho_{x_2y|x_3} = -0.5728$). We include x_2 and check if β_2 in the model

$$E(y) = \alpha + \beta_2 x_2 + \beta_3 x_3$$

can be assumed to be 0. We have the test statistic (see p. 163)

$$\frac{2.095/1}{4.2855/2} = 0.98 \simeq F(1, 2)_{0.55}.$$

Since we were using $\alpha = 10\%$, then this statistic is not significantly different from 0, and we stop the analysis here without including x_2 . The resulting model is

$$E(Y) = \alpha + \beta_3 x_3,$$

where α and β are estimated as earlier. We especially note that x_1 has not been included in the equation at all.

|||| Remark 3.12

If we had used $\alpha = 50\%$ we would have continued the analysis and considered the partial correlations given x_2 and x_3 . According to the matrix p. 164 the partial correlation coefficient between y and x_1 given that x_2 and x_3 are included in the equation

$$\rho_{x_1y|x_2x_3} = 0.8956.$$

Now x_1 is the only variable not included so it is trivially the one which has the greatest partial correlation with y . We now include x_1 in the equation and investigate if β_1 in the model $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ is significantly different from 0. The test statistic is (p. 163)

$$\frac{3.652/1}{0.634/1} = 5.76 \simeq F(1, 1)_{0.71}.$$

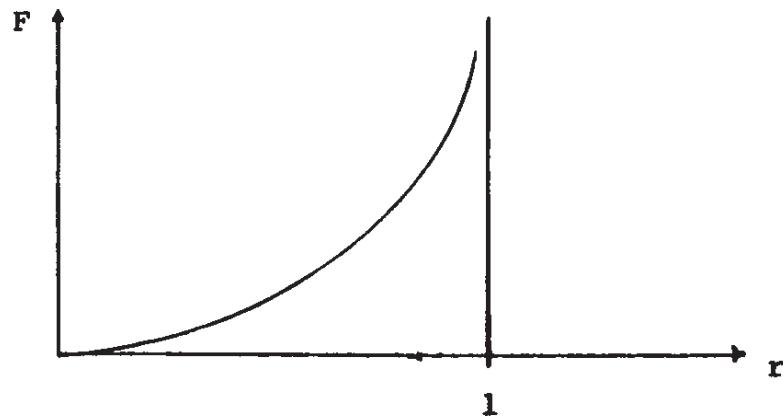
In the case we have seen that the equation was extended considerably just by changing α . It is important to note that changes in α can have drastic consequences for the resulting model.

|||| **Remark 3.13**

The procedure of choosing the variable which has the greatest partial correlation with the dependent variable at every step, is equivalent to choosing the variable which has the greatest F-value in the partial F-test. This result comes from the relation between the partial correlation coefficient and the F-statistic. This is of the form

$$F = g(r) = \frac{r^2}{1 - r^2} \cdot f,$$

where f is the number of degrees of freedom for the denominator (cf p. 128). This relation is monotonously increasing



If we e.g. in step 2 want to compute the F-test statistic from the correlation matrix we would get

$$F = \frac{(-0.5728)^2}{1 - (-0.5728)^2} \cdot 2 = 0.98.$$

It is further seen that the mentioned criterion is equivalent to at each step always taking the variable which gives the greatest reduction in residual sum of squares.

|||| **Remark 3.14**

In some of the existing standard regression programmes it is not possible to specify an α -value. We must then instead give a fixed number as the limit for the F-test statistics we will accept respectively reject. We must then by looking at a table over F-quantile find a suitable value. If we e.g. wish to have $\alpha = 5\%$, we see that we should use the value 4 since

$$F(1, n)_{0.95} \simeq 4,$$

for reasonably large values of n .

The 'forward selection' method has its merits compared to the backward elimination method in that we do not have to compute the total equation. The greatest drawback with the method is probably that we do not take into account that some of the variables could be redundant if others enter at a later stage. If we e.g. have that $x_1 = ax_2 + bx_3$ (approximately) and that x_1 has been chosen as the most important variable. If we then at a later stage in the analysis also include x_2 and x_3 then it is obvious that we no longer need x_1 . It should therefore be removed. This happens in the last method we mention.

3.3.5 Stepwise regression.

The name is badly chosen since we could equally well call the last two methods by this name. There are also many authors who use the name stepwise regression as a common name for a number of different procedures. In this text we will specifically have the following method in mind. Choice of the variable to enter the equation is performed like in the forward selection procedure, but at every single step we check each of the variables in the equation as if they were the last included variable. We then compute an F-test statistic for all the variables in the equation. If some of these are smaller than the $1 - \alpha$ quantile in the relevant F-distribution then the respective variable is removed. If we look at our standard example we get the following steps ($\alpha_{in} = 50\%$, $\alpha_{out} = 40\%$).

- 1) x_3 is included as in the forward selection procedure and we test if β_3 is significantly different from 0. The test statistic and the conclusion are as before.
- 2) We now include x_2 . We compute the partial F-test for β_2 (in the model

$E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$:

$$x_2 : \quad \text{F-value} = \frac{2.095/1}{4.285/2} = 0.98 \simeq F(1, 2)_{0.55}.$$

Then we compute a partial F-test for β_3 (in the model $E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$). Using the table p. 163 we find that

$$x_3 : \quad \text{F-value} = \frac{16.757/1}{4.285/2} = 7.82 \simeq F(1, 2)_{0.88}.$$

- 3) We now again remove x_2 from the equation since $0.55 < 0.60$. The difference at this step between the forward selection procedure and the stepwise procedure is that we also compute an F-value for x_3 and thereby have a possibility that x_3 again will be eliminated from the equation. This was not possible by the ordinary forward selection procedure.
- 4) The only remaining variable is x_1 . It has a partial F-value of

$$x_1 : \quad \text{F-value} = \frac{1.125/1}{5.255/2} = 0.43 < F(1, 2)_{0.50},$$

so it does not enter the equation at all.

The analysis stops and we have the model

$$E(Y) = \alpha + \beta_3 x_3.$$

||| Remark 3.15

The reason why we investigated the partial F-value under 2, but not under 4 is that x_1 does not enter the equation at all since

$$0.43 < F(1, 2)_{0.50} = F_{1-\alpha_{ind}}.$$

On the other hand x_2 was entered into the equation since

$$0.98 < F(1, 2)_{0.55} > F_{1-\alpha_{ind}}.$$

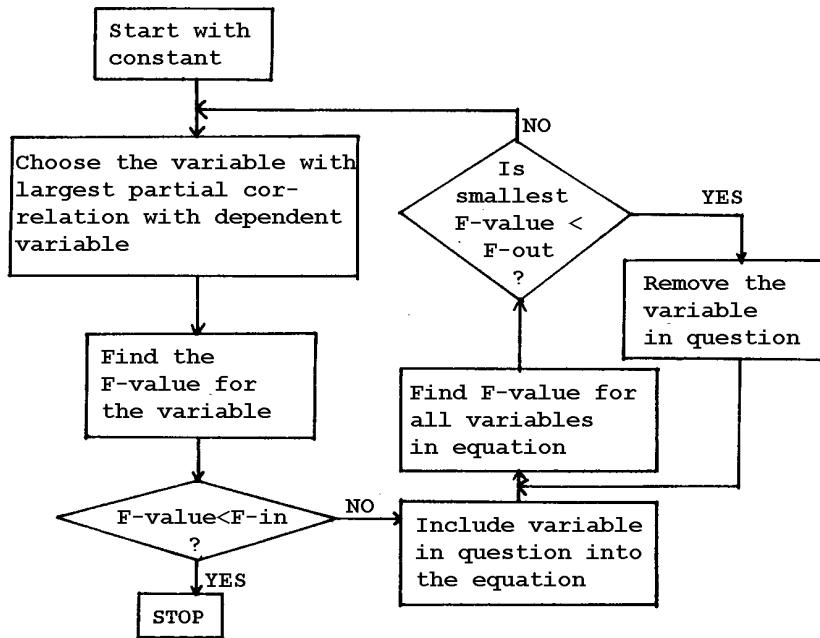


Figure 3.8: Flow diagram for Stepwise-Regression procedure in stepwise regression analysis.

||| Remark 3.16

Like the section on the forward selection procedure we can note that we are often forced to use fixed F-values instead of $1 - \alpha$ quantiles. If we do not use the same level when determining if we want to include more variables as we do when determining if some of the variables should be removed, we will often let the last value be about half as big as the first one i.e.

$$\text{F-out of equation} = \frac{1}{2}\text{F-into equation.}$$

(This is the opposite of what we actually used in the example).

3.3.6 Numerical appendix.

In this appendix we will show the calculation of the numbers used in the previous sections. It should not be necessary to go through all these computations but they are shown, so we with the help of these should be able to check our understanding of the different principles.

A. Data:

y	x_1	x_2	x_3
4.9	0	0	2
3.0	1	0	1
0.2	1	1	0
2.9	1	2	2
6.4	2	1	2

B. Basic Model: $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ or

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\varepsilon \in N(\mathbf{0}, \sigma^2 \mathbf{I})$$

C. Estimators in sub-models

i) **Model M:** $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} -0.175 \\ 1.450 \\ -1.400 \\ 2.375 \end{bmatrix}; p_M(\mathbf{y}) = \begin{bmatrix} 4.575 \\ 3.650 \\ -0.125 \\ 3.225 \\ 6.075 \end{bmatrix}; \mathbf{y} - p_M(\mathbf{y}) = \begin{bmatrix} 0.325 \\ -0.650 \\ 0.325 \\ -0.325 \\ 0.325 \end{bmatrix}$$

$$\frac{1}{5-4} \|\mathbf{y} - p_M(\mathbf{y})\|^2 = \frac{0.845}{1} = 0.845$$

$$R^2 = \frac{21.868 - 0.633750}{21.868} = 97.1\%$$

ii) **Model H_{12} :** $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3.026 \\ 1.243 \\ -0.987 \end{bmatrix}; p_{H_{12}}(\mathbf{y}) = \begin{bmatrix} 3.026 \\ 4.269 \\ 3.282 \\ 2.295 \\ 4.525 \end{bmatrix}; \mathbf{y} - p_{H_{12}}(\mathbf{y}) = \begin{bmatrix} 1.874 \\ -1.269 \\ -3.082 \\ 0.605 \\ -1.875 \end{bmatrix}$$

$$\frac{1}{5-3} \|\mathbf{y} - p_{H_{12}}(\mathbf{y})\|^2 = \frac{18.512611}{2} = 9.2563$$

$$R^2 = \frac{21.868 - 18.512611}{21.868} = 15.3\%$$

iii) **Model** H_{13} : $E(Y) = \alpha + \beta_1 x_1 + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} -0.350 \\ 0.750 \\ 2.200 \end{bmatrix}; p_{H_{13}}(\mathbf{y}) = \begin{bmatrix} 4.05 \\ 2.60 \\ 0.40 \\ 4.80 \\ 5.55 \end{bmatrix}; \mathbf{y} - p_{H_{13}}(\mathbf{y}) = \begin{bmatrix} 0.85 \\ 0.40 \\ -1.20 \\ -1.90 \\ 0.85 \end{bmatrix}$$

$$\frac{1}{5-3} \|\mathbf{y} - p_{H_{13}}(\mathbf{y})\|^2 = \frac{5.2250}{2} = 2.6275$$

$$R^2 = \frac{21.868 - 5.2550}{21.868} = 76.0\%$$

iv) **Model** H_{23} : $E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.945 \\ -0.872 \\ 2.309 \end{bmatrix}; p_{H_{23}}(\mathbf{y}) = \begin{bmatrix} 5.563 \\ 3.254 \\ 0.073 \\ 3.819 \\ 4.691 \end{bmatrix}; \mathbf{y} - p_{H_{23}}(\mathbf{y}) = \begin{bmatrix} -0.663 \\ -0.254 \\ 0.127 \\ -0.919 \\ 1.709 \end{bmatrix}$$

$$\frac{1}{5-3} \|\mathbf{y} - p_{H_{23}}(\mathbf{y})\|^2 = \frac{4.285456}{2} = 2.1427$$

$$R^2 = \frac{21.868 - 4.2855}{21.868} = 80.4\%$$

v) **Model** H_1 : $E(Y) = \alpha + \beta_1 x_1$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 2.73 \\ 0.75 \end{bmatrix}; p_{H_1}(\mathbf{y}) = \begin{bmatrix} 2.73 \\ 3.48 \\ 3.48 \\ 3.48 \\ 4.23 \end{bmatrix}; \mathbf{y} - p_{H_1}(\mathbf{y}) = \begin{bmatrix} 2.17 \\ -0.48 \\ -3.28 \\ -0.58 \\ 2.17 \end{bmatrix}$$

$$\frac{1}{5-2} \|\mathbf{y} - p_{H_1}(\mathbf{y})\|^2 = \frac{20.7430}{3} = 6.9143$$

$$R^2 = \frac{21.868 - 20.743}{21.868} = 5.1\%$$

vi) **Model** H_2 : $E(Y) = \alpha + \beta_2 x_2$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3.914 \\ -0.543 \end{bmatrix}; p_{H_2}(\mathbf{y}) = \begin{bmatrix} 3.914 \\ 3.914 \\ 3.371 \\ 2.828 \\ 3.371 \end{bmatrix}; \mathbf{y} - p_{H_2}(\mathbf{y}) = \begin{bmatrix} 0.986 \\ -0.914 \\ -3.171 \\ 0.072 \\ 3.029 \end{bmatrix}$$

$$\frac{1}{5-2} \|\mathbf{y} - p_{H_2}(\mathbf{y})\|^2 = \frac{21.042858}{3} = 7.0143$$

$$R^2 = \frac{21.868 - 21.043}{21.868} = 3.8\%$$

vii) Model H_3 : $E(Y) = \alpha + \beta_3 x_3$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 2.2 \end{bmatrix}; p_{H_3}(\mathbf{y}) = \begin{bmatrix} 4.8 \\ 2.6 \\ 0.4 \\ 4.8 \\ 4.8 \end{bmatrix}; \mathbf{y} - p_{H_3}(\mathbf{y}) = \begin{bmatrix} 0.1 \\ 0.4 \\ -0.2 \\ -1.9 \\ 1.6 \end{bmatrix}$$

$$\frac{1}{5-2} \|\mathbf{y} - p_{H_3}(\mathbf{y})\|^2 = \frac{6.38}{3} = 2.1267$$

$$R^2 = \frac{21.868 - 6.38}{21.868} = 70.8\%$$

viii) Model H_0 : $E(Y) = \alpha$

$$\hat{\alpha} = 3.48$$

$$p_{H_0}(\mathbf{y}) = \begin{bmatrix} 3.48 \\ 3.48 \\ 3.48 \\ 3.48 \\ 3.48 \end{bmatrix}; \mathbf{y} - p_{H_0}(\mathbf{y}) = \begin{bmatrix} 1.42 \\ -0.48 \\ -3.28 \\ -0.58 \\ 2.92 \end{bmatrix}$$

$$\frac{1}{5-1} \|\mathbf{y} - p_{H_0}(\mathbf{y})\|^2 = \frac{21.8680}{4} = 5.4670$$

D. Successive testings

1) $H \supseteq H_{12} \supseteq H_1 \supseteq H_0$ i.e. : $\beta_3 = 0, \beta_2 = 0, \beta_1 = 0$

Variation	SS	d.o.f.
$H_0 - H_1$ ($\beta_1 = 0$)	$21.868 - 20.7430 = 1.125$	1
$H_1 - H_{12}$ ($\beta_2 = 0$)	$20.7430 - 18.5126 = 2.230$	1
$H - H_{12}$ ($\beta_3 = 0$)	$18.5126 - 0.6338 = 17.879$	1
$M - \text{obs}$	$0.6338 = 0.634$	1
$H_0 - \text{obs}$	21.868	4

2) $M \supseteq H_{12} \supseteq H_2 \supseteq H_0$ d.v.s. : $\beta_3 = 0, \beta_1 = 0, \beta_2 = 0$

Variation	SS	d.o.f.
$H_0 - H_2$ ($\beta_2 = 0$)	$21.8680 - 21.0429 = 0.825$	1
$H_2 - H_{12}$ ($\beta_1 = 0$)	$21.0429 - 18.5126 = 2.530$	1
$H_{12} - M$ ($\beta_3 = 0$)	$18.5126 - 0.6338 = 17.879$	1
$M - \text{obs}$	$0.6338 = 0.634$	1
$H_0 - \text{obs}$	21.868	4

3) $M \supset H_{13} \supset H_1 \supset H_0$ d.v.s. : $\beta_2 = 0, \beta_3 = 0, \beta_1 = 0$

Variation	SS	d.o.f.
$H_0 - H_1$ ($\beta_1 = 0$)	$21.8680 - 20.7430 = 1.125$	1
$H_1 - H_{13}$ ($\beta_3 = 0$)	$20.7430 - 5.2550 = 15.488$	1
$H_{13} - M$ ($\beta_2 = 0$)	$5.2550 - 0.6338 = 4.621$	1
$M - \text{obs}$	$0.6338 = 0.634$	1
$H_0 - \text{obs}$	21.868	4

4) $M \supseteq H_{13} \supseteq H_3 \supseteq H_0$ d.v.s. : $\beta_2 = 0, \beta_1 = 0, \beta_3 = 0$

Variation	SS	d.o.f.
$H_0 - H_3$ ($\beta_3 = 0$)	$21.8680 - 6.38 = 15.488$	1
$H_3 - H_{13}$ ($\beta_1 = 0$)	$6.38 - 5.2550 = 1.125$	1
$H_{13} - M$ ($\beta_2 = 0$)	$5.2550 - 0.6338 = 4.621$	1
$M - \text{obs}$	$0.6338 = 0.634$	1
$H_0 - \text{obs}$	21.868	4

5) $M \supseteq H_{23} \supseteq H_2 \supseteq H_0$ d.v.s. : $\beta_1 = 0, \beta_3 = 0, \beta_2 = 0$

Variation	SS	d.o.f.
$H_0 - H_2$ ($\beta_2 = 0$)	$21.8680 - 21.0429 = 0.825$	1
$H_2 - H_{23}$ ($\beta_3 = 0$)	$21.0429 - 4.2855 = 16.757$	1
$H_{23} - M$ ($\beta_1 = 0$)	$4.2855 - 0.6338 = 3.652$	1
$M - \text{obs}$	$0.6338 = 0.634$	1
$H_0 - \text{obs}$	21.868	4

6) $M \supset H_{23} \supset H_3 \supset H_0$ d.v.s. : $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0$

Variation	SS	d.o.f.
$H_0 - H_3$ ($\beta_3 = 0$)	$21.8680 - 6.38 = 15.488$	1
$H_3 - H_{23}$ ($\beta_2 = 0$)	$6.38 - 4.2855 = 2.095$	1
$H_{23} - M$ ($\beta_1 = 0$)	$4.2855 - 0.6338 = 3.652$	1
$M - \text{obs}$	$0.6338 = 0.634$	1
$H_0 - \text{obs}$	21.868	4

E. Variance-covariance- and correlation- matrix for data.

$$\text{Variance-covariance matrix} = \frac{1}{5-1} \begin{pmatrix} 2 & 1 & 0 & 1.50 \\ 1 & 2.8 & 0.4 & -1.52 \\ 0 & 0.4 & 3.2 & 7.04 \\ 1.50 & -1.52 & 7.04 & 21.868 \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ y \end{matrix}$$

$$\text{correlation matrix} = \begin{pmatrix} 1 & 0.4225 & 0 & 0.2268 \\ 0.4225 & 1 & 0.1336 & -0.1942 \\ 0 & 0.1336 & 1 & 0.8416 \\ 0.2268 & -0.1942 & 0.8416 & 1 \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ y \end{matrix}$$

F. Partial correlations for given x_3 :

$$\begin{aligned} & \begin{pmatrix} 1 & 0.4225 & 0.2268 \\ 0.4225 & 1 & -0.1942 \\ 0.2268 & -0.1942 & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0.1336 \\ 0.8416 \end{pmatrix} [1]^{-1} \begin{bmatrix} 0 & 0.1336 & 0.8416 \end{bmatrix} \\ &= \begin{pmatrix} 1 & 0.4225 & 0.2268 \\ 0.4225 & 0.9822 & -0.3066 \\ 0.2268 & -0.3066 & 0.2917 \end{pmatrix}, \end{aligned}$$

i.e. the correlation matrix is

$$\begin{pmatrix} 1 & 0.4263 & 0.4199 \\ 0.4263 & 1 & -0.5728 \\ 0.4199 & -0.5728 & 1 \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ y \end{matrix}$$

First calculated using the above mentioned partial correlation matrix

$$\begin{aligned} & \begin{pmatrix} 1 & 0.4199 \\ 0.4199 & 1 \end{pmatrix} - \begin{pmatrix} 0.4263 \\ 0.5728 \end{pmatrix} [1]^{-1} [0.4263 - 0.5728] = \\ & \begin{pmatrix} 0.8183 & 0.6641 \\ 0.6641 & 0.6718 \end{pmatrix}, \end{aligned}$$

which results in the following correlation matrix

$$\begin{pmatrix} 1 & 0.8956 \\ 0.8956 & 1 \end{pmatrix} \begin{matrix} x_1 \\ y \end{matrix}$$

As a check we could compute it from the original covariance matrix

$$\begin{aligned} & \begin{pmatrix} 2 & 1.50 \\ 1.50 & 21.868 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ -1.52 & 7.04 \end{pmatrix} \begin{pmatrix} 2.8 & 0.4 \\ 0.4 & 3.2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -1.52 \\ 0 & 7.04 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 1.50 \\ 1.50 & 21.868 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ -1.52 & 7.04 \end{pmatrix} \begin{pmatrix} 0.3636 & -0.0455 \\ -0.0455 & 0.3182 \end{pmatrix} \begin{pmatrix} 1 & -1.52 \\ 0 & 7.04 \end{pmatrix} \\ &= \begin{pmatrix} 1.6363 & 2.3727 \\ 2.3727 & 4.2855 \end{pmatrix}, \end{aligned}$$

and the partial correlation matrix is then

$$\begin{pmatrix} 1 & 0.8960 \\ 0.8960 & 1 \end{pmatrix} \begin{matrix} x_1 \\ y \end{matrix}$$

The deviations in the elements off the diagonal are a result of truncation errors.

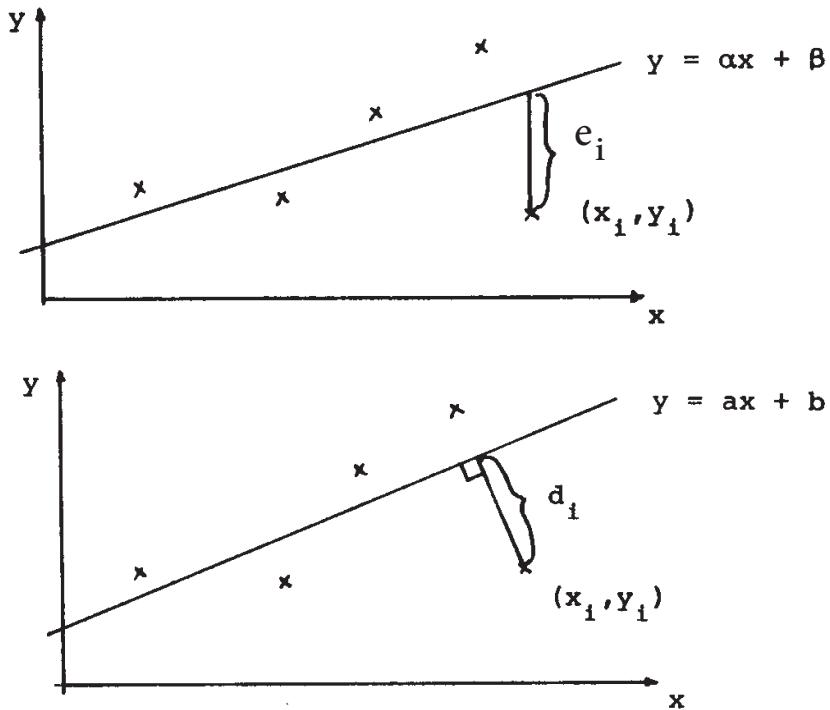


Figure 3.9: In ordinary one-dimensional regression α and β are determined by minimizing $\sum e_i^2$. In orthogonal regression a and b are determined by minimizing $\sum d_i^2$.

3.4 Other regression models and solutions

In this section we shall look at an alternative criterion for estimating a (linear) function of some independent variables. Furthermore we shall consider a linear regularization solution to the normal equations in the case where we have strong multicollinearity between the independent variables, the socalled ridge regression.

3.4.1 Orthogonal regression (linear functional relationship).

In the ordinary least squares estimation of a regression surface we minimise the sum of squares of the vertical distances between the regression surface and the observed points.

Often we will be in the situation that it would be more reasonable to minimise the orthogonal distances and then we talk about *orthogonal regression* (not to be confused with regression by orthogonal polynomials).

Let us assume the following variables μ_1, \dots, μ_p , which satisfy a linear relationship

$$\alpha_0 + \alpha_1\mu_1 + \dots + \alpha_p\mu_p = 0, \quad (3-6)$$

i.e. the variables lie in a hyper plane with the above mentioned equation. We are interested in determining this plane i.e. to determine $\alpha_0, \dots, \alpha_p$. Assume that it is not possible to observe the values μ_1, \dots, μ_p , but only measure

$$X_{ij} = \mu_{ij} + Z_{ij}, \quad j = 1, \dots, p, \quad i = 1, \dots, n,$$

where the Z_{ji} 's are random variables with mean value 0 and where $\mu_{i1}, \dots, \mu_{ip}$, $i = 1, \dots, n$, satisfies 3-6.

Estimation of the parameters α_i on the basis of such a set of observations is often called estimations of a *linear functional relationship* in the literature.

Here it would intuitively reasonable exactly to use the hyper plane which is found by minimising the orthogonal distances down to this. If the Z_{ij} 's are normally distributed with the same variance it can be shown (see e.g. [5] p. 392) that this plane gives the maximum likelihood estimator of the α 's. We formulate the solution to the problem in

|||| Theorem 3.17

We consider n points $x_1, \dots, x_n \in R^p$ and the hyperplane

$$\alpha_0 + \alpha_1x_1 + \dots + \alpha_px_p = 0$$

which minimise the sum of squares of the orthogonal distances from the points. Then $\alpha_1, \dots, \alpha_p$ are the coordinates of a normed eigenvector corresponding to the smmallest eigenvalue of the empirical variance variance-covariance matrix for the n x -points. The last coefficient is given by

$$\alpha_0 = -\alpha_1\bar{x}_1 - \dots - \alpha_p\bar{x}_p.$$

|||| Proof

We write the observations as

$$\begin{bmatrix} x'_1 \\ \vdots \\ x'_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

The distance from a point with the coordinates $x = (x_1, \dots, x_p)'$ to the hyperplane is

$$\left| \frac{\alpha_0 + \alpha_1 x_1 + \dots + \alpha_p x_p}{\sqrt{\alpha_1^2 + \dots + \alpha_p^2}} \right|.$$

Therefore we must determine $\alpha_0, \dots, \alpha_p$ so that

$$f(\alpha) = \sum_{i=1}^n \frac{(\alpha_0 + \alpha_1 x_{i1} + \dots + \alpha_p x_{ip})^2}{\alpha_1^2 + \dots + \alpha_p^2}$$

is minimised. If we introduce a zero'th coordinate x_0 by $x_{i0} = 1, i = 1, \dots, n$, we could write

$$f(\alpha) = \sum_{i=1}^n \left(\sum_{j=0}^p \alpha_j x_{ji} \right)^2 / \sum_{j=1}^p \alpha_j^2.$$

Solving this minimisation problem is equivalent to minimize

$$g(\alpha) = \sum_{i=1}^n \left(\sum_{j=0}^p \alpha_j x_{ij} \right)^2$$

subject to the constraint

$$\sum_{j=1}^p \alpha_j^2 = 1.$$

If we introduce a Lagrange multiplier λ we see that we must determine the global minimum of

$$\varphi(\alpha, \lambda) = \sum_{i=1}^n \left(\sum_{j=0}^p \alpha_j x_{ij} \right)^2 - \lambda \left(\sum_{j=1}^p \alpha_j^2 - 1 \right).$$

The coordinate of the gradient vector are for $v = 1, \dots, p$

$$\frac{\partial \varphi}{\partial \alpha_v} = 2 \sum_{i=1}^n \sum_{j=0}^p \alpha_j x_{ij} x_{iv} - 2\lambda \alpha_v,$$

and for $v = 0$

$$\frac{\partial \varphi}{\partial \alpha_0} = 2 \sum_{i=1}^n \sum_{j=0}^p \alpha_j x_{ij} x_{i0} = 2 \sum_{i=1}^n \sum_{j=0}^p \alpha_j x_{ij}.$$

Putting these partial derivatives = 0, the last equation becomes

$$\sum_{i=1}^n \sum_{j=0}^p \alpha_j x_{ij} = 0,$$

or

$$\alpha_0 = -\alpha_1 \bar{x}_1 - \dots - \alpha_p \bar{x}_p.$$

where

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

If this is inserted in the first set of equations they can be rewritten as

$$\sum_{i=1}^n \sum_{j=1}^p \alpha_j (x_{ij} - \bar{x}_j) (x_{iv} - \bar{x}_v) - \lambda \alpha_v = 0.$$

If we denote the empirical variance covariance-variance matrix for the observations

$$\hat{\Gamma} = (\hat{\gamma}_{jv}),$$

we see that the equations are now rewritten as

$$\sum_{j=1}^p \alpha_j \hat{\gamma}_{jv} - \frac{\lambda}{n-1} \alpha_v = 0, \quad v = 1, \dots, p.$$

If we let

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \boldsymbol{\alpha},$$

then the equations system can be written as

$$\hat{\Gamma} \boldsymbol{\alpha} = \frac{\lambda}{n-1} \boldsymbol{\alpha},$$

i.e. $\boldsymbol{\alpha}$ is an eigenvector of $\hat{\Gamma}$ corresponding to the eigenvalue $\frac{\lambda}{n-1}$.

The question is now, which of the p eigenvalues for $\hat{\Gamma}$ should be chosen. After some manipulations with the original equations it follows that we must choose the smallest eigenvalue. This concludes the proof of the theorem.

■

|||| Remark 3.18

The result which has been stated in the theorem has a close connection to the results which will be shown in a later chapter on principal components.

3.4.2 Regularization and Ridge Regression

In the analysis of regression models one often will have stability problems if the design matrix is ill-conditioned. This may be detected by suitable regression diagnostics. If we have a model with many independent variables a solution

to this problem may be various stepwise regression procedures. This will however not always work satisfactorily, so instead of excluding some variables and focus completely on others one might try utilize the information available in all independent variables in a different way.

We consider the usual model

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where \mathbf{x} is a known $n \times p$ matrix, $\boldsymbol{\beta}$ the unknown parameter vector and $\boldsymbol{\varepsilon}$ the error vector.

We assume that

$$\begin{aligned} E(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ D(\boldsymbol{\varepsilon}) &= \sigma^2 \mathbf{I}_n. \end{aligned}$$

The ordinary least squares estimator is - assuming that \mathbf{x} has maximum rank -

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}.$$

Furthermore we assume that the independent variables are scaled so that $\mathbf{x}'\mathbf{x}$ has correlation form (i.e. the single independent variables are reduced with their average and divided with their standard deviation). This normalization will help make the estimates more stable numerically. This normalization is often recommendable in a practical situation.

If $\mathbf{x}'\mathbf{x}$ in this form is close to a unity matrix, i.e. if the independent variables are near-orthogonal, the least squares estimator is fine. If we have *multicollinearity*, i.e. if the independent variables are strongly correlated, the estimates $\hat{\boldsymbol{\beta}}$ will be unstable.

We now analyze some properties of $\hat{\boldsymbol{\beta}}$ that has not been dealt with earlier. We have

$$D(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}.$$

If we put L equal to the distance from $\hat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}$,

$$L^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

we get

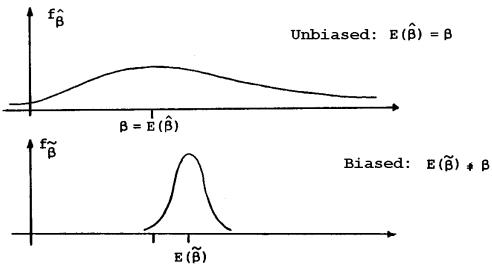
$$E(L^2) = \sum_{i=1}^p E[(\hat{\beta}_i - \beta_i)^2] = \sum_{i=1}^p V(\hat{\beta}_i) = \sigma^2 \text{tr}(\mathbf{x}'\mathbf{x})^{-1}.$$

Since

$$L^2 = \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}'\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{\beta},$$

we get that the expected value of the squared length of $\hat{\boldsymbol{\beta}}$ is

$$E(\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}) = \sigma^2 \text{tr}(\mathbf{x}'\mathbf{x})^{-1} + \boldsymbol{\beta}'\boldsymbol{\beta}.$$



If we denote the eigenvalues for $\mathbf{x}'\mathbf{x}$

$$\lambda_1 \geq \cdots \geq \lambda_p,$$

we obtain (accordingly to Theorem A.25 and the results on p. 376)

$$E(L^2) = \sigma^2 \left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_p} \right) > \frac{\sigma^2}{\lambda_p},$$

and

$$E(\hat{\beta}'\hat{\beta}) = \sigma^2 \left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_p} \right) + \beta'\beta > \frac{\sigma^2}{\lambda_p} + \beta'\beta.$$

If the independent variables are strongly correlated the eigenvalues of $\mathbf{x}'\mathbf{x}$ will vary a lot and consequently the smallest will be very small ($<< 1$). According to the relations above the squared distance between β and $\hat{\beta}$ will in this case have a large expected value, and the squared length of $\hat{\beta}$ will have an expectation by far exceeding the squared length of β .

This tendency to 'inflation' of $\hat{\beta}$ is caused by requiring unbiasedness of β . The question is therefore whether we by relaxing on this requirement may obtain estimates that in some sense are closer to β . The problem is been sketched in the figure below.

Here we may again refer to the *mean squared error* of an estimator $\tilde{\beta}$ (in the one-dimensional case)

$$\text{MSE}(\tilde{\beta}) = E[(\tilde{\beta} - \beta)^2] = V(\tilde{\beta}) + \{E(\tilde{\beta}) - \beta\}^2,$$

i.e. that the MSE of an estimator is equal to the variance plus the squared bias. If we therefore by allowing a small bias may obtain a great reduction in the variance, this would obviously be preferable. This is exactly what is obtained with the ridge estimator introduced in the definition below.

|||| Definition 3.19

A ridge estimator for β in the model

$$Y = x\beta + \varepsilon$$

is an estimator $\hat{\beta}_k^* = \hat{\beta}^*$, that is a solution to

$$(x'x + k \cdot I)\hat{\beta}^* = x'Y,$$

i.e.

$$\hat{\beta}^* = (x'x + k \cdot I)^{-1}x'Y.$$

Here k is a constant $\in [0, 1]$.

|||| Remark 3.20

In numerical mathematics such way of solving the normal equations is called a (Tikhonov) *regularization*. This is a very common way of solving ill-posed problems.

We are now listing some properties of $\hat{\beta}^*$. These properties are among other things used in determining k , a quantity not given in beforehand.

We have

|||| Theorem 3.21

Let the situation be like in the definition. We put $x'y = g$ and denote the observed residual sum of squares for an arbitrary estimator $\tilde{\beta}$ equal to

$$H(\tilde{\beta}) = (y - x\tilde{\beta})'(y - x\tilde{\beta}).$$

Then the gradient of H in $\tilde{\beta} = 0$ is proportional to g and has the opposite direction of g . $\hat{\beta}_k^*$ may be determined by, that it for a fixed length minimizes $H(\tilde{\beta})$, i.e.

$$\min_{\|\tilde{\beta}\|=\|\hat{\beta}_k^*\|} H(\tilde{\beta}) = H(\hat{\beta}_k^*).$$

Furthermore $H(\hat{\beta}_k^*)$ is an increasing function of k . The length of $\hat{\beta}_k^*$ is decreasing in k , and the angle γ between $\hat{\beta}_k^*$ and g is decreasing in k .

|||| Proof

Not very complicated but is omitted. The reader is referred to [8] and [9]. ■

The instructive figure 3.4.2 is taken from [10]

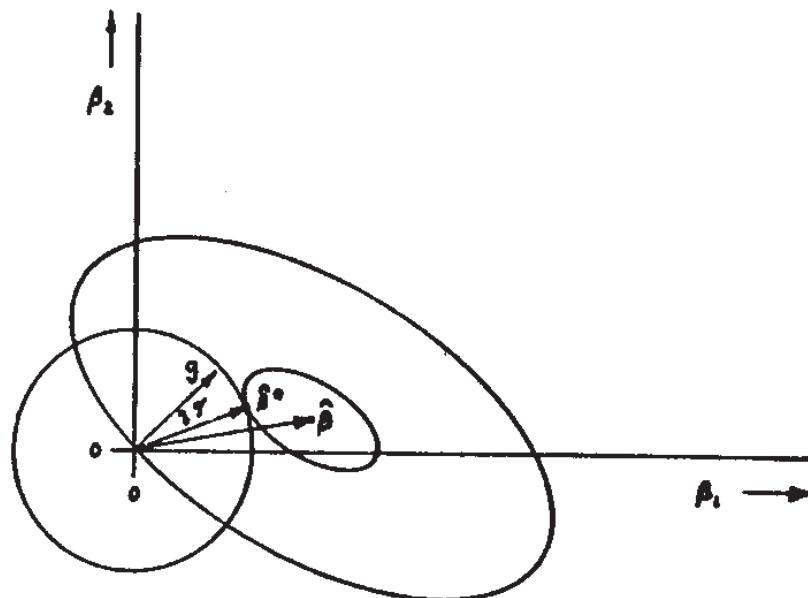


Figure 3.4.2.

It depicts the situation geometrically in the case $p = 2$. The point $\hat{\beta}$ in the center of the ellipses is the least squares solution. The ellipses are level curves for H . The circle with center in origo is a tangent to the small ellipse. We see that $\hat{\beta}^*$ is the shortest vector that has a residual sum of squares as small as the value of H on the small ellipse. Furthermore we see that $\hat{\beta}^*$ always lies between $\hat{\beta}$ and g .

Other properties of the ridge estimator are given in the following theorem

|||| Theorem 3.22

Let the situation be as described above. Then $\hat{\beta}_k^* = \hat{\beta}^*$ is a linear transformation of $\hat{\beta}$ since

$$\hat{\beta}^* = \mathbf{z}_k \hat{\beta} = (\mathbf{x}' \mathbf{x} + k \mathbf{I})^{-1} (\mathbf{x}' \mathbf{x}) \hat{\beta}.$$

$\hat{\beta}^*$ is biased since

$$E(\hat{\beta}^*) = \mathbf{z}_k \beta.$$

The dispersion matrix of $\hat{\beta}^*$ is

$$D(\hat{\beta}^*) = \sigma^2 [\mathbf{x}' \mathbf{x} + k \mathbf{I}]^{-1} (\mathbf{x}' \mathbf{x}) [\mathbf{x}' \mathbf{x} + k \mathbf{I}]^{-1},$$

and the expected squared distance to β is

$$E[(\hat{\beta}^* - \beta)' (\hat{\beta}^* - \beta)] = \text{tr}(D(\hat{\beta}^*)) + \beta' (\mathbf{z}_k - \mathbf{I})' (\mathbf{z}_k - \mathbf{I}) \beta.$$

In the last expression the first term is equal to the variance of the squared length of $\hat{\beta}^*$ and the last term is equal to the squared bias.

|||| Proof

Omitted. Follows by elementary matrix manipulations.

■

From the theorem follows an important corollary

|||| Corollary 3.23

If $\beta' \beta$ is limited there exists a $k > 0$, so that the expected squared distance between β and $\hat{\beta}^*$ is strictly smaller than the expected distance between β and $\hat{\beta}$.

|||| Proof

This follows by noting that $\text{tr}(D(\hat{\beta}^*))$ is decreasing in k whereas $\beta' (\mathbf{z}_k - \mathbf{I})' (\mathbf{z}_k - \mathbf{I}) \beta$ is increasing in k . Since $k \rightarrow 0 \Rightarrow \hat{\beta}^* \rightarrow \hat{\beta}$ the result follows immediately.

The only remaining problem is determining a reasonable k . Historically the so-called *ridge trace* is used. There are other alternatives (see e.g. [11] for results on using *cross validation*), but the ridge trace is a straight forward method.

||| **Definition 3.24**

By the ridge trace we understand the mapping of the individual coefficients in the ridge estimator as a function of k .

||| **Remark 3.25**

The philosophy behind using the ridge trace in determining k is a sensitivity argument. From the ridge trace it follows which coefficients that are sensitive to variations in k . One then selects the smallest value of k giving a stable sequence of the coefficients.

We illustrate the principles in the next example.

||| Example 3.26

([10]). In the example we consider the relation between ASTM and gas chromatography distillation of gasoline. We shall not dwell on the differences but only state that gas chromatography is far more precise in assessing the volatility than the ASTM standard used in 1975. It is therefore of interest to use gas chromatography in controlling the distillation but at the same time being able to state what ASTM result this would correspond to. So we must predict an ASTM result based on gas chromatography measurements.

The ASTM method involves measuring the volatility as the fraction that has evaporated at different temperature levels. We shall only consider one single ASTM temperature, namely 158°F. The fraction evaporated at this temperature is called y .

We now want to predict y based on determinations by gas chromatography of fractions evaporated at 15 different temperature levels. Those fractions are called x_1, \dots, x_{15} . We apply a linear model

$$E(y) = \beta_1 x_1 + \dots + \beta_{15} x_{15}.$$

The independent variables add up to 1 and therefore we have not introduced an intercept.

The main use of the model should be partly predictions of ASTM distillations of gasoline, and it was requested that the prediction standard deviation should be smaller than 1.5%. Furthermore the results should be used as input for determining optimal mixing procedures by means of linear programming. Earlier results using ordinary least squares and stepwise regression procedures had given coefficients that were unacceptable from a physical/chemical point of view.

There were 59 observations of the 16 variables. These were split in two parts - one consisting of 29 observations used for estimation, and one consisting of 30 used for determining the prediction error. In the next figure we have shown the ridge trace. It is seen that the system stabilizes for k -values around $k = 0.005 - 0.01$. In the case presented $k = 0.006$ was chosen.

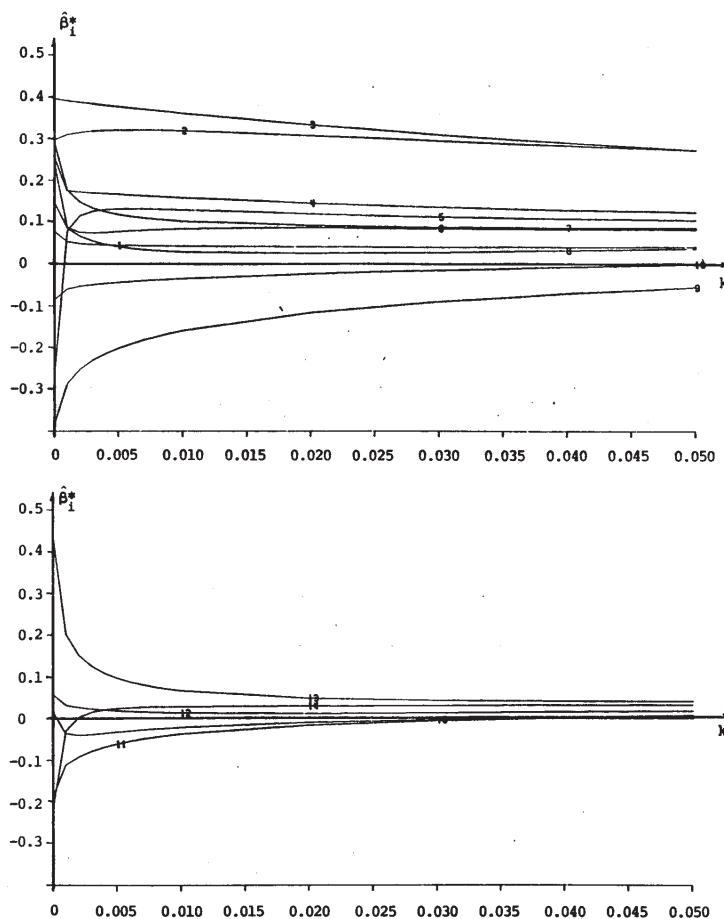


Figure 3.26: Ridge trace for the data from example 3.26. The coefficients are shown in a relative scale (\bar{Y} is equal to 0).

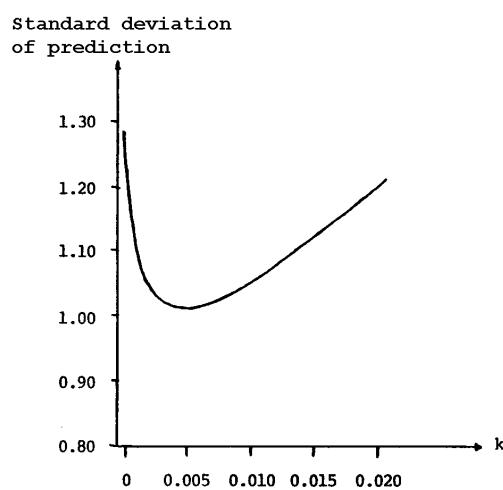


Figure 3.26: Prediction standard deviation as a function of k .

In figure 3.26 we show the prediction standard deviation found for the different values of k . Note that the minimum occurs for $k = 0.006$, the value obtained by

analysing the ridge trace. The least squares analysis has a prediction standard deviation of 1.28 and the ridge model 1.01. In the following figure 3.26 the coefficients are compared with coefficients resulting from the theoretical considerations. These are of course not necessarily the 'truth' but of course you expect a considerable similarity.

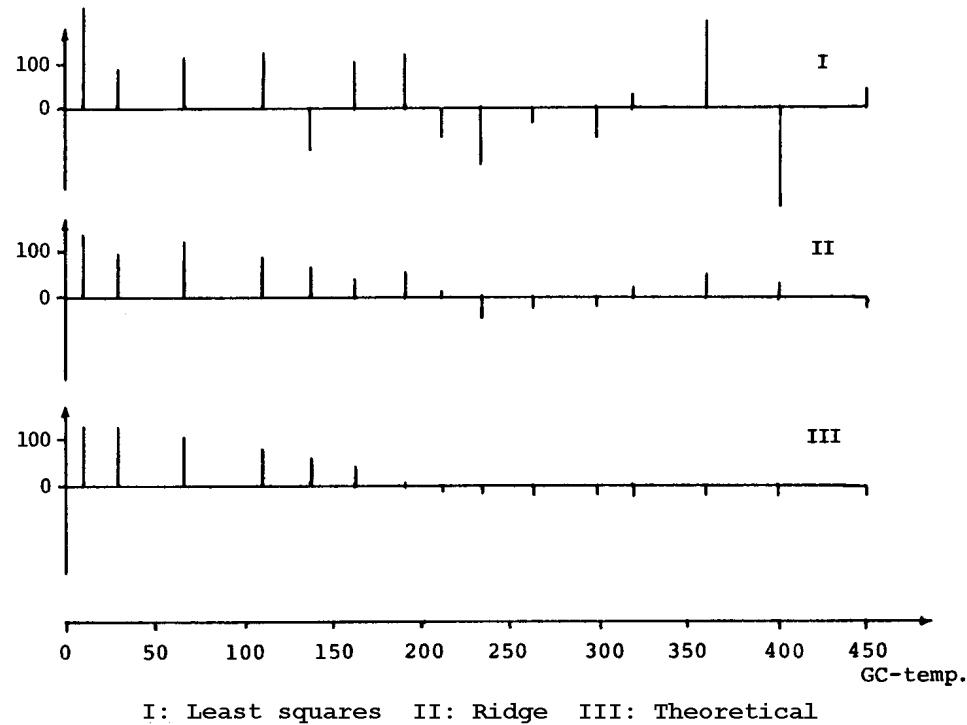


Figure 3.26: Comparison of coefficients.

It is seen that the ridge estimates show a more steady progression with increasing gas chromatography temperatures. Furthermore they are closer to the theoretical values.

3.4.3 Non-linear regression and curve fitting.

Often we will have to analyse regression situations which give rise to non linear normal equations or likelihood equations.

We could of course then use a general programme for maximising non linear functions or an iterative method to solve the non linear equations. The variance covariance matrix for the estimates can then be estimated using the inverse information matrix. We shall not pursue this further in the present context.

We give a few examples.

||| Example 3.27

We consider some data which concern conserving of iron items from the Iron Age. The data are from Eva Salomonsen (1977) [12]. On the National Museum's conservation laboratory they have for 63 years used Rosenberg's annealing method to remove chlorides from the iron. In order to investigate the effectiveness of this method 295 iron items conserved in the years 1913-1974, have been investigated and the number of defect items i.e. items where a continuing disintegration has been found is summed up for each year. The numbers are shown below.

Period	Number investigated	number defects	number of defects in % of investigated.
1913	52	14	26.9
1921-24	34	11	32.4
1933-34	53	10	18.9
1940-43	47	13	27.7
1953-54	56	4	7.1
1961-69	46	4	8.7
1972-74	7	0	0
Total	295	56	19.0

Number of defects

annealed iron items in comparison to the total amount investigated for each specific year.

As seen in the table the defect percent grows with time and this growth is what we want to model. A reasonable model would be to let

$$\begin{aligned} X_i &= \text{number of defect for age } t_i \\ n_i &= \text{number of investigated for age } t_i \\ p_i &= \text{the probability of an item with age } t_i \text{ being defect} \end{aligned}$$

and then claiming that

$$X_i \in B(n_i, p_i)$$

i.e. binomially distributed with parameters (n_i, p_i) .

As age we of course choose the time which has elapsed since the annealing treatment. For the periods, which cover several years the annealing time has been set to the middle of the considered time interval.

The remaining problem is to find the dependence of the defect percent p_t of time. Here a very often used model is the logistic curve:

$$p_i = p(t_i) = \frac{1}{1 + \exp(-\alpha - \beta t_i)}.$$

The curve has asymptotes in $p = 0$ and $p = 1$ and is continuously growing so it of course satisfies the basic requirements we might have. If we define the so called logit

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i},$$

we find that

$$\text{logit}(p_i) = \alpha + \beta t_i,$$

i.e. that the model is linear in these logits. The model has been used quite a lot in connection with bioassays especially by Berkson.

The likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^n \binom{n_i}{x_i} \left\{ \frac{1}{1 - \exp(-\alpha - \beta t_i)} \right\}^{x_i} \left\{ \frac{\exp(-\alpha - \beta t_i)}{1 + \exp(-\alpha - \beta t_i)} \right\}^{n_i - x_i}$$

and then

$$\begin{aligned} \ln L(\alpha, \beta) &= \\ &= c - \sum_i x_i \ln(1 + \exp(-\alpha - \beta t_i)) - \sum_i (n_i - x_i)(\alpha + \beta t_i) \\ &\quad - \sum_i (n_i - x_i) \ln(1 + \exp(-\alpha - \beta t_i)) \\ &= c - \sum_i n_i \ln(1 + \exp(-\alpha - \beta t_i)) - \sum_i (n_i - x_i)(\alpha + \beta t_i). \end{aligned}$$

We can now either differentiate this expression with respect to α, β and let the differential coefficients equal 0 or we can maximise $\ln L(\alpha, \beta)$ by means of a general minimising program. By doing the first the following estimates are found

$$\begin{aligned} \hat{\alpha} &= -2.99984 \\ \hat{\beta} &= 0.03813 \end{aligned}$$

The resulting logistic curve has been drawn in figure 3.27. Furthermore 95% confidence intervals around the single observations have been drawn.

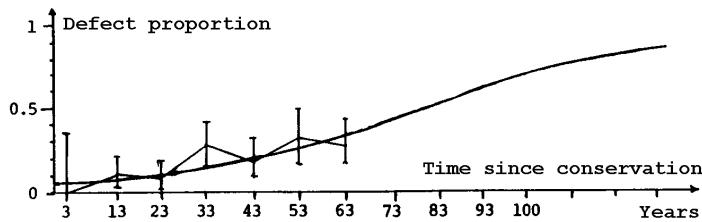


Figure 3.27: Defect proportion for annealed iron items.

A fair agreement is found but we should, however, be careful to extrapolate hundreds of years into the future even though the figure might lead us to.

Often we are in the situation where we wish to fit a given data with a suitable smooth curve but where there does not seem to be any possibility for it is very difficult to determine a universal law which can cover all of the considered area. We could then think of performing a piecewise approximation with different functions. Here a very appropriate class is the so called *spline functions* which are named after a special drawing device. These are introduced in

|||| **Definition 3.28**

Let there be given an interval $[a, b]$ and observations or points x_1, \dots, x_n which all lie in the interval. Furthermore let there be a value y_i for each x_i . As a spline of order $2m-1$ with knots x_1, \dots, x_n we will understand the function φ which satisfies

- 1) φ is a polynomial of degree $2m-1$ in $[x_i, x_{i+1}]$
- 2) φ is a polynomial of degree $m - 1$ in $[a, x_1]$ and $[x_n, b]$
- 3) $\varphi, \varphi', \dots, \varphi^{(2m-2)}$ are continuous in x_1, \dots, x_n
- 4) $\varphi(x_j) = y_j$

|||| **Remark 3.29**

A spline of order $2m-1$ is put together or smoothly of $(2m-1)$ degree polynomials. It can be shown that the obtained curve very much resembles the one we would obtain by nailing two nails into each knot and then forcing a very elastic steel rug through these (a so called drawing spline). We will not pursue this further but just refer the reader to the literature on the area.

||| Example 3.30

In figure 3.30 the level for a number of points on a line of length 5 km in Dyrehaven are shown. The data is made available by Poul Frederiksen, Department of Surveying, DTU. In different projects it is interesting to compute an expression for a variation around a suitable chosen smooth trend curve. It is therefore obvious to use a cubic spline function. This is done by means of the Harwell programme VB05B which on the basis of observations $(x_1, y_1), \dots, (x_m, y_m)$ minimises

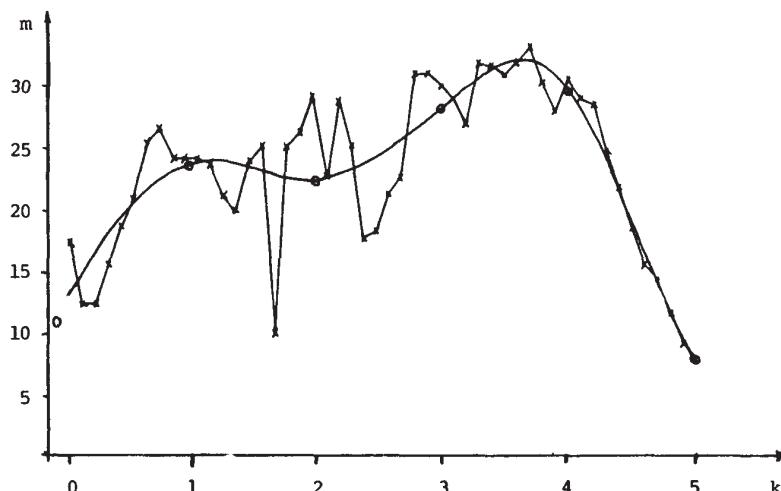


Figure 3.30: Levels

and approximating cubic splines.

$$F = \sum_{i=1}^m w_i^2 \{y_i - S(x_i)\}^2,$$

where w_i 'er are users specified weights and S is a cubic spline function with knots $j, j=1, \dots, n$, which abscesses are specified by the users.

The resulting spline and its knots are also given. We note the very nice fit between the very irregular observations and the spline function.

|||| Chapter 4

Tests in the multidimensional normal distribution

In this chapter we will give a number of generalisations to some of the well known test statistics based on one dimensional normally distributed random variables. In most cases the test statistics will be analogues to the well known ones, except for multiplication being substituted with matrix multiplication, numerical values by the determinant of the matrix etc.

4.1 Test for mean value.

4.1.1 Hotelling's T^2 in the One-Sample Situation

In this section we will consider independent random variables X_1, \dots, X_n , where

$$X_i \in N_p(\mu, \Sigma),$$

i.e. p-dimensionally normally distributed with mean vector μ and variance-covariance matrix Σ . We assume that Σ is regular and unknown. We want to test a hypothesis about the mean vector μ being equal to a given vector μ_0 against all alternatives i.e.

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0$$

We first repeat some results on the estimators. From theorem 1.55 p. 54 we have the following results on the empirical mean vector \bar{X} and the empirical

variance-covariance matrix \mathbf{S}

$$\begin{aligned}\bar{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i && \in \mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n} \boldsymbol{\Sigma}) \\ \mathbf{S} &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' && \in \mathcal{W}(n-1, \frac{1}{n-1} \boldsymbol{\Sigma})\end{aligned}$$

$\bar{\mathbf{X}}$ and \mathbf{S} are stochastically independent.

In the following we will furthermore need the following results on the distribution of certain functions of normally distributed and Wishart distributed stochastic variables.

||| Lemma 4.1

Let \mathbf{Y} be a p -dimensional stochastic variable and let \mathbf{U} be a $p \times p$ stochastic matrix with

$$\begin{aligned}\mathbf{Y} &\in \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ m\mathbf{U} &\in \mathcal{W}(m, \boldsymbol{\Sigma}),\end{aligned}$$

furthermore let \mathbf{Y} and \mathbf{U} be stochastically independent. We now let

$$T^2 = \mathbf{Y}'\mathbf{U}^{-1}\mathbf{Y}.$$

Then the following holds

$$\frac{m-p+1}{mp} T^2 \in \mathcal{F}(p, m-p+1; \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}),$$

i.e. the left hand side is non-centrally F-distributed with non-centrality parameter $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ and degrees of freedom equal to $(p, m-p+1)$. If $\boldsymbol{\mu} = \mathbf{0}$, then the non-centrality parameter is 0 i.e. we then have the special case

$$\frac{m-p+1}{mp} T^2 \in \mathcal{F}(p, m-p+1).$$

||| Proof

Omitted. See e.g. [2], p. 106.

We now have the following main result

|||| Theorem 4.2

We will use the notation

$$T^2 = n(\bar{\mathbf{X}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{X}} - \boldsymbol{\mu}_0),$$

where $\bar{\mathbf{X}}$, $\boldsymbol{\mu}_0$ and \mathbf{S} are as stated in the introduction to this section. Then the critical area for a ratio test of H_0 against H_1 at level α is

$$C = \{x_1, \dots, x_n \mid \frac{n-p}{(n-1)p} T^2 > F(p, n-p)_{1-\alpha}\},$$

where T^2 is the observed value of T^2 .

|||| Proof

From Lemma 6.1 we find that

$$\frac{n-p}{(n-1)p} T^2 \in F(p, n-p)$$

under H_0 . From this follows that C is the critical region for a test of H_0 versus H_1 at level α . That this corresponds to a ratio test follows from direct computation by using theorem A.7 among other things.

■

|||| Remark 4.3

The quantity T^2 is often called Hotelling's T^2 after Harold Hotelling, who first considered this test statistic.

|||| Remark 4.4

In the one dimensional case we use the test statistic

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}.$$

We now have that Z^2 can be written

$$Z^2 = n(\bar{X} - \mu_0)[S^2]^{-1}(\bar{X} - \mu_0),$$

i.e. precisely the same as T^2 reduces to in the one-dimensional case. Furthermore note that the square of a student distributed variable $t(\nu)$ is $F(1, \nu)$ distributed which means that there (of course) also is a relation between the distribution of the two test statistics.

In order to compute the test statistic it is useful to remember the follow theorem where it is seen that inversion of a matrix can be substituted by the calculation of some determinants.

|||| Theorem 4.5

Let the notation be as above. Then the following holds

$$T^2 = \frac{\det[\mathbf{S} + n(\bar{X} - \mu_0)(\bar{X} - \mu_0)']}{\det[\mathbf{S}]} - 1$$

|||| Proof

Omitted. Purely technical and follows by using theorem A.7 p. 349 on the matrix

$$\begin{bmatrix} -1 & \sqrt{n}(\bar{X} - \mu_0)' \\ \sqrt{n}(\bar{X} - \mu_0) & \mathbf{S} \end{bmatrix}$$

We now give an illustrative

||| Example 4.6

In the following table values for silicium and aluminium (in %) in 7 samples collected on the moon are given

	Sample						
	1	2	3	4	5	6	7
Silicium	19.4	21.5	19.2	18.4	20.6	19.8	18.7
Aluminium	5.9	4.0	4.0	5.4	6.2	5.7	6.0

We are now very interested in testing if these samples can be assumed to come from a population with the same mean values as basalt from our own planet earth. These are

$$\mu_0 = \begin{pmatrix} 22.10 \\ 7.40 \end{pmatrix}.$$

It seems sensible to use Hotelling's T^2 to help answer the above question. If we call the observations x_1, \dots, x_7 , we find

$$\bar{x} = \begin{pmatrix} 19.66 \\ 5.31 \end{pmatrix},$$

$$\mathbf{s} = \begin{pmatrix} 1.1795 & -0.3076 \\ -0.3076 & 0.8681 \end{pmatrix}.$$

Since

$$\bar{x} - \mu_0 = \begin{pmatrix} -2.44 \\ -2.09 \end{pmatrix},$$

then

$$n(\bar{x} - \mu_0)(\bar{x} - \mu_0)' = \begin{pmatrix} 41.68 & 35.70 \\ 35.70 & 30.58 \end{pmatrix},$$

and

$$\mathbf{s} + n(\bar{x} - \mu_0)(\bar{x} - \mu_0)' = \begin{pmatrix} 42.86 & 35.39 \\ 35.39 & 31.45 \end{pmatrix}.$$

Then

$$t^2 = \frac{95.49}{0.9293} - 1 = 101.75.$$

The F-test statistic is

$$\frac{7-2}{6 \cdot 2} t^2 = 42.8 > F(2, 5)_{0.999} = 37.1,$$

and the hypothesis is therefore rejected at least at all levels α larger than 0.1%. It therefore does not seem reasonable to assume that the 7 moon samples originate from a population with the same mean value of silicium and aluminium as basalt from our planet earth.

From the result of theorem 4.2 we can easily construct a confidence region for μ . We have with the usual notation

|||| Theorem 4.7

A $(1 - \alpha)$ -confidence region for the expectation $E(X)$ is

$$\{\mu | n(\bar{x} - \mu)'s^{-1}(\bar{x} - \mu) \leq \frac{(n-1)p}{n-p} F(p, n-p)_{1-\alpha}\},$$

i.e. an ellipsoid with centre in \bar{x} and main axes determined by the eigenvectors in the inverse empirical variance-covariance matrix.

|||| Proof

Trivial from the definition of a confidence area and theorem 4.2.

■

We now continue example 4.6 in the following

|||| Example 4.8

We will now determine a 95% confidence area for the mean vector. According to theorem 4.7 the confidence area is ordered by the ellipse

$$7(19.66 - \mu_1, 5.31 - \mu_2)s^{-1} \begin{pmatrix} 19.66 - \mu_1 \\ 5.31 - \mu_2 \end{pmatrix} = \frac{12}{5}F(2, 5)_{0.95}$$

or

$$(19.66 - \mu_1, 5.31 - \mu_2)s^{-1} \begin{pmatrix} 19.66 - \mu_1 \\ 5.31 - \mu_2 \end{pmatrix} = 1.9851.$$

We find

$$s^{-1} = \begin{pmatrix} 0.9341 & 0.3310 \\ 0.3310 & 1.2692 \end{pmatrix}$$

with the eigenvalues 1.4727 and 0.7307 and the corresponding (normed) eigenvectors

$$\begin{pmatrix} 0.5236 \\ 0.8520 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -0.8520 \\ 0.5236 \end{pmatrix}.$$

In the coordinate system with origin in \bar{x} and the above mentioned vectors as unity vectors the ellipse has the equation

$$1.4727y_1^2 + 0.7307y_2^2 = 1.9851$$

or

$$\frac{y_1^2}{1.1610^2} + \frac{y_2^2}{1.6482^2} = 1$$

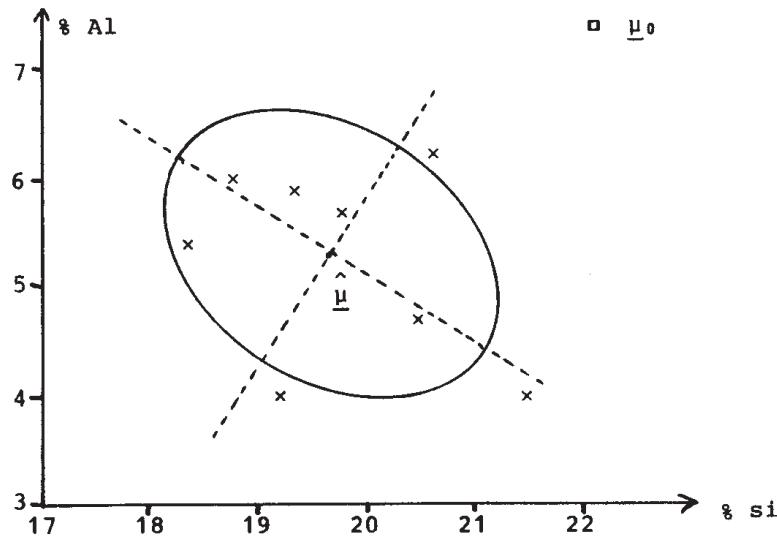


Figure 4.8: Observations and confidence region.

In figure 4.8 the confidence region and the observations are shown. Furthermore $\mu_0 = (22.10, 7.40)'$ is given. It is seen that this observation lies outside the confidence region corresponding to the hypothesis $\mu = \mu_0$ against $\mu \neq \mu_0$ being rejected at all levels greater than 0.01% and therefore especially for $\alpha = 5\%$.

4.1.2 Hotelling's T^2 in the two-sample situation.

Quite analogous to the t-test in the one dimensional case Hotelling's T^2 can be used to investigate if samples from two normal distributions (with the same variance-covariance structure) can be assumed to have the same expected values. We consider independent stochastic variables X_1, \dots, X_n and Y_1, \dots, Y_m , where

$$\begin{aligned} X_i &\in N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ Y_i &\in N_p(\boldsymbol{\nu}, \boldsymbol{\Sigma}), \end{aligned}$$

and we wish to test

$$H_0 : \mu = \nu \quad \text{against} \quad H_1 : \mu \neq \nu.$$

We use the notation

$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \\ \bar{Y} &= \frac{1}{m} \sum_{i=1}^m Y_i \\ \mathbf{S}_1 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' \\ \mathbf{S}_2 &= \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})(Y_i - \bar{Y})' \\ \mathbf{S} &= \frac{(n-1)\mathbf{S}_1 + (m-1)\mathbf{S}_2}{n+m-2}\end{aligned}$$

From theorem 1.55 and theorem 1.54 we have

$$\begin{aligned}\bar{X} &\in N_p(\mu, \frac{1}{n}\Sigma) \\ \bar{Y} &\in N_p(\nu, \frac{1}{m}\Sigma) \\ \mathbf{S} &\in W(n+m-2, \frac{1}{n+m-2}\Sigma).\end{aligned}$$

We now give the main result on testing H_0 against H_1 in

||| Theorem 4.9

We use the same notation as given above. Now, let

$$T^2 = \frac{nm}{n+m} (\bar{X} - \bar{Y})' \mathbf{S}^{-1} (\bar{X} - \bar{Y}).$$

Then the critical region for a test of H_0 against H_1 at level α is equal to

$$C = \{x_1, \dots, x_n, y_1, \dots, y_m \mid \frac{n+m-p-1}{(n+m-2)p} t^2 > F(p, n+m-p-1)_{1-\alpha}\}$$

Here t^2 is the observed value of T^2 .

|||| Proof

From lemma 5.1 and from the above mentioned relationships we find that

$$\frac{n+m-p-1}{(n+m-2)p} T^2 \in F(p, n+m-p-1; (\mu - \nu)' \Sigma^{-1} (\mu - \nu)),$$

and the result follows readily. ■

Analogous to the one-sample situation we can use the results to determine a confidence region for the difference between mean vectors. We have

|||| Theorem 4.10

We still consider the above mentioned situation and let $\mu - \nu = \delta_o$. Then a $(1 - \alpha)$ confidence region for δ_o is equal to

$$\{\delta | \frac{nm}{n+m} (\bar{x} - \bar{y} - \delta)' s^{-1} (\bar{x} - \bar{y} - \delta) \leq \frac{(n+m-2)p}{n+m-p-1} F(p, n+m-p-1)_{1-\alpha}\}.$$

|||| Proof

Follows directly from the definition of a confidence region and from theorem 4.9. ■

|||| Remark 4.11

The confidence region is an ellipsoid with centre in $\bar{x} - \bar{y}$ and main axes determined by the eigenvectors in s^{-1} .

|||| Remark 4.12

As mentioned the test results and confidence intervals require that the variance-covariance matrices for the X - and for the Y -observations are equal. If this is not the case the above mentioned results are not exact and a different procedure should be used. We will not consider this here but refer to e.g. [2], p. 118.

We will now consider an example on the use of T^2 in the two-sample situation.

|||| Example 4.13

At the Laboratory of Heating- and Climate-technique, DTU, one has measured the following in an experiment

- i) the height in cm.
- ii) evaporation loss in g/m^2 skin during a 3 hour period
- iii) mean temperature in $^\circ\text{C}$. This temperature is found by measuring the skin temperature at 14 different locations every minute for 5 minutes (same locations every time). The mean temperature is then an average of all $14 \times 5 = 70$ measurements,

on 16 men and 16 women. The result of the experiment is given in the table p. 192.

Person No.	Height in cm	Evaporation loss in g/m ² skin	Mean temperature in °C
1	177	18.1	33.9
2	189	18.8	33.2
3	181	20.4	33.9
4	184	19.5	33.8
5	183	30.5	33.3
6	178	22.2	33.6
7	162	19.4	39.2
8	176	26.7	33.2
9	190	16.6	33.2
10	180	45.4	33.5
11	179	24.0	33.9
12	175	34.6	33.8
13	183	21.3	33.5
14	177	33.3	33.9
15	185	22.9	33.8
16	176	18.6	33.5
1	160	14.6	32.9
2	171	27.0	33.5
3	168	27.6	32.3
4	171	20.2	33.1
5	169	30.8	33.4
6	169	17.4	33.5
7	167	21.1	33.0
8	170	19.3	34.1
9	162	21.5	33.8
10	160	15.2	33.0
11	168	15.4	33.7
12	157	25.2	33.9
13	161	13.9	34.8
14	164	20.2	31.9
15	161	25.3	39.0
16	180	12.6	33.5

Table 4.13: Data from indoor-climate experiments, Laboratory for Heating- and Climate-technique, DTU.

We consider these numbers as realisations of stochastic variables

$$X_1, \dots, X_{16} \quad \text{and} \quad Y_1, \dots, Y_{16}.$$

We furthermore assume, that the variables are stochastic independent and that

$$X_i \in N(\mu, \Sigma)$$

and

$$Y_i \in N(\nu, \Sigma),$$

i.e. the variance-covariance matrices are assumed equal. Later we will discuss whether this hypothesis is reasonable or not.

The estimates for μ and ν are the empirical mean vectors i.e.

$$\hat{\mu} = \bar{x} = \begin{pmatrix} 179.7 \\ 24.5 \\ 33.6 \end{pmatrix}$$

and

$$\hat{\nu} = \bar{y} = \begin{pmatrix} 166.1 \\ 20.5 \\ 33.4 \end{pmatrix}.$$

We will now check if the difference between $\hat{\mu}$ and $\hat{\nu}$ is significant, i.e. whether μ and ν can be assumed equal.

With the notation chosen in theorem 4.9 we find

$$\mathbf{s} = \begin{pmatrix} 38.5 & -4.3 & -0.8 \\ -4.3 & 45.5 & -0.3 \\ -0.8 & -0.3 & 0.3 \end{pmatrix},$$

and

$$t^2 = \frac{16 \cdot 16}{16 + 16} (\bar{x} - \bar{y})' \mathbf{s}^{-1} (\bar{x} - \bar{y}) = 52.4.$$

The test statistic then becomes

$$\frac{16 + 16 - 3 - 1}{(16 + 16 - 2)3} 52.4 = 16.3.$$

Since

$$F(3, 28)_{0.999} = 7.19$$

a hypothesis that $\mu = \nu$ will at least be rejected at all levels greater than 0.1%. We will therefore conclude that there is a fairly large (simultaneous) difference in the three variables for men and for women, a result which probably will not shock anyone when it is remembered that the first variable gives the height.

If we instead only consider the second and third coordinates, i.e. the values for evaporation loss and mean temperature we get the test statistic

$$\frac{16 \cdot 16}{16 + 16} \frac{16 + 16 - 2 - 1}{(16 + 16 - 2)2} (4.0, 0.2) \begin{pmatrix} 45.5 & -0.3 \\ -0.3 & 0.3 \end{pmatrix}^{-1} \begin{pmatrix} 4.0 \\ 0.2 \end{pmatrix} \simeq 0.2.$$

This quantity is to be compared with the quantiles in an $F(2, 29)$ -distribution and it is readily seen that a hypothesis that the mean vectors are equal can be accepted at all reasonable levels.

4.2 The multidimensional general linear model.

In the previous section we have looked at the one- and two-sample situation for the multidimensional normal distribution. We have seen that the multidimensional results are quite analogous to the one dimensional ones. In this section and in the following we will continue this analogy and derive the results regarding regression and analysis of variance of multidimensional variables.

We consider independently distributed variables $\mathbf{Y}_1, \dots, \mathbf{Y}_n$,

$$\mathbf{Y}_i \in \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}).$$

The variance-covariance matrix $\boldsymbol{\Sigma}$ (and the mean vectors $\boldsymbol{\mu}_i$) are assumed unknown. We arrange the observations in an $n \times p$ data matrix

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}'_1 \\ \vdots \\ \mathbf{Y}'_n \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}.$$

Here the *single rows represent e.g. repetitions of measurements of a p-dimensional phenomenon*. In full analogy with the model which we considered in the univariate general linear model we will assume that the mean parameter $\boldsymbol{\mu}_i$ can be written as known linear functions of other (and fewer) unknown parameters $\boldsymbol{\theta}$, i.e.

$$E(\mathbf{Y}) = \mathbf{x}\boldsymbol{\theta} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{bmatrix}.$$

It is seen that we assume \mathbf{x} known and $\boldsymbol{\theta}$ unknown. This model can be viewed from different angles. If we let the j 'th column in the \mathbf{Y} matrix equal

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}'_1 \\ \vdots \\ \mathbf{Y}'_n \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{|1} & \cdots & \mathbf{Y}_{|p} \end{bmatrix}$$

then we can write

$$E(\mathbf{Y}_{|j}) = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_{1j} \\ \vdots \\ \theta_{kj} \end{bmatrix} = \mathbf{x}\boldsymbol{\theta}_{|j}.$$

The n measurements on the j 'th "property" (attribute/variable) will therefore follow an ordinary one dimensional general linear model.

If we instead write the mean value of a single observation \mathbf{Y}_i , we find

$$\mathbb{E}(\mathbf{Y}'_i) = (x_{i1} \cdots x_{ik}) \begin{pmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{pmatrix} = \mathbf{x}'_i \boldsymbol{\theta},$$

where $\mathbf{x}'_i = \mathbf{x}_{-i}$ is the i 'th row in the \mathbf{x} -matrix. This readily gives

$$\mathbb{E}(\mathbf{Y}_i) = \boldsymbol{\theta}' \mathbf{x}_i.$$

If the observations are rearranged into a column vector

$$\underline{\mathbf{Y}} = \text{vc}(\mathbf{Y}) = \begin{bmatrix} \mathbf{Y}_{1|} \\ \vdots \\ \mathbf{Y}_{p|} \end{bmatrix},$$

we find from theorem 1.9, p. 10, that

$$\mathbf{D}(\mathbf{Y}) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n = \begin{bmatrix} \sigma_1^2 \mathbf{I}_n & \cdots & \sigma_{1p}^2 \mathbf{I}_n \\ \vdots & & \vdots \\ \sigma_{p1}^2 \mathbf{I}_n & \cdots & \sigma_p^2 \mathbf{I}_n \end{bmatrix},$$

where $\boldsymbol{\Sigma} \otimes \mathbf{I}_n$ is the tensor product of $\boldsymbol{\Sigma}$ and \mathbf{I}_n , cf. section A.5.

The first problem is to estimate $\boldsymbol{\theta}$. We have

||| Theorem 4.14

We consider the above mentioned situation. If the observations \mathbf{Y}_i are normally distributed the maximum likelihood estimate of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}' \mathbf{x})^{-1} \mathbf{x}' \mathbf{Y}.$$

||| Proof

Omitted. See e.g. [2].



|||| **Remark 4.15**

We see that

$$\hat{\theta}_{j\mid} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}_{j\mid},$$

i.e. the estimate for the j 'th column in θ is simply equal to the result we get by only considering the one dimensional general linear model for the j 'th "property".

|||| **Remark 4.16**

If the observations are not normally distributed one will still be able to use the estimate $\hat{\theta}$, since this of course just like the one dimensional case has a Gauss-Markov property. We will not go into details with this but just mention a couple of results. The least squares properties are that

$$M = (\mathbf{Y} - \mathbf{x}\theta)'(\mathbf{Y} - \mathbf{x}\theta) - (\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta})$$

is positive semidefinite. From this follows that

$$\text{ch}_i(\mathbf{Y} - \mathbf{x}\theta)'(\mathbf{Y} - \mathbf{x}\theta) \geq \text{ch}_i(\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta}),$$

where ch_i corresponds to the i 'th largest eigenvalue. From this follows again that $\hat{\theta}$ minimises

$$\det(\mathbf{Y} - \mathbf{x}\theta)'(\mathbf{Y} - \mathbf{x}\theta)$$

and

$$\text{tr}(\mathbf{Y} - \mathbf{x}\theta)'(\mathbf{Y} - \mathbf{x}\theta).$$

|||| **Remark 4.17**

Above we have silently assumed that $\mathbf{x}'\mathbf{x}$ has full rank i.e. $\text{rk}(\mathbf{x}) = k < n$. If this is not the case one can by analogy to the one dimensional (univariate) results find solutions by means of pseudo inverse matrices.

After these considerations on the estimation of $\hat{\theta}$ we turn to the estimation of Σ .

||| **Theorem 4.18**

We consider the situation from theorem 4.14. Then the maximum likelihood estimate for Σ equals

$$\begin{aligned}\hat{\Sigma}^* &= \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \hat{\boldsymbol{\theta}}' \mathbf{x}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\theta}}' \mathbf{x}_i)' \\ &= \frac{1}{n} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) \\ &= \frac{1}{n} [\mathbf{Y}'\mathbf{Y} - (\mathbf{x}\hat{\boldsymbol{\theta}})'(\mathbf{x}\hat{\boldsymbol{\theta}})].\end{aligned}$$

The (i, j) 'th element can also be written

$$\hat{\sigma}_{ij}^* = \frac{1}{n} (\mathbf{Y}_{i|} - \mathbf{x}\hat{\boldsymbol{\theta}}_{i|})'(\mathbf{Y}_{j|} - \mathbf{x}\hat{\boldsymbol{\theta}}_{j|}).$$

||| **Proof**

The many identities between $\hat{\Sigma}$'s elements are found by simple matrix manipulations. For the results we refer to [2].

■

The distribution of the estimators mentioned are given in

||| Theorem 4.19

We consider the situation from theorems 4.14 and 4.18 and we introduce the usual notations

$$\begin{aligned}\tilde{\theta} &= \text{vc}(\theta) = \begin{bmatrix} \theta_{|1} \\ \vdots \\ \theta_{|p} \end{bmatrix} \\ \hat{\theta} &= \text{vc}(\hat{\theta}) = \begin{bmatrix} \hat{\theta}_{|1} \\ \vdots \\ \hat{\theta}_{|p} \end{bmatrix}.\end{aligned}$$

Then we have that $\hat{\theta}$ is normally distributed

$$\hat{\theta} = \text{vc}(\hat{\theta}) \in N_{pk}(\tilde{\theta}, \Sigma \otimes (\mathbf{x}'\mathbf{x})^{-1}),$$

and $n\hat{\Sigma}^*$ is Wishart distributed

$$n\hat{\Sigma}^* \in W(n - k, \Sigma).$$

Finally $\hat{\Sigma}^*$ and $\hat{\theta}$ and therefore also $\hat{\Sigma}^*$ and $\hat{\theta}$ are stochastically independent.

||| Proof

It is trivial that

$$E(\hat{\theta}) = E[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{Y}] = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{x}\theta = \theta$$

and from this it follows that $E(\hat{\theta}) = \tilde{\theta}$. Furthermore $\hat{\theta}$ is of course normally distributed.

Finally we have that

$$D(\hat{\theta}_{|i}) = \sigma_{ii}(\mathbf{x}'\mathbf{x})^{-1}$$

and

$$C(\hat{\theta}_{|i}, \hat{\theta}_{|j}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'C(\mathbf{Y}_{|i}, \mathbf{Y}_{|j})\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1} = \sigma_{ij}(\mathbf{x}'\mathbf{x})^{-1}.$$

From this the result concerning the variance covariance matrix for $\hat{\theta}$ is readily seen.

The result concerning the distribution of $\hat{\Sigma}^*$ and concerning the independence of $\hat{\theta}$ and $\hat{\Sigma}^*$ are quite analogous to the corresponding one dimensional results but we will not look further into these matters here. The reader is referred to e.g. [2].

■

From the theorem we readily find

||| Corollary 4.20

The unbiased estimate for Σ is equal to

$$\hat{\Sigma} = \frac{n}{n-k} \hat{\Sigma}^* = \frac{1}{n-k} (\mathbf{Y} - \mathbf{x} \hat{\theta})' (\mathbf{Y} - \mathbf{x} \hat{\theta}).$$

||| Proof

Trivial when you remember that

$$E(W(k, \Delta)) = k\Delta.$$

■

We now turn to testing the parameters in the model.

We have

|||| **Theorem 4.21**

We consider the above mentioned situation including the assumption of the normality of the observations. Furthermore we consider the hypothesis

$$H_0 : \mathbf{A} \boldsymbol{\theta} \mathbf{B}' = \mathbf{C} \quad \text{against} \quad H_1 : \mathbf{A} \boldsymbol{\theta} \mathbf{B}' \neq \mathbf{C},$$

where $\mathbf{A}(r \times k)$, $\mathbf{B}(s \times p)$ and $\mathbf{C}(r \times s)$ are given (known) matrices. We introduce

$$\begin{aligned}\Delta &= \mathbf{A} \hat{\boldsymbol{\theta}} \mathbf{B}' - \mathbf{C} \\ \mathbf{R} &= n \hat{\Sigma}^* = (\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}})'(\mathbf{Y} - \mathbf{x} \hat{\boldsymbol{\theta}}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\theta}}'(\mathbf{x}'\mathbf{x})\hat{\boldsymbol{\theta}}\end{aligned}$$

and

$$\begin{aligned}\mathbf{E} &= \mathbf{B} \mathbf{R} \mathbf{B}' \\ \mathbf{H} &= \Delta'[\mathbf{A}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{A}']^{-1}\Delta.\end{aligned}$$

The likelihood ratio test for testing H_0 against H_1 is equivalent to the test given by the critical region

$$\{\mathbf{y} \mid \frac{\det(\mathbf{e})}{\det(\mathbf{e} + \mathbf{h})} \leq U(s, r, n - k)_\alpha\},$$

where $U(s, r, n - k)_\alpha$ is the α quantile in the null-hypothesis distribution of the test statistic (see below).

|||| **Proof**

Omitted. The essential part of the proof is that it can be shown that \mathbf{S} and \mathbf{H} are independent Wishart distributed variables if H_0 is true. For more detail we refer to the literature. As it is seen indirectly from the formulation of the theorem the null-hypothesis distribution of

$$\Lambda = \mathbf{U} = \frac{\det(\mathbf{E})}{\det(\mathbf{E} + \mathbf{H})}$$

only depends on s, r and $n - k$. The quantity is termed in the literature as *Wilks's Λ* or *Anderson's U* . Since the distribution contains three parameters it is somewhat difficult to use in practise and we therefore give an approximation to an F-distribution in the following

■

|||| **Theorem 4.22**

Let U be $U(p,q,r)$ -distributed and let

$$\begin{aligned} t &= \begin{cases} \frac{1}{\sqrt{\frac{p^2q^2-4}{p^2+q^2-5}}} & p^2 + q^2 = 5 \\ \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} & p^2 + q^2 \neq 5 \end{cases} \\ v &= \frac{1}{2}(2r + q - p - 1). \end{aligned}$$

Then

$$F = \frac{1 - U^{\frac{1}{t}}}{U^{\frac{1}{t}}} \cdot \frac{vt + 1 - \frac{1}{2}pq}{pq}$$

is approximately distributed as

$$F(pq, vt + 1 - \frac{1}{2}pq).$$

If either p or q are equal to 1 or 2, then the approximation is exact.

|||| **Proof**

Omitted.

■

|||| **Remark 4.23**

We see that the test statistic in Theorem 4.21 compares the "size" of the matrices \mathbf{E} and $\mathbf{E} + \mathbf{H}$. We shall now present the test statistic as a function of the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ and give three other functions of those eigenvalues that are also commonly considered as test statistics for the hypothesis given in Theorem 4.21.

We let $\lambda_1 \geq \dots \geq \lambda_n$ be the ordered eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ and let \mathbf{z}_i be the corresponding eigenvectors, i.e.

$$\mathbf{E}^{-1}\mathbf{H}\mathbf{z}_i = \lambda_i \mathbf{z}_i$$

By straight forward calculations we obtain

$$\begin{aligned} (\mathbf{E} + \mathbf{H})^{-1}\mathbf{E}\mathbf{z}_i &= \frac{1}{1 + \lambda_i} \mathbf{z}_i \\ (\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}\mathbf{z}_i &= \frac{\lambda_i}{1 + \lambda_i} \mathbf{z}_i \end{aligned}$$

Thus getting the eigenvalues of the matrices we see that *Wilks' Lambda* is equal to

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{E} + \mathbf{H})} = \prod_{i=1}^n \frac{1}{1 + \lambda_i}$$

and we introduce the *Pillai's Trace*

$$v = \text{tr}((\mathbf{E} + \mathbf{H})^{-1}\mathbf{H}) = \sum_{i=1}^n \frac{\lambda_i}{1 + \lambda_i},$$

Hotelling-Lawley's Trace

$$H = \text{tr}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^n \lambda_i$$

and finally *Roy's Maximum Root*

$$R = \lambda_1.$$

Earlier we presented an F-approximation to Wilks' Lambda. There exist similar expressions for the other, and all statistics are computed in the multivariate procedures of SAS. SAS may also produce exact or near-exact p-values in the multivariate tests.

We shall now illustrate the introduced concept in the following example.

||| Example 4.24

In the period 1968-69 the Royal Veterinary and Agricultural University's Experimental Farm for crop growing, Højbakkegård, conducted an experiment concerning the growth of lucerne. They investigated the offsprings from 176 crossings. In order to establish the "quality" of the single crossings 9 properties were measured on each one. The 9 variables are given in the following table.

As mentioned, the 5 first variables are graded on a numerical scale. This method is chosen since it is very difficult to measure the respective variables directly, and experience shows that it gives satisfactory results.

Variable No. & name	Unit of measure	Explanation
1: Type of growth	Grade 1 – 9	1 = growth is lying down, 9 = growth is upright
2: Regrowth after winter	"	1 = worst, 9 = best
3: Ability to creep	"	1 = no runners, 9 = most runners
4: Activity	"	1 = weakest, 9 = strongest
5: Time of blooming	"	1 = latest blooming, 9 = earliest blooming
6: Plant height	cm	
7: Seed weight	g per plant	
8: Plant weight	g per plant after drying	
9: Percent seed	%	Calculated per plant by means of (7) and (8)

The following analyses are based on the average values for the 9 variables based on numbers from between 15 and 20 plants (most of the results are based on 20 plants). In the following table a section of these numbers is shown.

Obs.No. = No. of cross- ing	Variable No. and name								
	1 Type of growth	2 Re- growth	3 Ability to creep	4 Activity	5 Bloom- ing	6 Plant- height	7 Seed weight	8 Plant weight	9 Per- cent seed
1	4.11	5.00	3.05	6.17	3.67	50.00	3.47	120.10	2.75
2	3.08	4.75	4.17	7.50	5.17	61.50	0.82	111.33	0.75
3	3.12	4.00	3.35	6.53	3.99	55.29	0.86	97.47	0.81
:									
176	4.00	4.40	4.60	7.40	2.90	50.00	0.66	153.50	0.44

The main goal with the experiment was to examine the variation among the 9 variables. More specifically one was e.g. interested in how variable 3 (ability to creep) and variable 4 (activity) varies together with the others. The two variables mentioned are usually of great importance for the development of a plant and it is therefore of importance what the relation is to the other variables.

As a first orientation we will compute the empirical correlation matrix. It is found to be

	1	2	3	4	5	6	7	8	9
1	1.000	-0.033	0.116	0.018	0.131	-0.207	0.035	-0.087	0.041
2	-0.033	1.000	0.711	0.515	0.125	0.199	-0.025	0.348	-0.066
3	0.116	0.711	1.000	0.440	0.022	0.039	-0.133	0.218	-0.157
4	0.018	0.515	0.440	1.000	0.201	0.517	0.071	0.689	-0.081
5	0.131	0.125	0.022	0.201	1.000	0.496	0.987	0.168	0.486
6	-0.207	0.199	0.039	0.517	0.496	1.000	0.453	0.559	0.367
7	0.035	-0.025	-0.133	0.071	0.487	0.453	1.000	0.360	0.947
8	-0.087	0.348	0.218	0.689	0.168	0.559	0.360	1.000	0.128
9	0.041	-0.066	-0.157	-0.081	0.486	0.367	0.947	0.128	1.000

We note that variable 1 (type of growth) is only vaguely correlated with the other variables. On the other hand e.g. variables 2 and 3 (re-growth and ability to creep) and (of course) 7 and 9 (weight of seed and percentage of seed) are very strongly correlated.

As mentioned we are especially interested in variable 3's and variable 4's variation with the other variables. We note that there are a number of fairly large correlations but it is difficult to get an impression solely based on these. We will therefore try if it is possible to express these two variables as linear functions of the others i.e.

$$\begin{aligned} E(Y_1) &= \sum_{i=1}^k \theta_{i1} x_i \\ E(Y_2) &= \sum_{i=1}^k \theta_{i2} x_i \end{aligned}$$

where we now have used the variable notations

- Y_1 = Ability to "creep"
 Y_2 = Activity
 x_1 = Type of growth
 x_2 = Re growth after winter
 x_3 = Time of blooming
 x_4 = Height of plant
 x_5 = Weight of seed
 x_6 = Weight of plant
 x_7 = Percentage of seed

We are obviously talking about a multidimensional general linear model. If we let $\theta = (\theta_{ij})$, we get

$$\hat{\theta} = \begin{bmatrix} 0.28400 & 0.42731 \\ 0.79508 & 0.22230 \\ -0.02573 & 0.02607 \\ -0.01151 & 0.06290 \\ -0.14467 & -0.16756 \\ 0.00307 & 0.01103 \\ 0.10614 & 0.03463 \end{bmatrix}.$$

If we assume

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \in N(\mu_i, \Sigma),$$

then the unbiased estimate of Σ is

$$\hat{\Sigma} = \begin{bmatrix} 0.85897 & 0.07870 \\ 0.07870 & 0.29444 \end{bmatrix}.$$

The matrix $(\mathbf{x}'\mathbf{x})^{-1}$ is found to be

1	2	3	4	5	6	7
1.55920	-0.16549	-0.47258	-0.05010	0.41826	-0.00235	-0.42289
-0.16549	0.85139	-0.17981	-0.01327	0.63774	-0.01759	-0.69467
-0.47258	-0.17981	1.77862	-0.10728	-0.29340	0.01164	-0.02184
-0.05010	-0.01327	-0.10728	0.02253	0.12325	-0.00441	-0.17012
0.41826	0.63774	-0.29340	0.12325	5.25546	-0.08437	-7.04885
-0.00235	-0.01759	0.01164	-0.00441	-0.08437	0.00243	0.11182
-0.42289	-0.69467	-0.02184	-0.17012	-7.04885	0.11182	10.11541

From this we can easily compute the variance and covariance on the single θ -values. Because we have

$$D(\hat{\theta}) = \Sigma \otimes (\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} \sigma_{11}(\mathbf{x}'\mathbf{x})^{-1} & \sigma_{12}(\mathbf{x}'\mathbf{x})^{-1} \\ \sigma_{21}(\mathbf{x}'\mathbf{x})^{-1} & \sigma_{22}(\mathbf{x}'\mathbf{x})^{-1} \end{pmatrix},$$

and therefore e.g.

$$\hat{V}(\hat{\theta}_{42}) = 0.2944 \cdot 0.02253 = 0.0066.$$

These results can be used in the construction of ordinary t-tests for the single coefficients. We will, however, not consider this here. Instead we will give a couple of examples of how to construct simultaneous tests. Let us e.g. consider the hypothesis

$$H_0 : \theta_{41} = \theta_{42} = 0$$

against all alternatives. This hypotheses must be brought into the form given in theorem 4.21. This can be done by choosing

$$\begin{aligned}\mathbf{A} &= (0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0) \\ \mathbf{B} &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\end{aligned}$$

and

$$\mathbf{C} = (0 \ 0).$$

Then we will have

$$\mathbf{A} \boldsymbol{\theta} \mathbf{B}' = (\theta_{41} \ \theta_{42}).$$

By the use of a standard programme we get the F-test statistic

$$F = 53.66$$

with degrees of freedom

$$(f_1, f_2) = (2, 168).$$

The test statistic is in this case exact F-distributed, since $s = 2$ and $r = 1$. It is seen that the observed F-value is significant at all reasonable levels.

As another example consider the hypothesis

$$\boldsymbol{\theta}_1 = \begin{bmatrix} \theta_{51} & \theta_{52} \\ \theta_{61} & \theta_{62} \\ \theta_{71} & \theta_{72} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

against all alternatives. This hypothesis can be transformed into the form of theorem 4.21 by choosing

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \\ \mathbf{B} &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\end{aligned}$$

and

$$\mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix};$$

since then we obtain

$$\mathbf{A} \boldsymbol{\theta} \mathbf{B}' = \boldsymbol{\theta}_1.$$

Asain using a standard programme we find

$$F = 10.63; \quad (f_1, f_2) = (6, 336).$$

Once again we have a clear significance.

As a last example consider the hypothesis

$$\theta_{62} = \theta_{72} = 0$$

against all alternatives. This is brought into the standard form by choosing

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \\ \mathbf{B} &= (0 \ 1) \end{aligned}$$

and

$$\mathbf{C} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The F-test statistic has (2, 169) degrees of freedom and is found to be 27.4. The values shown are therefore significant.

4.3 Multivariate Analyses of Variance (MANOVA)

We will now specialise the results from the previous section to generalisations of the univariate one- and two-sided analysis of variance. First

4.3.1 One-sided multi-dimensional analysis of variance

We consider observations

$$\begin{gathered} Y_{11}, \dots, Y_{1n_1} \\ \vdots \qquad \vdots . \\ Y_{k1}, \dots, Y_{kn_k} \end{gathered}$$

These are assumed to be stochastically independent with

$$Y_{ij} \in N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \quad i = 1, \dots, k; \quad j = 1, \dots, n_i,$$

i.e. p -dimensional normal distributed with the same variance-covariance matrix. We wish to test hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_k$$

against

$$H_1 : \exists i, j (\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j).$$

Analogously to the univariate one-sided analysis of variance we define sums of squares deviation matrices

$$\begin{aligned} \mathbf{T} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{Y}})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}})' \\ \mathbf{W} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)' \\ \mathbf{B} &= \sum_{i=1}^k n_i (\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})' \end{aligned}$$

Here we have with $n = \sum_i n_i$

$$\begin{aligned} \bar{\mathbf{Y}}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{Y}_{ij} \\ \bar{\mathbf{Y}} &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} \mathbf{Y}_{ij}. \end{aligned}$$

After a bit of algebra we see that “total” matrix \mathbf{T} is the sum of the “between groups” matrix \mathbf{B} and the “within groups” matrix \mathbf{W} i.e.

$$\mathbf{T} = \mathbf{W} + \mathbf{B},$$

i.e. as in the one-dimensional case we have a partitioning of the total variation in the variation between groups and the variation within groups.

It is trivial that we as an unbiased estimate of the variance-covariance matrix Σ can use

$$\hat{\Sigma} = \frac{1}{n-k} \mathbf{W}.$$

If the hypothesis is true then \mathbf{T} will also be proportional with such an estimate. If the hypothesis is not true then \mathbf{T} will be “larger”. Therefore the following theorem seems intuitively reasonable.

||| **Theorem 4.25**

The ratio test for the test of the hypothesis H_0 against H_1 is given by the critical region

$$\{\mathbf{y}_{11}, \dots, \mathbf{y}_{kn_k} \mid \frac{\det(\mathbf{w})}{\det(\mathbf{t})} \leq U(p, k-1, n-k)_\alpha\}.$$

||| **Proof**

Omitted. Is found by special choices of \mathbf{A} , \mathbf{B} and \mathbf{C} matrices in theorem 4.21. ■

Just as the case for the one-dimensional analysis of variance the results are displayed using an analysis of variance table.

Source of variation	SS – matrix	Degrees of freedom
Deviation from hypothesis = variation between groups	$\mathbf{B} = \sum_i n_i (\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})'$	$k - 1$
Error = variation within groups	$\mathbf{W} = \sum_i \sum_j (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)'$	$n - k$
Total	$\mathbf{T} = \sum_i \sum_j (\mathbf{Y}_{ij} - \bar{\mathbf{Y}})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}})'$	$n - 1$

As it is done in univariate ANOVA it is of course possible to determine expected values of the \mathbf{B} and \mathbf{T} matrices even without H_0 being true. We will, however, not pursue this further here.

4.3.2 Two-sided multidimensional analysis of variance

In this case we will only look at a two-sided analysis of variance with 1 observation per cell. We will therefore assume that we have observations

$$\begin{aligned} Y_{11}, \dots, Y_{1m} \\ \vdots \quad \vdots , \\ Y_{k1}, \dots, Y_{km} \end{aligned}$$

which are assumed to be p -dimensional normal distributed with the same variance-covariance matrix Σ and with mean values

$$E(Y_{ij}) = \mu_{ij} = \mu + \alpha_i + \beta_j,$$

where the parameters α_i β_j satisfy

$$\sum_i \alpha_i = \sum_j \beta_j = \mathbf{0}.$$

We now want to test the hypothesis

$$H_0 : \alpha_1 = \dots = \alpha_k = \mathbf{0}$$

against

$$H_1 : \exists i (\alpha_i \neq \mathbf{0})$$

and

$$K_0 : \beta_1 = \dots = \beta_m = \mathbf{0}$$

against

$$K_1 : \exists j (\beta_j \neq \mathbf{0}).$$

Analogous to the sums of squares of the one-dimensional (univariate) analysis of variance we define the matrices

$$\begin{aligned} \mathbf{T} &= \sum_{i=1}^k \sum_{j=1}^m (Y_{ij} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{..})' \\ \mathbf{Q}_1 &= \sum_{i=1}^k \sum_{j=1}^m (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})' \\ \mathbf{Q}_2 &= m \sum_{i=1}^k (\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{i.} - \bar{Y}_{..})' \\ \mathbf{Q}_3 &= k \sum_{j=1}^m (\bar{Y}_{.j} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})'. \end{aligned}$$

Here we have used the usual notation

$$\begin{aligned}\bar{Y}_{..} &= \frac{1}{km} \sum_{i=1}^k \sum_{j=1}^m Y_{ij} \\ \bar{Y}_{i\cdot} &= \frac{1}{m} \sum_{j=1}^m Y_{ij}, \quad i = 1, \dots, k \\ \bar{Y}_{\cdot j} &= \frac{1}{k} \sum_{i=1}^k Y_{ij}, \quad j = 1, \dots, m.\end{aligned}$$

We see in this case that we also have the usual partitioning of the total variation

$$\mathbf{T} = \mathbf{Q}_1 + \mathbf{Q}_2 + \mathbf{Q}_3,$$

i.e. the total variation (\mathbf{T}) is partitioned in the variation between rows (\mathbf{Q}_2), and the variation between columns (\mathbf{Q}_3) and the residual variation (interaction variation) (\mathbf{Q}_1).

We now have

||| Theorem 4.26

The ratio test at level α for test of H_0 against H_1 is given by the critical region

$$\{y_{11}, \dots, y_{km} \mid \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_2)} \leq U(p, k-1, (k-1)(m-1))_\alpha\}.$$

The ratio test at level α for test of K_0 against K_1 is given by the critical region

$$\{y_{11}, \dots, y_{km} \mid \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_3)} \leq U(p, m-1, (k-1)(m-1))_\alpha\}.$$

||| Proof

Omitted. Follows readily from theorem 4.21. See e.g. [2].

We collect the results in a usual analysis of variance table

Source of variation	SS-matrix	Degrees of freedom	Test statistic
Differences between columns	$\mathbf{Q}_3 = k \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})'$	$m - 1$	$\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1 + \mathbf{Q}_3)}$
Differences between rows	$\mathbf{Q}_2 = m \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{i.} - \bar{Y}_{..})'$	$k - 1$	$\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1 + \mathbf{Q}_2)}$
Residual	$\mathbf{Q}_1 = \sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}) \times (\bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})'$	$(k - 1)(m - 1)$	
Total	$\mathbf{T} = \sum_i \sum_j (\bar{Y}_{ij} - \bar{Y}_{..})(\bar{Y}_{ij} - \bar{Y}_{..})'$	$km - 1$	

The matrix $\frac{1}{(k-1)(m-1)} \mathbf{Q}_1$ can be used as a unbiased estimate of Σ .

We now give an illustrative example.

||| Example 4.27

At the Royal Veterinary and Agricultural University's experimental farm, Højbakkegård, an experiment concerning the yield of crops was conducted in the period 1956-58 as part of an international study. Experiments on 10 plant types were performed. The kinds of yield which were of interest were the amounts of

dry matter
green matter
nitrogen.

Each type of plant was grown in 6 blocks (i.e. plots of soil with different quality). In order to reduce the amount of data we will limit ourselves to three plants and to the year 1957. The results of the experiment considered are given below.

Type of plant	Type of yield	Block No.						Yield
		1	2	3	4	5	6	
Marchigiana	Dry matter	9.170	10.683	10.063	8.104	10.018	9.570	
	nitrogen	0.286	0.335	0.315	0.259	0.319	0.304	
	green matter	40.959	47.677	44.950	36.919	45.859	43.838	
Kayseri	Dry matter	9.403	10.914	11.018	11.385	13.387	12.848	
	nitrogen	0.285	0.330	0.333	0.339	0.400	0.383	
	green matter	42.475	49.546	50.152	51.718	60.758	58.334	
Atlantic	Dry matter	11.349	10.971	9.794	8.944	11.715	11.903	
	nitrogen	0.369	0.357	0.319	0.291	0.379	0.386	
	green matter	52.475	50.757	45.151	42.221	55.505	56.364	

in 1000 kg/ha

We wish to analyse how the yield varies with the blocks, the type of plants and the type of yield.

We will first analyse each type of yield by itself. For this we base the analysis on a two-sided analysis of variance. The model is

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1, 2, 3, \dots, 6, j = 1, \dots, 6),$$

and we are therefore assuming that each observation y_{ij} can be written as a sum of μ (level), α_i (effect of plant), β_j (effect of block) and ε_{ij} (residual, being a small randomly varying quantity).

If we first consider dry matter we get

$$y_{11} = 9.170, \quad y_{12} = 10.683, \dots, \quad y_{36} = 11.903.$$

The analysis of variance table was (found by means of SSP-ANOVA)

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F-values
A	11.218244	5	2.243648	2.25
B	10.945597	2	5.472798	5.49
AB	9.970109	10	0.997010	
Total	32.133936	17		

The test statistic for the hypothesis $\beta_1 = \dots = \beta_6 = 0$ is

$$F = \frac{s_3^2}{s_1^2} = 2.25 < 3.33 = F_{95\%}(5, 10)$$

i.e. we cannot reject that the β 's equal 0.

Correspondingly the test statistic for the hypothesis $\alpha_1 = \alpha_2 = \alpha_3 = 0$ equals

$$F = \frac{s_2^2}{s_1^2} = 5.49 > 4.10 = F_{95\%}(2, 10).$$

At a 5% level we therefore reject that the α 's all equal 0. However, we note that

$$F_{97.5\%}(2, 10) = 5.46,$$

so there is no significance at the 2.5% level.

If we perform the corresponding computations on the nitrogen yield we get, using as observations: $y'_{ij} = y_{ij} \cdot 1000$:

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F-values
A	10802.27734	5	2160.45532	2.60
B	8030.77734	2	4015.38867	4.83
AB	8310.55469	10	831.05542	
Total	27143.60938	17		

Here we again find that there is no difference between blocks but there is possibly a difference between plants. This difference is, however, not significant at the 2.5% level.

The corresponding computations on yield of green matter was (again using coded observations: $y'_{ij} = 1000y_{ij}$):

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F-values.
A	261702416	5	52340480	2.75
B	260173824	2	130086912	6.83
AB	190600448	10	19060032	
Total	712476672	17		

Here we again have that there is no difference between blocks. We also find a difference between plants at the 5% level but not at the 1% level since

$$F_{99\%}(2, 10) = 7.56.$$

We therefore see that the three types of yield show more or less the same sort of variation: There is no difference between blocks but there is difference between plants. These are, however, not significant at a small levels of α .

Now the three forms of yield are known to be strongly interdependent. Therefore we will expect that the analysis of variance would give more or less similar results and it would therefore be interesting to examine the variation and the yield when we take this dependency into consideration. Such a type of analysis can be performed by a three dimensional two-sided analysis of variance i.e. we use the model

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad i = 1, 2, 3, \quad j = 1, \dots, 6,$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \quad \alpha_i = \begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \alpha_{3i} \end{pmatrix}, \quad \beta_j = \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix},$$

and the observations are

$$Y_{ij} = \begin{pmatrix} \text{content of green matter} & \text{in plant } i \text{ in blok } j \\ \text{content of nitrogen} & \text{--- ---} \\ \text{content of dry matter} & \text{--- ---} \end{pmatrix}.$$

The observed values are

$$\mathbf{y}_{11} = \begin{pmatrix} 40.959 \\ 0.286 \\ 9.170 \end{pmatrix}, \dots, \mathbf{y}_{36} = \begin{pmatrix} 56.364 \\ 0.386 \\ 11.903 \end{pmatrix}.$$

In this way we can aggregate the three analysis of variances shown above into one.

With the notation from p. 210 the matrices \mathbf{Q}_1 , \mathbf{Q}_2 and \mathbf{Q}_3 are found to be

$$\begin{aligned}\mathbf{q}_2 &= \begin{bmatrix} 260.18359 \\ 1.38547 & 0.00803 \\ 52.37032 & 0.26262 & 10.94564 \end{bmatrix} \\ \mathbf{q}_3 &= \begin{bmatrix} 261.70239 \\ 1.67129 & 0.01080 \\ 53.97473 & 0.34801 & 11.21827 \end{bmatrix} \\ \mathbf{q}_1 &= \begin{bmatrix} 190.59937 \\ 1.25512 & 0.00831 \\ 43.45444 & 0.28667 & 9.97013 \end{bmatrix}\end{aligned}$$

The matrices have been found by means of the BMD-programme BMDX69. Still by means of the programme mentioned we find

Source	ln(Generalized variance)	U-statistic	Degrees of freedom	Approximate F-statistic	Degrees of freedom
I	-1.89908	0.003315	3 2 10	43.6455	6 16.00
J	-4.84194	0.062894	3 5 10	2.5843	15 22.49
Full model	-7.60824				

Here I corresponds to the variation between plants and J to the variation between blocks.

The (in this case exact) F-test statistic for a test of the hypothesis $\alpha_1 = \alpha_2 = \alpha_3 = 0$, (i.e.. the hypothesis that all plants are equal) is 43.6. The number of degrees of freedom is (6,16). Since

$$F(6, 16)_{0.9995} = 7.74,$$

we therefore have a very strong rejection of the hypothesis.

Since

$$F(15, 22)_{0.975} = 2.50,$$

we see that now also the hypothesis of the blocks being equal is rejected at the level $\alpha = 2.5\%$.

The conclusion on the multi-dimensional analysis of variance is therefore that there is a clear difference in the yield for the three types of plants. It is on the other hand more uncertain if there are differences between the blocks.

We note a difference from three one-dimensional analyses. In these cases we only have moderate or no significance for the hypothesis of the plant yields being equal. We therefore have different results by considering the simultaneous analysis instead of the three marginal ones.

4.4 Tests regarding variance-covariance matrices

In this section we will briefly give some of the tests for hypothesis on variance covariance matrices. On one hand corresponding to a hypothesis about the variance covariance matrix having a given structure or is equal to a given matrix, or on the other hand corresponding to a hypothesis that several variance covariance matrices are equal.

4.4.1 Tests regarding a single variance-covariance matrix

First we will give a test that k -groups of normally distributed variables are independent. We are considering a $\mathbf{X} \in N_p(\boldsymbol{\mu}, \Sigma)$, and we divide \mathbf{X} in k components we the dimensions p_1, \dots, p_k , i.e.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_k \end{bmatrix}.$$

The corresponding partitioning of the parameters is

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \vdots \\ \boldsymbol{\mu}_k \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \cdots & \Sigma_{1k} \\ \vdots & & \vdots \\ \Sigma_{k1} & \cdots & \Sigma_{kk} \end{bmatrix}.$$

Our hypothesis is now that X_1, \dots, X_k are independent i.e. that the variance-covariance matrix has the form

$$\Sigma = \Sigma_0 = \begin{bmatrix} \Sigma_{11} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \Sigma_{kk} \end{bmatrix}.$$

If we define $\hat{\Sigma}$ computed on the basis of n realisations of X in the usual way and if we partition $\hat{\Sigma}$ analogously to the partitioning of Σ , we have

|||| Theorem 4.28

We consider the above mentioned situation and let

$$V = \frac{\det(\hat{\Sigma})}{\prod_{i=1}^k \det(\hat{\Sigma}_{ii})}.$$

Then the coefficient test for test of the hypothesis $\Sigma = \Sigma_0$ is given by the critical region

$$\{V \leq v_\alpha\}.$$

When finding the boundary of the critical region we can use that

$$\begin{aligned} P\{-m \ln V \leq v\} \\ \simeq P\{\chi^2(f) \leq v\} + \frac{\gamma_2}{m^2}[P\{\chi^2(f+4) \leq v\} - P\{\chi^2(f) \leq v\}], \end{aligned}$$

where

$$\begin{aligned} m &= n - \frac{3}{2} - \frac{p^3 - \sum p_i^3}{3(p^2 - \sum p_i^2)} \\ \gamma_2 &= \frac{p^4 - \sum p_i^4}{48} - \frac{5(p^2 - \sum p_i^2)}{96} - \frac{(p^3 - \sum p_i^3)^2}{72(p^2 - \sum p_i^2)}. \end{aligned}$$

$$f = \frac{1}{2}[p^2 - \sum p_i^2], \quad p = \sum p_i$$

If $k = 2$, the V is distributed as $U(p_1, p_2, n - 1 - p_2)$.

|||| Proof

Omitted. See e.g. [2].

In the above mentioned situation we looked at a test for a variance covariance matrix having a certain structure. We will now turn around and look at a test for the hypothesis that a variance covariance matrix is proportional with a given matrix. We briefly give the result in

||| Theorem 4.29

We consider independent observations X_1, \dots, X_n with $X_i \in N_p(\mu, \Sigma)$, and we let

$$\mathbf{A} = \sum (X_i - \bar{X})(X_i - \bar{X})'.$$

The likelihood ratio test statistic for a test of $H_0 : \Sigma = \sigma^2 \Sigma_0$, where Σ_0 is known and σ^2 unknown against all alternatives is

$$W = \frac{[\det(\mathbf{A} \Sigma_0^{-1})]^{\frac{n}{2}}}{[\text{tr } \mathbf{A} \Sigma_0^{-1} / p]^{\frac{pn}{2}}}.$$

When determining the critical region we can use that

$$\begin{aligned} P\{-(n-1)\rho \ln W \leq z\} \\ \simeq P\{\chi^2(f) \leq z\} + \omega_2 [P\{\chi^2[f+4] \leq z\} - P\{\chi^2(f) \leq z\}], \end{aligned}$$

where

$$\begin{aligned} \rho &= 1 - \frac{2p^2 + p + 2}{6p(n-1)} \\ f &= \frac{1}{2}p(p+1) - 1 \\ \omega_2 &= \frac{(p+2)(p-1)(p-2)(2p^3 + 6p^2 + 3p + 2)}{288p^2n^2\rho^2}. \end{aligned}$$

||| Proof

Omitted. See e.g. [2].

Finally we will consider the situation where we wish to test that a variance

covariance matrix is equal to a given matrix. Then the following holds true

|||| Theorem 4.30

We consider independent observations X_1, \dots, X_n with $X_i \in N_p(\mu, \Sigma)$, and we let

$$\mathbf{A} = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'.$$

The quotient test statistic for a test of $H_0 : \Sigma = \Sigma_0$, where Σ_0 is known against all alternatives is

$$\lambda_1 = \left(\frac{e}{n}\right)^{pn/2} [\det(\mathbf{A} \Sigma_0^{-1})]^{\frac{n}{2}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{A} \Sigma_0^{-1})\right).$$

When determining the critical region we can use that

$$P\{-2 \ln \lambda_1 \leq v\} \simeq P\{\chi^2(\frac{1}{2}p(p+1)) \leq v\}.$$

|||| Proof

Omitted. See e.g. [2].

■

4.4.2 Test for equality of several variance-covariance matrices

We will in this section consider the problem of testing the assumption of equal variance covariance matrices in Hotelling's two sample situation and in the multidimensional analysis of variance.

We will assume that there are independent observations

$$\begin{aligned} X_{11}, \dots, X_{1n_1}, \quad & X_{1j} \in N_p(\mu_1, \Sigma_1) \\ \vdots & \\ X_{k1}, \dots, X_{kn_k}, \quad & X_{kj} \in N_p(\mu_k, \Sigma_k) \end{aligned}$$

and we wish to test the hypothesis

$$H_0 : \Sigma_1 = \dots = \Sigma_k \quad \text{against} \quad H_1 : \exists i, j : \Sigma_i \neq \Sigma_j.$$

We let

$$\begin{aligned} n &= \sum n_i, \\ \mathbf{W}_i &= \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)', \end{aligned}$$

and

$$\mathbf{W} = \sum_{i=1}^k \mathbf{W}_i,$$

cf. section 4.3.1.

We then have

|||| Theorem 4.31

As a test statistic for the test of H_0 against H_1 we can use

$$L = \frac{\prod_{i=1}^k [\det(\mathbf{W}_i)]^{\frac{(n_i-1)}{2}}}{[\det \mathbf{W}]^{\frac{(n-k)}{2}}} \cdot \frac{(n-k)^{\frac{p(n-k)}{2}}}{\prod_{i=1}^k (n_i - 1)^{\frac{p(n_i-1)}{2}}}.$$

The critical region is of the form

$$\{L \leq l_\alpha\}$$

and in the determination of this we can use that

$$\begin{aligned} P\{-2\rho \ln L \leq z\} &\approx \\ P\{\chi^2(f) \leq z\} + \omega_2 [P\{\chi^2(f+4) \leq z\} - P\{\chi^2(f) \leq z\}], \end{aligned}$$

where

$$\begin{aligned} f &= \frac{1}{2}(k-1)p(p+1), \\ \rho &= 1 - \left(\sum_i \frac{1}{n_i} - \frac{1}{n} \right) \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}, \\ \omega_2 &= \frac{1}{48\rho^2} p(p+1) [(p-1)(p+2) \left(\sum_i \frac{1}{n_i^2} - \frac{1}{n^2} \right) - 6(k-1)(1-\rho)^2]. \end{aligned}$$

|||| **Proof**

Omitted. See e.g. [2].

■

|||| Chapter 5

Discriminant analysis and classification

In this section we will address the problem of characterizing different populations and classifying an individual in one of those known populations based on measurements of some characteristics of the individual.

We may think of the question of classifying e.g. cell tissue obtained from stained samples of biopsies. In figure 5.1 we have an example of such a sample, where the dark parts are marking tumor tissue. It will of course be very relevant to be able to make an (semi)automated algorithm that can identify tumor tissue in such samples and then compute various descriptors based on such a classification. The other half of the figure shows a part of a salami sausage obtained from a study on monitoring the fermentation process of salamis. A first task will be to classify the individual pixels as either fat or meat. Having done so, we may proceed with some more elaborate analyses.

In Figure 5.2 we have shown scatterplots and histograms of three variables obtained from images of cells taken from samples from 4 individuals. The variables considered are the log of the area of the cell, the log of the ratio between the area of the nucleus and the area of the cell, and finally the average image value of the cytoplasm in the cell. It is of interest whether those values are characteristic for the individuals. The points have been color coded in cyan, red, green, and blue corresponding to the four individuals.

It is seen that the points are mixed together pretty much, so there does not seem to be a direct connection between the values and the individuals. The $\log(\text{area})$ variable seems to follow a univariate normal distribution fairly well, whereas the two other variables show a bimodal behavior, but not related to the individuals! It would be of interest to identify the factors that control these

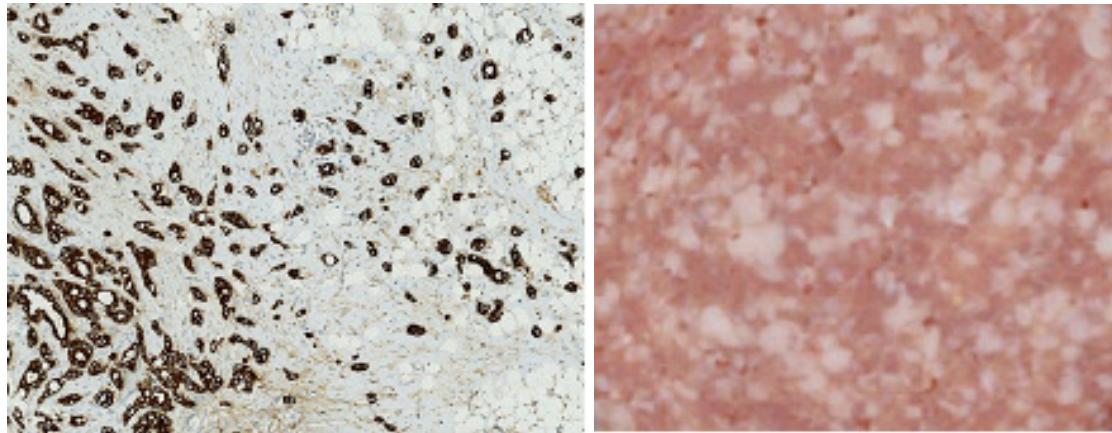


Figure 5.1: Left. Tissue Micro Array core slice from breast carcinomas stained with PCK marking tumor tissue. The core were digitalized using a high resolution optical scanner. Source Lindberg et al (2017). Right. Salami section two days after fermentation start showing fat and meat fractions. Source Trinderup et al (2017).

bimodalities: Do we have measurements taken by two different operators, are there a time dependence etc.

In Figure 5.3 we have shown a Landsat false color composite satellite image. Furthermore we have mapped values of Landsat bands 1 and 6, and bands 1, 6, and 3 in scatterplots color coded with the same color as used for delineating some training areas. In contrast to the previous image we see a fairly good grouping of the observations corresponding to the different training areas. Therefore it will be feasible to use values of the Landsat bands to classify a pixel as coming from one of the geological units considered.

One should note however, that the green points fall into two, rather distinct clusters. The pixels corresponding to the green points come from the geological unit called 'Deltas and young alluvial fans'. This unit is composed of two separate areas, and it turns out that one is sunlit, the other is in shadow. One may therefore consider defining two populations corresponding to the different light conditions.

We shall see that it actually makes a difference whether we do one or the other, indicating that in a classification task, the homogeneity of the training areas is important.

We first consider the problem of discriminating between two populations. In the sequel we shall use the terms populations, groups or classes interchangeably.

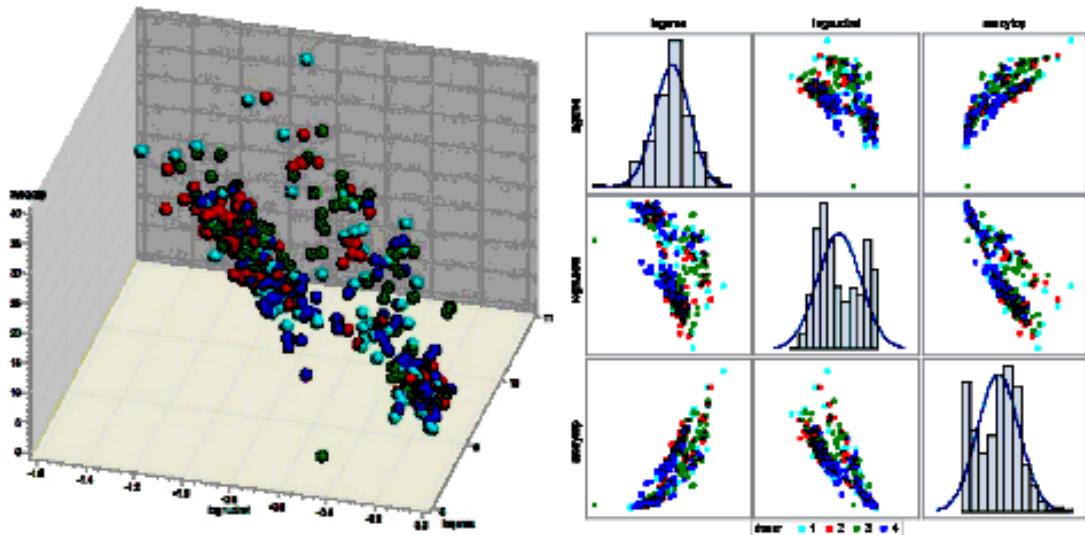


Figure 5.2: Scatter plots and histograms of cell measurements (log area, log (area nucleus/area), average cytoplasm intensity. The color code corresponds to different individuals. Source Ottosen (2015)

5.1 Discrimination between two populations

5.1.1 Bayes and minimax solutions

We consider the populations π_1 and π_2 and wish to conclude whether a given individual is a member of population one or population two. We perform measurements of p different characteristics of the individual and hereby get the result

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}.$$

If the individual comes from π_1 the frequency function of \mathbf{X} is $f_1(\mathbf{x})$ and if it comes from π_2 it is $f_2(\mathbf{x})$.

Let us furthermore assume that we have given a *loss function* L :

From	Classify as	
	π_1	π_2
Nature	π_1	0
	π_2	$L(\pi_2, \pi_1) = L_{21}$
		$L(\pi_1, \pi_2) = L_{12}$
		0

Table 5.1: The loss function connected with the classification problem

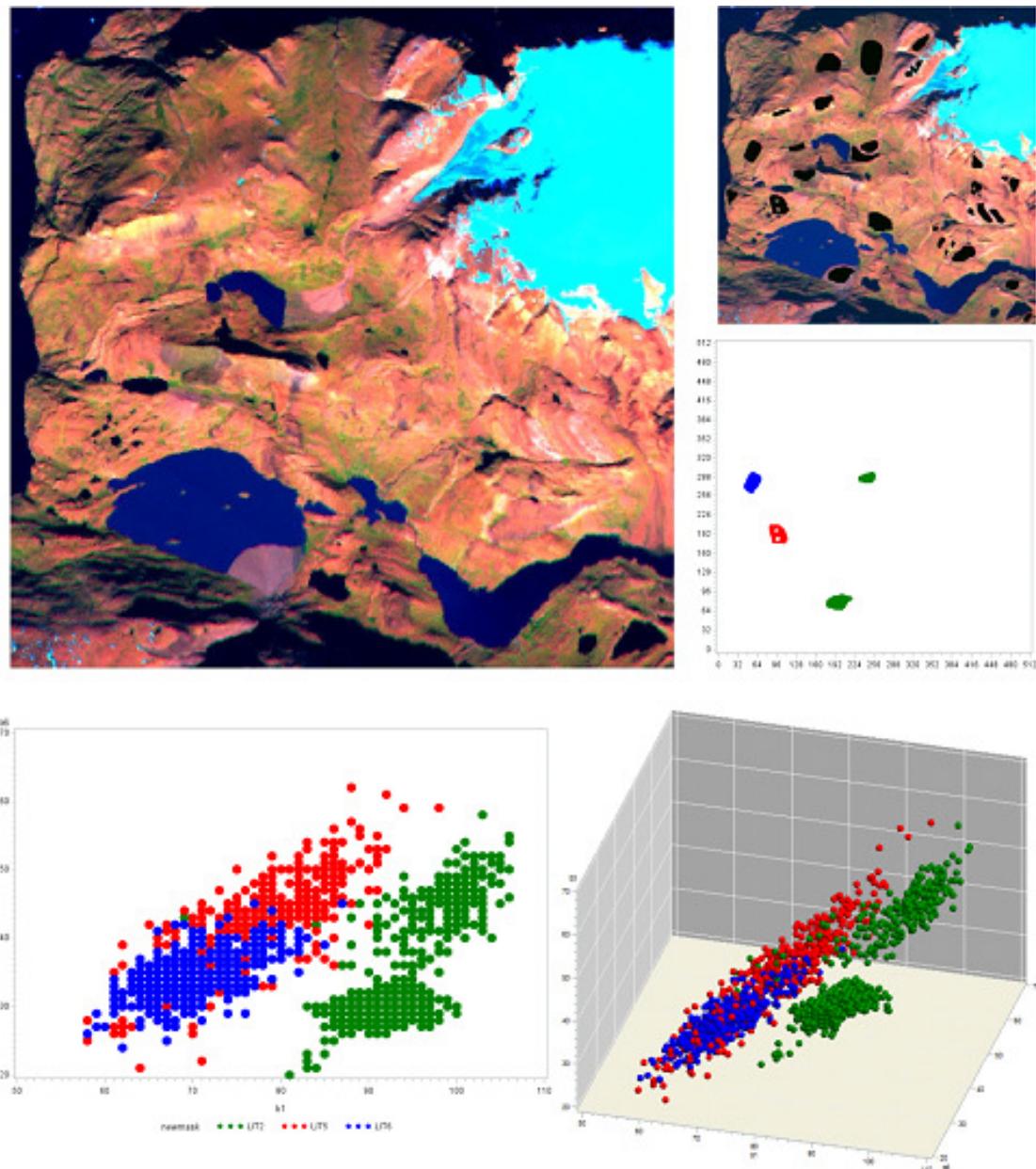


Figure 5.3: Top. Landsat false color composite (RGB \sim bands 4, 6, 1) of area on Ymer Ø, Central East Greenland. Training areas from different geological units; locality of training areas for unit 2 (deltas and young alluvial fans, green), unit 5 (quartzites, red), and unit 6 (black shales, blue).
Bottom. Scatterplots of values of Landsat bands 1 and 6, and bands 1, 6, and 3 color coded with the same color as used for delineating the training areas.
Source Conradsen et al (1987).

We will assume that there is no loss if we select the correct population.

In certain situations one also knows approximately what the prior probability is to have an individual from each of the groups, i.e. we have given a *prior distribution* g :

$$P\{\Pi = \pi_i\} = g(\pi_i) = p_i, \quad i = 1, 2,$$

where the random variable Π is designating the population we have observations from.

We now seek a *decision function* $d : \mathbb{R}^p \rightarrow \{\pi_1, \pi_2\}$ of the form

$$d(\mathbf{x}) = d_{R_1}(\mathbf{x}) = \begin{cases} \pi_1 & \text{if } \mathbf{x} \in R_1 \\ \pi_2 & \text{if } \mathbf{x} \in R_2 = R_1^c \end{cases}$$

where R_1^c is the complement set of R_1 . We thus divide \mathbb{R}^p into two regions R_1 and R_2 . If our observation lies in R_1 we will choose π_1 and if our observation lies in R_2 we will choose π_2 .

For each π_i $d_{R_1}(\mathbf{X})$ is therefore a binary random variable assuming the values π_1 or π_2 with probabilities $P\{\mathbf{X} \in R_1 | \pi_i\}$ and $P\{\mathbf{X} \in R_2 | \pi_i\}$.

If we have a prior distribution we define the *posterior distribution* k by

$$k(\pi_i | \mathbf{x}) = \frac{g(\pi_i) f_i(\mathbf{x})}{g(\pi_1) f_1(\mathbf{x}) + g(\pi_2) f_2(\mathbf{x})} = \frac{p_i f_i(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})}$$

which is the conditional distribution of the random variable Π given that the observation $\mathbf{X} = \mathbf{x}$. The result follows from Bayes' Theorem.

The expected loss in this distribution is

$$\begin{aligned} E_{\mathbf{x}}(L(\Pi, d_{R_1}(\mathbf{x}))) &= L(\pi_1, d_{R_1}(\mathbf{x})) k(\pi_1 | \mathbf{x}) + L(\pi_2, d_{R_1}(\mathbf{X})) k(\pi_2 | \mathbf{x}) \\ &= \begin{cases} L(\pi_2, \pi_1) k(\pi_2 | \mathbf{x}) & \text{if } \mathbf{x} \in R_1 \\ L(\pi_1, \pi_2) k(\pi_1 | \mathbf{x}) & \text{if } \mathbf{x} \in R_2 \end{cases}. \end{aligned}$$

The Bayes solution is defined by minimizing this quantity for any \mathbf{x} , i.e. we define R_1 by

$$\begin{aligned} \mathbf{x} \in R_1 &\iff L(\pi_2, \pi_1) k(\pi_2 | \mathbf{x}) \leq L(\pi_1, \pi_2) k(\pi_1 | \mathbf{x}) \\ &\iff \frac{L_{12} p_1 f_1(\mathbf{x})}{L_{21} p_2 f_2(\mathbf{x})} \geq 1 \\ &\iff \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{L_{21} p_2}{L_{12} p_1} \end{aligned}$$

These considerations are collected in

|||| **Theorem 5.1**

The *Bayes solution* to the classification problem is given by the region

$$R_1 = \left\{ \mathbf{x} \mid \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{L_{21}p_2}{L_{12}p_1} \right\}.$$

If we do not have a prior distribution we can instead determine a minimax strategy i.e. determine R_1 so that the maximal risk is minimised. For a given decision function d_{R_1} , the risk R is - for each population π_i - defined as the mean loss in the distribution of \mathbf{X} , i.e.

$$\begin{aligned} R(\pi_1, d_{R_1}) &= E_{\pi_1} L(\pi_1, d_{R_1}(\mathbf{X})) \\ &= L_{11} \times P\{\mathbf{X} \in R_1 \mid \pi_1\} + L_{12} \times P\{\mathbf{X} \in R_2 \mid \pi_1\} = L_{12} \times P\{\mathbf{X} \in R_2 \mid \pi_1\} \\ R(\pi_2, d_{R_1}) &= E_{\pi_2} L(\pi_2, d_{R_1}(\mathbf{X})) \\ &= L_{21} \times P\{\mathbf{X} \in R_1 \mid \pi_2\} + L_{22} \times P\{\mathbf{X} \in R_2 \mid \pi_2\} = L_{21} \times P\{\mathbf{X} \in R_1 \mid \pi_2\} \end{aligned}$$

One can now show

|||| **Theorem 5.2**

The *minimax solution* for the classification problem is given by the region

$$R_1 = \left\{ \mathbf{x} \mid \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c \right\}.$$

where c is determined by

$$L_{12}P\left\{ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < c \mid \pi_1 \right\} = L_{21}P\left\{ \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c \mid \pi_2 \right\}.$$

|||| **Remark 5.3**

The relation for determining c can be written

$$\begin{aligned} L_{12} \times (\text{the probability of misclassification if } \pi_1 \text{ is true}) \\ = L_{21} \times (\text{the probability of misclassification if } \pi_2 \text{ is true}) \end{aligned}$$

Since the left hand side is an increasing and the right hand side is a decreasing function of c it is obvious that we will minimize the maximal risk when we have equality. If we do not have any idea about the size of the losses we can let them both equal one. The minimax solution then gives us the region which minimizes the maximal probability of misclassification.

We will now consider the important special case where f_1 and f_2 are normal distributions.

5.1.2 Discrimination between two normal populations

If f_1 and f_2 are normal with the same dispersion matrix we have

|||| **Theorem 5.4**

Let $\pi_1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\pi_2 \sim N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. Then we have

$$\begin{aligned} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c &\Leftrightarrow \mathbf{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \geq \ln c \\ &\Leftrightarrow \left[\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 \right] - \left[\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 - \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \right] \geq \ln c. \end{aligned}$$

|||| **Proof**

We introduce the inner product ($|$) and the norm $\|\cdot\|$ by

$$(\mathbf{x} | \mathbf{y}) = (\mathbf{x} | \mathbf{y})_{\boldsymbol{\Sigma}^{-1}} = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{y}$$

and

$$\|\mathbf{x}\|^2 = \|\mathbf{x}\|_{\boldsymbol{\Sigma}^{-1}}^2 = (\mathbf{x} | \mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}.$$

We then have

$$f_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^p \sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}\|\mathbf{x} - \boldsymbol{\mu}_i\|^2\right).$$

From this we readily get

$$\begin{aligned} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c &\Leftrightarrow \ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \ln c \\ &\Leftrightarrow -\|\mathbf{x} - \boldsymbol{\mu}_1\|^2 + \|\mathbf{x} - \boldsymbol{\mu}_2\|^2 \geq 2\ln c \\ &\Leftrightarrow (\mathbf{x} - \boldsymbol{\mu}_1 | \mathbf{x} - \boldsymbol{\mu}_1) + (\mathbf{x} - \boldsymbol{\mu}_2 | \mathbf{x} - \boldsymbol{\mu}_2) \geq 2\ln c \\ &\Leftrightarrow 2(\mathbf{x} | \boldsymbol{\mu}_1) - 2(\mathbf{x} | \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 | \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_2 | \boldsymbol{\mu}_2) \geq 2\ln c \\ &\Leftrightarrow 2(\mathbf{x} | \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 | \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_2 | \boldsymbol{\mu}_2) \geq 2\ln c. \end{aligned}$$

By using the connection between $(|)$ and Σ^{-1} we find that the theorem readily follows.

■

|||| Remark 5.5

The expression $\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c$ is seen to define a subset of \mathbb{R}^p which is delimited by a hyper-plane (for $p = 2$ a straight line and for $p = 3$ a plane).

The vector $\overrightarrow{p_1 p_2}$ is the orthogonal projection with respect to Σ^{-1} of \mathbf{x} onto the line which connects $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. (It can be shown that the slope of the projection lines etc. are equal to the slope of the ellipse- (ellipsoid-) tangents at the points where they intersect the line $(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2)$. Since the length of a projection of a vector is equal to the inner product between the vector and a unit vector on the line we see that we classify an observation as coming from π_1 iff the projection of \mathbf{x} is large enough (computed with sign). Otherwise we will classify the observation as coming from π_2 .

The functions

$$\mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \ln p_i, \quad i = 1, 2$$

are called **linear discriminant functions** for π_i . If we do not have a prior distribution, the term $\ln p_i$ is omitted. The function

$$\mathbf{x}^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 - \ln c$$

is called the *linear discriminator* between π_1 and π_2 .

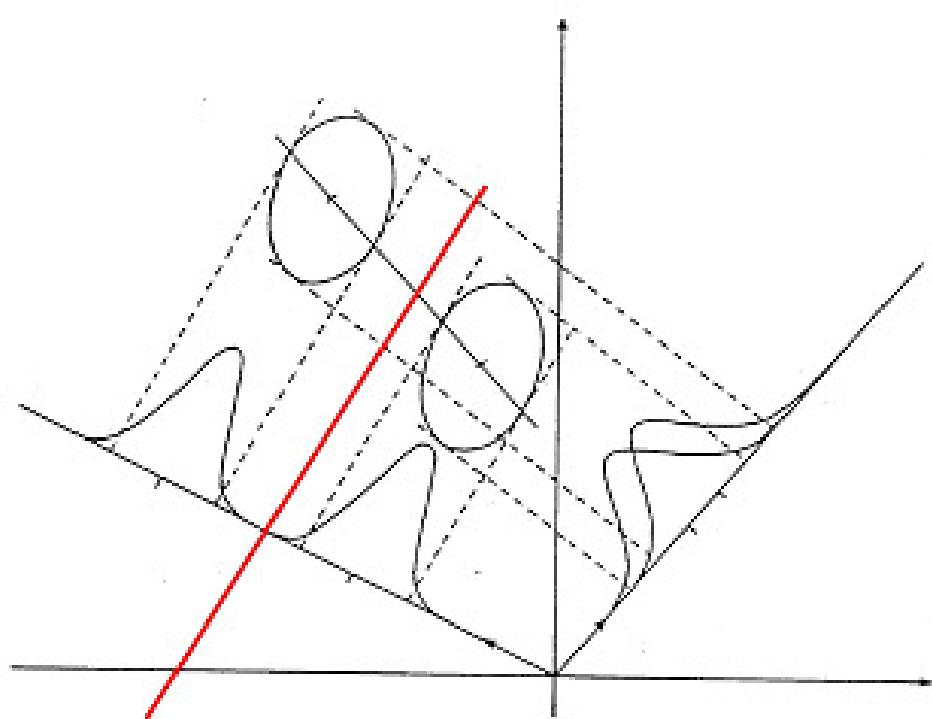


Figure 5.4: Classification example

We then have that the discriminator is the linear projection which - after the addition of suitable constants - minimizes the expected loss (the Bayes situation) or the probability of misclassification (the minimax situation).

In order to elucidate the content of the theorem, we will now give a slightly different interpretation of a discriminator. If we define

$$\delta = \Sigma^{-1} (\mu_1 - \mu_2),$$

we have the following

||| Theorem 5.6

The vector δ has the property that it maximizes the function

$$g(\mathbf{d}) = \frac{[E_1(\mathbf{X}^T \mathbf{d}) - E_2(\mathbf{X}^T \mathbf{d})]^2}{V(\mathbf{X}^T \mathbf{d})} = \frac{[(\mu_1 - \mu_2)^T \mathbf{d}]^2}{\mathbf{d}^T \Sigma \mathbf{d}}$$

||| Proof

The proof is straightforward. Since we readily get that $g(k\mathbf{d}) = g(\mathbf{d})$ we can determine extremes for g by determining extremes for the numerator under the constraint

$$\mathbf{d}^T \Sigma \mathbf{d} = 1$$

We introduce a Lagrange multiplier λ and seek the maximum of

$$\psi(\mathbf{d}) = [(\mu_1 - \mu_2)^T \mathbf{d}]^2 - \lambda (\mathbf{d}^T \Sigma \mathbf{d} - 1).$$

Now we have that

$$\frac{\partial \psi}{\partial \mathbf{d}} = 2(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \mathbf{d} - 2\lambda \Sigma \mathbf{d}$$

If we let this equal 0, we have

$$(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \mathbf{d} = \lambda \Sigma \mathbf{d}$$

i.e.

$$\mathbf{d} = \left\{ \frac{1}{\lambda} (\mu_1 - \mu_2)^T \mathbf{d} \right\} \Sigma^{-1} (\mu_1 - \mu_2) = k\delta$$

where k is a scalar.

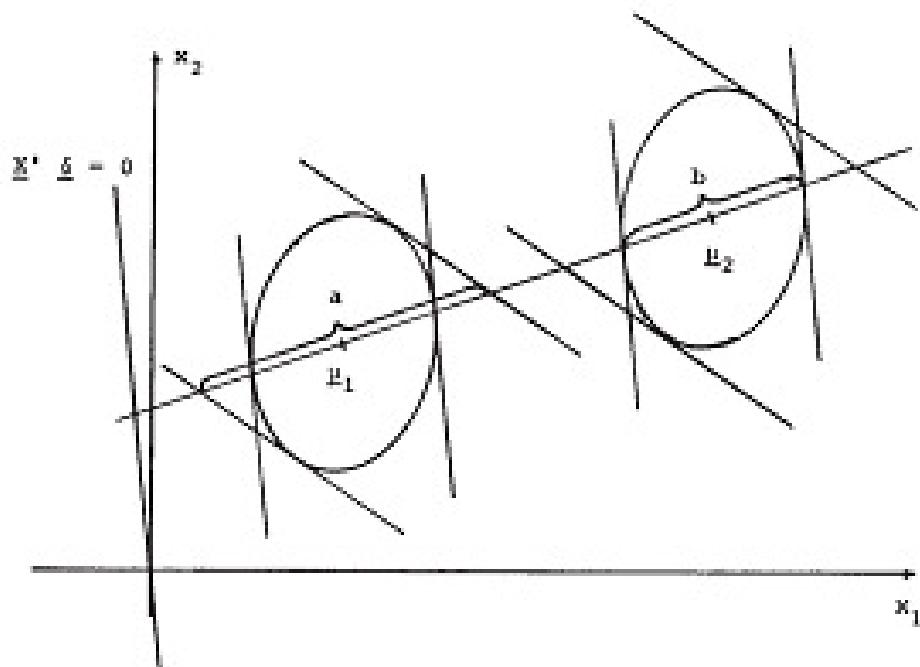


|||| **Remark 5.7**

The content of the theorem is that the linear function determined by

$$\mathbf{X}^T \boldsymbol{\delta} = \delta_1 X_1 + \cdots + \delta_p X_p$$

is the projection that “moves” π_1 furthest possible away from π_2 measured in units of the standard deviation of the projected distributions or - in analysis of variance terms - the projection which maximizes the variance between populations divided by the total variance.



The geometrical content of the theorem is indicated in the above figure where

- b: is the projection of the ellipse onto the line $(\mu_1; \mu_2)$ in the direction determined by $\mathbf{X}^T \boldsymbol{\delta} = 0$
- a: is the projection of the ellipse onto the line $(\mu_1; \mu_2)$ in a different direction.

It is seen that the projection determined by $\boldsymbol{\delta}$ onto the line which connects μ_1 and μ_2 is the one which “moves” the projection of the contour ellipsoids of the distributions corresponding to the two populations furthest possible away from each other.

We now give a theorem which is very useful in the determination of misclassification probabilities.

||| Theorem 5.8

We consider the random variable defined by the linear discriminator (omitting the term $-\ln c$), i.e.

$$Z = \mathbf{X}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 .$$

Then

$$Z \sim \begin{cases} N\left(+\frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}^2, \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}^2\right) & \text{if } \pi_1 \text{ is true} \\ N\left(-\frac{1}{2}\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}^2, \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}^2\right) & \text{if } \pi_2 \text{ is true} \end{cases} .$$

||| Proof

The proof is straight forward. Let us e.g. consider the case π_1 true. We then have that $E(\mathbf{X}) = \boldsymbol{\mu}_1$ and then

$$\begin{aligned} E(Z) &= \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 \\ &= \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \frac{1}{2} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}^2 \end{aligned}$$

$$\begin{aligned} V(Z) &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ &= \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_{\boldsymbol{\Sigma}^{-1}}^2 \end{aligned}$$

The result regarding π_2 is shown analogously. ■

We will now consider some examples.

||| Example 5.9

We consider the case where

$$\pi_1 \leftrightarrow N\left(\begin{bmatrix} 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\right)$$

$$\pi_2 \leftrightarrow N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\right)$$

and we want to determine a “best” discriminator function. Since we know nothing about the prior probabilities and losses, we will use the function which corresponds to the constant c in theorem 5.4 being 1. Since

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$$

we get the following function

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} - \frac{1}{2}(2 \cdot 16 + 1 \cdot 4 - 2 \cdot 8) + \frac{1}{2}(2 \cdot 1 + 1 \cdot 1 - 2 \cdot 1) = 0$$

or

$$5x_1 - 2x_2 - 9\frac{1}{2} = 0$$

If we enter an arbitrary point, e.g. $\begin{bmatrix} 5 \\ 6 \end{bmatrix}$ we get

$$5 \cdot 5 - 2 \cdot 6 - 9\frac{1}{2} = 3\frac{1}{2} > 0 .$$

This point is therefore classified as coming from π_1 .

If we have a loss function, the procedure is a bit different which is seen from

||| Example 5.10

Let us assume that we have losses assigned to the different decisions:

From		Classify as	
		π_1	π_2
Nature	π_1	0	$L_{12} = 2$
	π_2	$L_{21} = 1$	0

Since we have no prior probabilities we will determine the minimax solution. We will need

$$\|\mu_1 - \mu_2\|_{\Sigma^{-1}}^2 = [3 \ 1] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 2 \cdot 9 + 1 \cdot 1 - 2 \cdot 3 \cdot 1 = 13$$

From theorem 5.2 follows that we must determine c so

$$2P\left\{\frac{f_1(x)}{f_2(x)} < c \mid \pi_1\right\} = P\left\{\frac{f_1(x)}{f_2(x)} \geq c \mid \pi_2\right\}$$

$$\Leftrightarrow 2P\{Z < \ln c \mid \pi_1\} = P\{Z \geq \ln c \mid \pi_2\}$$

$$\begin{aligned} &\Leftrightarrow 2P\left\{N\left(\frac{1}{2} \cdot 13, 13\right) < \ln c\right\} = P\left\{N\left(-\frac{1}{2} \cdot 13, 13\right) \geq \ln c\right\} \\ &\Leftrightarrow 2P\left\{N(0, 1) < \frac{\ln c - 6.5}{\sqrt{13}}\right\} = P\left\{N(0, 1) \geq \frac{\ln c + 6.5}{\sqrt{13}}\right\} \end{aligned}$$

By trying with different values of c we see that

$$c \simeq 0.5617 .$$

Using this value, the misclassification probabilities are

$$\text{If } \pi_1 \text{ is true : } P\left\{N(0, 1) < \frac{\ln 0.5617 - 6.5}{\sqrt{13}}\right\} \simeq 0.025$$

$$\text{If } \pi_2 \text{ is true : } P\left\{N(0, 1) \geq \frac{\ln 0.5617 + 6.5}{\sqrt{13}}\right\} \simeq 0.050$$

The discriminating line is now determined by

$$5x_1 - 2x_2 - 9.5 = \ln 0.5617$$

or

$$5x_1 - 2x_2 - 8.92 = 0$$

This line intersects the line connecting μ_1 and μ_2 in $(2.36, 1.46)^T$ i.e. it is moved towards μ_2 compared to the mid-point $(2.5, 1.5)^T$. It is also obvious that the line is moved parallelly in this direction since we see from the loss matrix that it is more serious to be wrong if π_1 is true than if π_2 is true. Therefore we must expand R_1 i.e. move the limiting line towards μ_2 .

We must stress that it is of importance that the dispersion matrices for the two populations are equal. If this is not the case we will get a completely different result which will be seen from the following example.

||| Example 5.11

Let us assume that the dispersion matrix for population 2 is changed to an identity matrix i.e.

$$\pi_1 \leftrightarrow N\left(\begin{bmatrix} 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\right)$$

$$\pi_2 \leftrightarrow N\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

Again we want to classify an observation X which comes from one of the above mentioned distributions. Since the dispersion matrices are not equal we cannot use the result in theorem 5.4 but have to start from the beginning with theorem 5.2.

For $c > 0$ we have

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c \Leftrightarrow -(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + (\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \geq 2\ln c$$

Since

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) &= 2(x_1 - 4)^2 + (x_2 - 2)^2 - 2(x_1 - 4)(x_2 - 2) \\ &= 2x_1^2 + x_2^2 - 2x_1x_2 - 12x_1 + 4x_2 + 20, \end{aligned}$$

and

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) &= (x_1 - 1)^2 + (x_2 - 1)^2 \\ &= x_1^2 + x_2^2 - 2x_1 - 2x_2 + 2, \end{aligned}$$

then

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c \Leftrightarrow -x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 \geq 2\ln c$$

If we choose $c = 1$, we note that the curve which separates R_1 and R_2 is the hyperbola

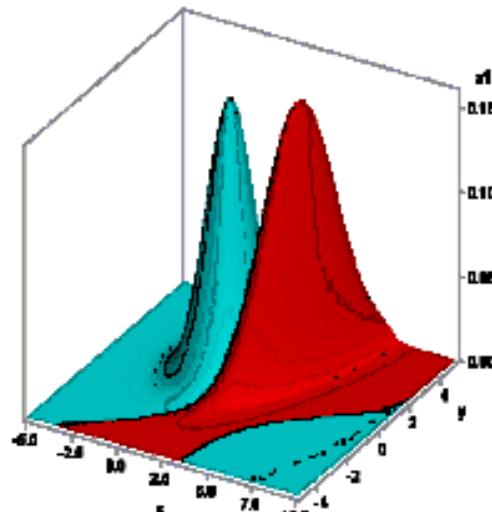
$$\{\mathbf{x} \mid -x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 = 0\}$$

It has center in $(3, -2)^T$ and asymptotes

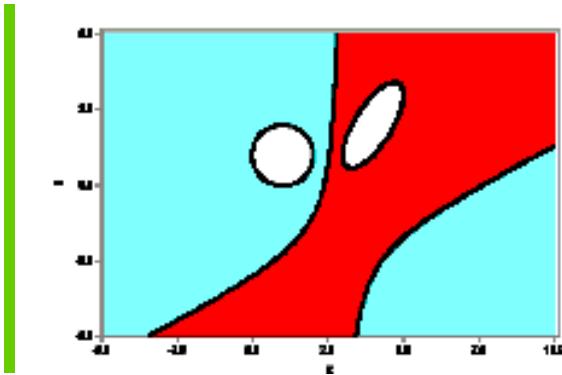
$$x_1 - 3 = 0$$

$$x_1 - 2x_2 - 7 = 0$$

These curves are shown in the figure below together with the contour ellipses for the two normal distributions. Note e.g. that a point such as $(9, 0)^T$ is in R_2 and therefore will be classified as coming from the distribution with center in $(1, 1)^T$. Further-



more the frequency functions are shown.



We will not consider the problem of misclassification probabilities in cases as the above mentioned where we have quadratic discriminators.

5.1.3 Discrimination with unknown parameters

If one does not know the two distributions f_1 and f_2 one must estimate them based on some observations and then construct discriminators from the estimated distributions the same way we did for the exact distributions.

Let us consider the normal case

$$\pi_1 \leftrightarrow N(\mu_1, \Sigma)$$

$$\pi_2 \leftrightarrow N(\mu_2, \Sigma)$$

where the parameters are unknown. If we have observations X_{11}, \dots, X_{1n_1} which we know come from π_1 and observations X_{21}, \dots, X_{2n_2} which we know come from π_2 we can estimate the parameters as usual:

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} X_{1j} = \bar{X}_1 \quad \text{and} \quad \hat{\Sigma}_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)(X_{1j} - \bar{X}_1)^T$$

$$\hat{\mu}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{2j} = \bar{X}_2 \quad \text{and} \quad \hat{\Sigma}_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)(X_{2j} - \bar{X}_2)^T$$

and with $N = n_1 + n_2$ the **pooled estimate** of the dispersion matrix

$$\hat{\Sigma} = \frac{1}{N - 2} \left[(n_1 - 1)\hat{\Sigma}_1 + (n_2 - 1)\hat{\Sigma}_2 \right]$$

We now estimate the appropriate decision rule by plugging these estimators into the formula given by theorem 5.4, i.e.

$$\mathbf{x}^T \widehat{\Sigma}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) - \frac{1}{2} \widehat{\boldsymbol{\mu}}_1^T \widehat{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_1 + \frac{1}{2} \widehat{\boldsymbol{\mu}}_2^T \widehat{\Sigma}^{-1} \widehat{\boldsymbol{\mu}}_2 .$$

The exact distribution of this quantity if we substitute \mathbf{x} with a random variable $\mathbf{X} \sim N(\boldsymbol{\mu}_i, \Sigma)$ is fairly complicated but for large sample sizes it is asymptotically equal to the distribution of Z in theorem 5.8 so for reasonable sample sizes we can use the theory we have derived.

The estimated norm between the expected values is

$$D^2 = \| \widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2 \|_{\widehat{\Sigma}^{-1}}^2 = (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^T \widehat{\Sigma}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$$

This is called the **empirical Mahalanobis' distance** as opposed to the **(theoretical) Mahalanobis' distance**

$$\Delta^2 = \| \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \|_{\Sigma^{-1}}^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

It should here be noted that a number of authors as well as statistical software packages use the expression Mahalanobis' distance also about the empirical Mahalanobis distance. The name is in honor of the Indian statistician P.C. Mahalanobis who developed discriminant analysis at the same time as the English statistician R.A. Fisher in the 1930's.

We see that D^2 is closely related to Hotelling's T^2 statistic for the two sample situation. More specifically

$$D^2 = \frac{n_1 + n_2}{n_1 n_2} T^2$$

Therefore we can test whether $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ by means of D^2 . We give the results in the following theorem.

||| Theorem 5.12

Using the significance level α , the critical area for a test of the hypothesis $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ against all alternatives becomes

$$C = \left\{ \mathbf{x}_{11}, \dots, \mathbf{x}_{1n_1}, \mathbf{x}_{21}, \dots, \mathbf{x}_{2n_2} \mid \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \cdot \frac{n_1 n_2}{n_1 + n_2} d^2 > F(p, n_1 + n_2 - p - 1)_{1-\alpha} \right\}$$

Here d^2 is the observed value of D^2 .

|||| **Proof**

An immediate consequence of the relation to Hotellings T^2 statistic.

■

5.1.4 Test for best discrimination function

We remind ourselves that the best discriminator

$$\hat{\delta} = \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$$

can be found by maximizing the function

$$\hat{g}(\mathbf{d}) = \frac{[(\hat{\mu}_1 - \hat{\mu}_2)^T \mathbf{d}]^2}{\mathbf{d}^T \hat{\Sigma} \mathbf{d}}$$

The maximum value is

$$\hat{g}(\hat{\delta}) = \frac{[(\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)]^2}{(\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)} = D^2$$

i.e. Mahalanobis' D^2 is the maximum value of $\hat{g}(\mathbf{d})$.

For an arbitrary (fixed) \mathbf{d}_0 we now let

$$D_0^2 = \hat{g}(\mathbf{d}_0) = \frac{[(\hat{\mu}_1 - \hat{\mu}_2)^T \mathbf{d}_0]^2}{\mathbf{d}_0^T \hat{\Sigma} \mathbf{d}_0}$$

|||| **Theorem 5.13**

The statistic

$$Z = \frac{n_1 + n_2 - p - 1}{p - 1} \cdot \frac{n_1 n_2 (D^2 - D_0^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_0^2}$$

may be used in testing the hypothesis that the linear projection determined by d_0 is the best discriminator against all alternatives. Z is $F(p - 1, n_1 + n_2 - p - 1)$ -distributed under the hypothesis and large values of Z are critical, i.e., the critical region is

$$C = \{x_{11}, \dots, x_{1n_1}, x_{21}, \dots, x_{2n_2} \mid z > F(p - 1, n_1 + n_2 - p - 1)_{1-\alpha}\}$$

if we use the significance level α . Here z is the observed value of Z .

|||| **Proof**

We shall not go into details with the proof but just note that Z gives a measure of how much the “distance” between the two populations is reduced by using d_0 instead of $\hat{\delta}$. If this reduction is too big i.e. if Z is large, we will not be able to assume that d_0 gives essentially as good a discrimination between the two populations as $\hat{\delta}$.

■

5.2 Discrimination between several populations

5.2.1 The Bayes solution

The main idea of the generalization in this section is that one compares the populations pairwise in the same way as in the previous section and finally selects the most probable population.

We consider the populations

$$\pi_1, \dots, \pi_k.$$

Based on measurements of p characteristics (or variables) of a given individual we wish to classify it as coming from one of the populations π_1, \dots, π_k . The

observed measurement is

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$$

If the individual comes from π_k then the frequency function for \mathbf{X} is $f_i(\mathbf{x})$.

We assume that a loss function L is given as shown in the following table.

		Classify as				
		π_1	\cdots	π_i	\cdots	π_k
Nature	π_1	0	\cdots	L_{1i}	\cdots	L_{1k}
	\vdots	\vdots				\vdots
	π_v	L_{v1}	\cdots	L_{vi}	\cdots	L_{vk}
	\vdots	\vdots		\vdots		\vdots
	π_k	L_{k1}	\cdots	L_{ki}	\cdots	0

Finally we assume we have a prior distribution

$$g(\pi_i) = p_i, \quad i = 1, \dots, k$$

The posterior distribution becomes

$$k(\pi_v | \mathbf{x}) = \frac{g(\pi_v)f_v(\mathbf{x})}{g(\pi_1)f_1(\mathbf{x}) + \dots + g(\pi_k)f_k(\mathbf{x})} = \frac{p_v f_v(\mathbf{x})}{p_1 f_1(\mathbf{x}) + \dots + p_k f_k(\mathbf{x})} = \frac{p_v f_v(\mathbf{x})}{h(\mathbf{x})}$$

For an individual with the observation \mathbf{x} we define the **discriminant value** or **discriminant score** for the i 'th population as

$$S_i^*(\mathbf{x}) = S_i^* = -[p_1 f_1(\mathbf{x})L_{1i} + \dots + p_k f_k(\mathbf{x})L_{ki}]$$

(note that $L_{ii} = 0$ so the sum has no term $p_i f_i(\mathbf{x})L_{ii}$).

We see that for the i 'th population, S_i^* is a constant ($-h(\mathbf{x})$) times the expected loss with respect to the posterior distribution. Since the proportionality factor $-h(\mathbf{x})$ is negative it follows that the Bayes' solution to the decision problem is to select the population which has the largest discriminant value (discriminant score) i.e.

$$\text{we select } \pi_v \text{ if } S_v^* \geq S_i^* \quad \forall i$$

If all losses L_{ij} ($i \neq j$) are equal, we can simplify the expression for the discriminant score:

$$\text{we select } \beta_i \text{ rather than } \beta_j \text{ if } S_v^* > S_i^*$$

$$\Leftrightarrow -\left(\sum_{\nu} p_{\nu} f_{\nu}(\boldsymbol{x}) - p_i f_i(\boldsymbol{x})\right) > -\left(\sum_{\nu} p_{\nu} f_{\nu}(\boldsymbol{x}) - p_j f_j(\boldsymbol{x})\right)$$

$$\Leftrightarrow p_i f_i(\boldsymbol{x}) > p_j f_j(\boldsymbol{x}).$$

In this case we can therefore choose the discriminant score

$$S_i = S_i(\boldsymbol{x}) = p_i f_i(\boldsymbol{x})$$

In this case the Bayes' rule is that we select the population which has the largest posterior distribution, i.e.

$$\text{select } \operatorname{argmax}_i k(\pi_i | \boldsymbol{x})$$

This rule is not only used where the losses are equal but also where it has not been possible to determine losses. If the prior probabilities p_i are unknown and it is not possible to estimate them one usually uses the discriminant score

$$S'_i = S'_i(\boldsymbol{x}) = f_i(\boldsymbol{x})$$

i.e. we select the population with the largest value of the probability density function.

The minimax solutions are determined by choosing the strategy which makes all the misclassification probabilities equally large (still assuming that all losses are equal). However, we shall not go into any further detail on that matter.

5.2.2 The Bayes' solution in the case with several normal distributions

For $i = 1, \dots, k$ we consider populations

$$\pi_i \leftrightarrow N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \text{ with prior probabilities } p_i$$

i.e. the density functions are

$$f_i(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{|\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)\right),$$

for $i = 1, \dots, k$.

Since we get the same decision rule by choosing monotone transformations of our discriminant scores we will take the logarithm of the f_i 's and disregard the common factor $1/\sqrt{2\pi^p}$. This gives (assuming that the losses are equal)

$$S_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \ln p_i$$

If the p_i are equal or unknown it is customary to remove the last term from the expression, and with a slight abuse of notation we still use the terminology $S_i^Q(\mathbf{x})$. This function is quadratic in \mathbf{x} and is called a **quadratic discriminant function (QDF)**. The relationship with the posterior distribution becomes

$$k^Q(\pi_v | \mathbf{x}) = \frac{\exp(S_v^Q(\mathbf{x}))}{\sum_{i=1}^k \exp(S_i^Q(\mathbf{x}))}.$$

If all the Σ_i are equal then the terms

$$-\frac{1}{2}\ln|\Sigma| - \frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$$

are common for all $S_i^Q(\mathbf{x})$ and can therefore be omitted. We then get

$$S_i^L(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \ln p_i$$

Similar remarks on equal or unknown priors as given above apply here. $S_i^L(\mathbf{x})$ is a linear (affine) function in \mathbf{x} and is called a **linear discriminant function or LDF**. Like in the quadratic case we have the posterior probability

$$k^L(\pi_v | \mathbf{x}) = \frac{\exp(S_v^L(\mathbf{x}))}{\sum_{i=1}^k \exp(S_i^L(\mathbf{x}))}.$$

If there are only two groups we note that we choose group 1 if

$$S_1^L(\mathbf{x}) - S_2^L(\mathbf{x}) \geq 0 \Leftrightarrow \mathbf{x}^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 > \ln \frac{p_2}{p_1}$$

i.e. the same result as in theorem 5.4.

It is of course possible to describe the decision rules by dividing \mathbb{R}^p into sets R_1, \dots, R_k so that we choose π_i exactly when $\mathbf{x} \in R_i$. Among other things this can be seen from the following

||| Example 5.14

We consider populations π_1 , π_2 and π_3 given by normal distributions with expected values

$$\mu_1 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mu_3 = \begin{bmatrix} 2 \\ 6 \end{bmatrix},$$

and common dispersion matrix

$$\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

Assuming that all p_i are equal so that we may disregard them in the discriminant scores, we get

$$S_1^L(\mathbf{x}) = [x_1 \ x_2] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} - \frac{1}{2} [4 \ 2] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \end{bmatrix} = 6x_1 - 2x_2 - 10$$

$$S_2^L(\mathbf{x}) = [x_1 \ x_2] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{2} [1 \ 1] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = x_1 - \frac{1}{2}$$

$$S_3^L(\mathbf{x}) = [x_1 \ x_2] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \end{bmatrix} - \frac{1}{2} [2 \ 6] \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 6 \end{bmatrix} = -2x_1 + 4x_2 - 10$$

We prefer π_1 to π_2 if

$$u_{12}(\mathbf{x}) = (6x_1 - 2x_2 - 10) - \left(x_1 - \frac{1}{2}\right) = 8x_1 - 6x_2 > 0 -$$

We prefer π_1 to π_3 if

$$u_{13}(\mathbf{x}) = (6x_1 - 2x_2 - 10) - (-2x_1 + 4x_2 - 10) = 5x_1 - 2x_2 - 9\frac{1}{2} > 0$$

and finally we prefer π_2 to π_3 if

$$u_{23}(\mathbf{x}) = \left(x_1 - \frac{1}{2}\right) - (-2x_1 + 4x_2 - 10) = 3x_1 - 4x_2 + 9\frac{1}{2} > 0.$$

It is now evident that we will choose π_1 if both $u_{12}(\mathbf{x}) > 0$ and $u_{13}(\mathbf{x}) > 0$ and analogously with the others. We can therefore define the regions

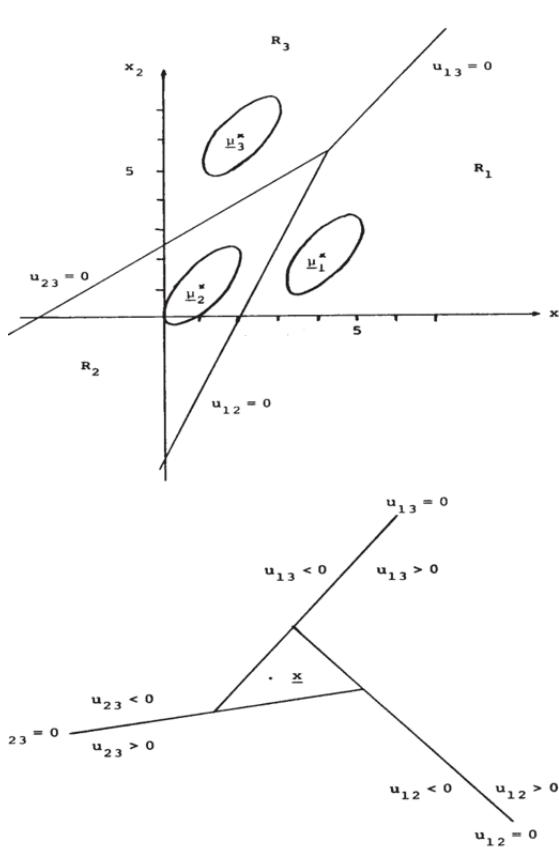
$$R_1 = \{\mathbf{x} \mid u_{12}(\mathbf{x}) > 0 \wedge u_{13}(\mathbf{x}) > 0\}$$

$$R_2 = \{\mathbf{x} \mid u_{12}(\mathbf{x}) < 0 \wedge u_{23}(\mathbf{x}) > 0\}$$

$$R_3 = \{\mathbf{x} \mid u_{13}(\mathbf{x}) < 0 \wedge u_{23}(\mathbf{x}) < 0\}$$

and we have that we choose π_i exactly when $\mathbf{x} \in R_i$. We have sketched the situation in the figure below.

One can easily prove that the lines will intersect in a point. It is, however, also possible to make a simple reasoning for this. Let us assume that the situation is as in the figure below.



We now note that

$$u_{ij}(x) > 0 \iff f_i(x) > f_j(x)$$

For the point x we have

$$\left. \begin{array}{l} u_{23}(x) < 0 \quad \text{i.e.} \quad f_2(x) < f_3(x) \\ u_{13}(x) > 0 \quad \text{i.e.} \quad f_1(x) > f_3(x) \end{array} \right\} \Rightarrow f_1(x) > f_2(x)$$

$$u_{12}(x) < 0 \quad \text{i.e.} \quad f_1(x) < f_2(x)$$

We have now established a contradiction i.e. the three lines determined by $u_{12}(x)$, $u_{13}(x)$ and $u_{23}(x)$ must intersect each other in one single point.

5.2.3 The case with several normal distributions and unknown parameters

As in the case with only two populations we estimate the unknown parameters and plug the estimates into the expressions obtained in the case with known parameters.

For $i = 1, \dots, k$ we consider populations

$$\pi_i \leftrightarrow N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \text{ with prior probabilities } p_i$$

and observations

$$X_{i1}, \dots, X_{in_i}$$

We define the within groups, between groups and total sums of squares matrices

$$\begin{aligned} W &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T \\ B &= \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T \\ T &= \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X})^T. \end{aligned}$$

and reminding of the fundamental equation

$$T = B + W.$$

We have the following estimates of means and dispersion matrices:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij} \\ \hat{\boldsymbol{\Sigma}}_i &= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T = \frac{1}{n_i - 1} W_i \end{aligned}$$

If the hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_k$ is true, we use the **pooled estimate** of the dispersion matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N - k} \sum_{i=1}^k (n_i - 1) \hat{\boldsymbol{\Sigma}}_i = \frac{1}{N - k} \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T = \frac{1}{N - k} W$$

where $N = \sum n_i$. Still assuming that the hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \dots = \boldsymbol{\Sigma}_k$ is true, we define the **squared generalized distance from $\hat{\boldsymbol{\mu}}_j$ to population π_i** as

$$D_i^2(\hat{\boldsymbol{\mu}}_j) = \begin{cases} (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_j) & \text{if the priors are equal} \\ (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_j)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_i - \hat{\boldsymbol{\mu}}_j) - 2 \ln p_i & \text{if the priors are not all equal} \end{cases}$$

If the hypothesis is not true, we define the **squared generalized distance from $\hat{\boldsymbol{\mu}}_j$ to population π_i** as

$$D_i^2(\hat{\mu}_j) = \begin{cases} (\hat{\mu}_i - \hat{\mu}_j)^T \hat{\Sigma}_i^{-1} (\hat{\mu}_i - \hat{\mu}_j) & \text{if the priors are equal} \\ (\hat{\mu}_i - \hat{\mu}_j)^T \hat{\Sigma}_i^{-1} (\hat{\mu}_i - \hat{\mu}_j) + \ln |\hat{\Sigma}_i| - 2\ln p_i & \text{if the priors are not all equal} \end{cases}$$

We see that the first of those 4 expressions is equal to the estimated squared **Mahalanobis distance** between populations π_i and π_j , using the estimate $\hat{\Sigma}$ of the dispersion matrix not only based on observations from populations π_i and π_j but from all k groups. This estimate has $k - 1$ degrees of freedom, and according to lemma 5.1, the quantity

$$\frac{n_i n_j}{n_i + n_j} \frac{N - k - p + 1}{(N - k) p} D_i^2(\hat{\mu}_j) = \frac{n_i n_j}{n_i + n_j} \frac{N - k - p + 1}{(N - k) p} (\hat{\mu}_i - \hat{\mu}_j)^T \hat{\Sigma}^{-1} (\hat{\mu}_i - \hat{\mu}_j)$$

will follow an $F(p, N - k - p + 1)$ distribution if $\mu_i = \mu_j$. This result can be used in testing the hypothesis $\mu_i = \mu_j$ against the alternative $\mu_i \neq \mu_j$. Critical values are large values of the test statistic. For $k = 2$ the test statistic coincides with the test statistic based on the usual version of Mahalonobis' distance presented in section 5.1.3.

Remark: If we look at the most general expression of the squared distance form an observation x to population π_i we get (replacing $\hat{\mu}_j$ with x)

$$D_i^2(x) = (x - \hat{\mu}_i)^T \hat{\Sigma}_i^{-1} (x - \hat{\mu}_i) + \ln |\hat{\Sigma}_i| - 2\ln p_i$$

which is the estimate of minus two times the formula for $S_i^Q(x)$ given earlier. Thus the discriminant score becomes

$$-\frac{1}{2} D_i^2(x)$$

and the estimated posterior probability for group i becomes

$$\hat{k}^Q(\pi_i | x) = \frac{\exp(-\frac{1}{2} D_i^2(x))}{\sum_{j=1}^k \exp(-\frac{1}{2} D_j^2(x))}$$

5.2.4 Short about kernel estimates and nearest neighbor estimates

Often a multivariate normal distribution will not be adequate for modelling the group specific probability density functions. In the last couple of decades kernel based methods have become popular as a non-parametric alternative. We shall briefly touch upon the use of Gaussian kernels in discrimination and classification. Related to the use of uniform kernels are the k-nearest neighbor estimates of the different group densities. We shall very briefly illustrate the use of such methods.

For each group π_i we define a kernel that is the density function of a multivariate normal distribution with mean 0 and dispersion matrix $r^2 \hat{\Sigma}_i$, i.e.

$$K_i(z) = \frac{1}{r^p (2\pi)^{p/2} |\hat{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2r^p} z^T \hat{\Sigma}_i^{-1} z\right)$$

We then estimate the density function for group π_i as

$$\hat{f}_i(x) = \frac{1}{n_i} \sum_{j=1}^{n_i} K_i(x - x_{ij})$$

In this expression, r acts as a smoothing parameter that must be fixed in one way or another. Often it will suffice evaluate the outcome of different choices. A possible choice for the smoothing parameter r is

$$\left[\frac{4}{n_i (2p+1)} \right]^{1/(p+4)}$$

This value possesses some optimality properties, but we shall not go into any further detail here. This value may be used in the SAS procedure DISCRIM.

Another way of obtaining a non-parametric estimate of the pdf for the i 'th group is to use a **k nearest neighbor** method (k-NN). For a given point x the squared distance to the k 'th nearest neighbor from the training set, say x_{st} , is given by

$$r_k^2(x) = \|x - x_{st}\|_{\hat{\Sigma}^{-1}}^2 = (x - x_{st})^T \hat{\Sigma}^{-1} (x - x_{st})$$

The number of observations from the training set that are within the ellipsoid

$$E_{r_k(x)}(x) = \left\{ z \mid (z - x)^T \hat{\Sigma}^{-1} (z - x) \leq r_k^2(x) \right\}$$

is thus k (we ignore the problem with possible ties). Let k_i of those come from group π_i . Then we have the estimate

$$\hat{f}_i(x) = \frac{1}{v(r_k(x))} \frac{k_i}{n_i} = \frac{\Gamma(1 + \frac{p}{2})}{\pi^{p/2} |\hat{\Sigma}|^{1/2}} \frac{1}{[r_k(x)]^p} \frac{k_i}{n_i}$$

where $v(r_k(\mathbf{x}))$ is the volume of the ellipsoid $E_{r_k(\mathbf{x})}(\mathbf{x})$.

For classification purposes it is important to note that the only class dependent part of this pdf is the last fraction $\frac{k_i}{n_i}$ which simply is the relative fraction of the training set observations that is within the ellipsoid $E_{r_k(\mathbf{x})}(\mathbf{x})$ around \mathbf{x} . In the posterior distribution this fraction will of course be modified by the prior probabilities. In these expressions k also acts as a smoothing constant and may likewise be determined by evaluating outcomes from using different values of k .

|||| Remark 5.15

The above represents a statistical approach to the k-NN method. In computer science, the k-NN method will in general be simpler based on the k nearest neighbors using the Euclidian distance and the classify using a simple majority vote principle.

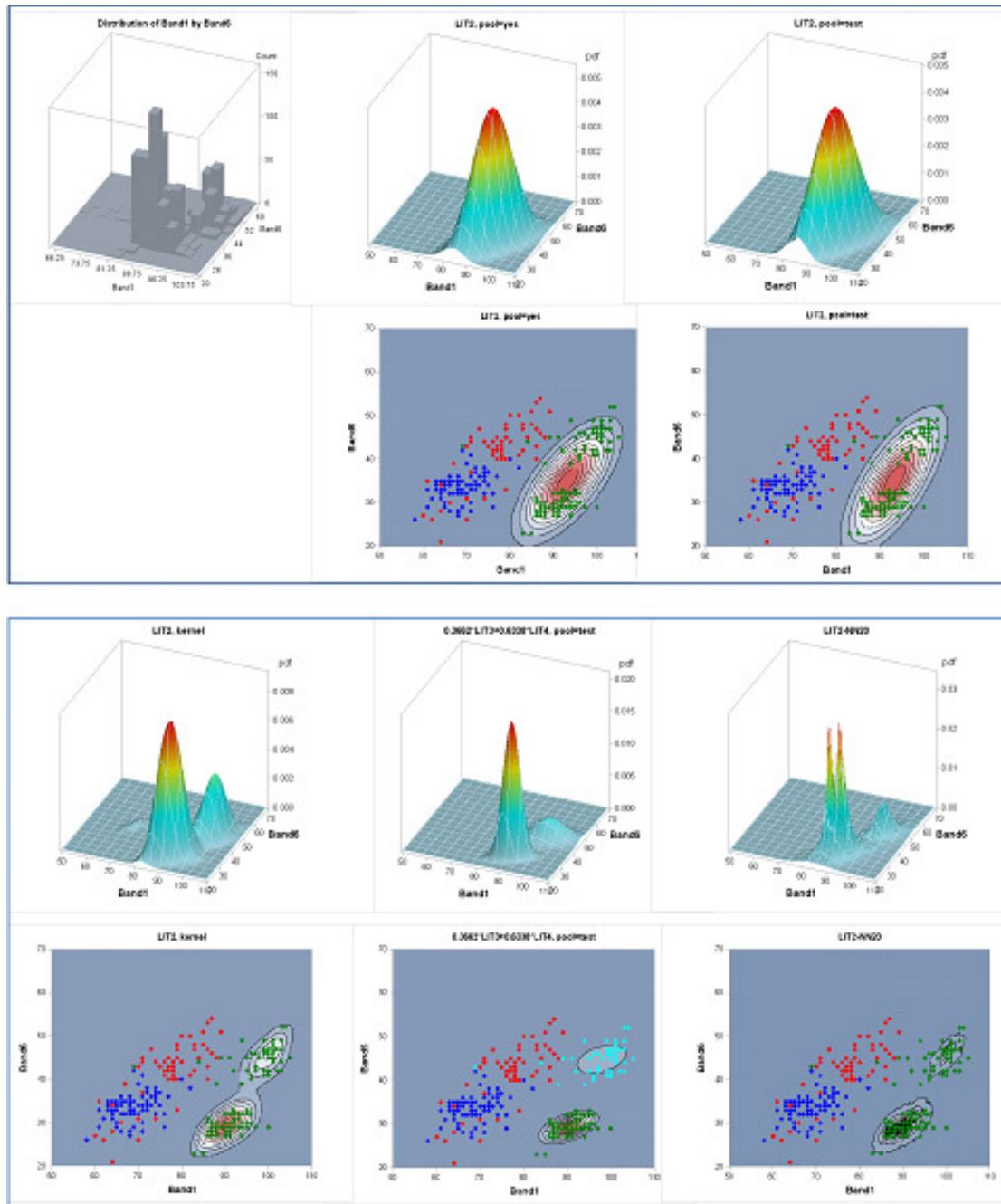


Figure 5.5: Histogram and estimated distributions for unit 2, deltas and young alluvial fans using Landsat Band1 and band6. In the top frame the distribution is estimated as a single normal distribution, in the fist case with the pooled dispersion matrix, in the second with the within class 2 estimated dispersion. In the lower frame: Left: Using a kernel estimate. Middle: Using a compound distribution. Right: Using a nearest neighbor estimate. All pdf-plots are accompanied by a scatterplot overlayed on a contour plot.

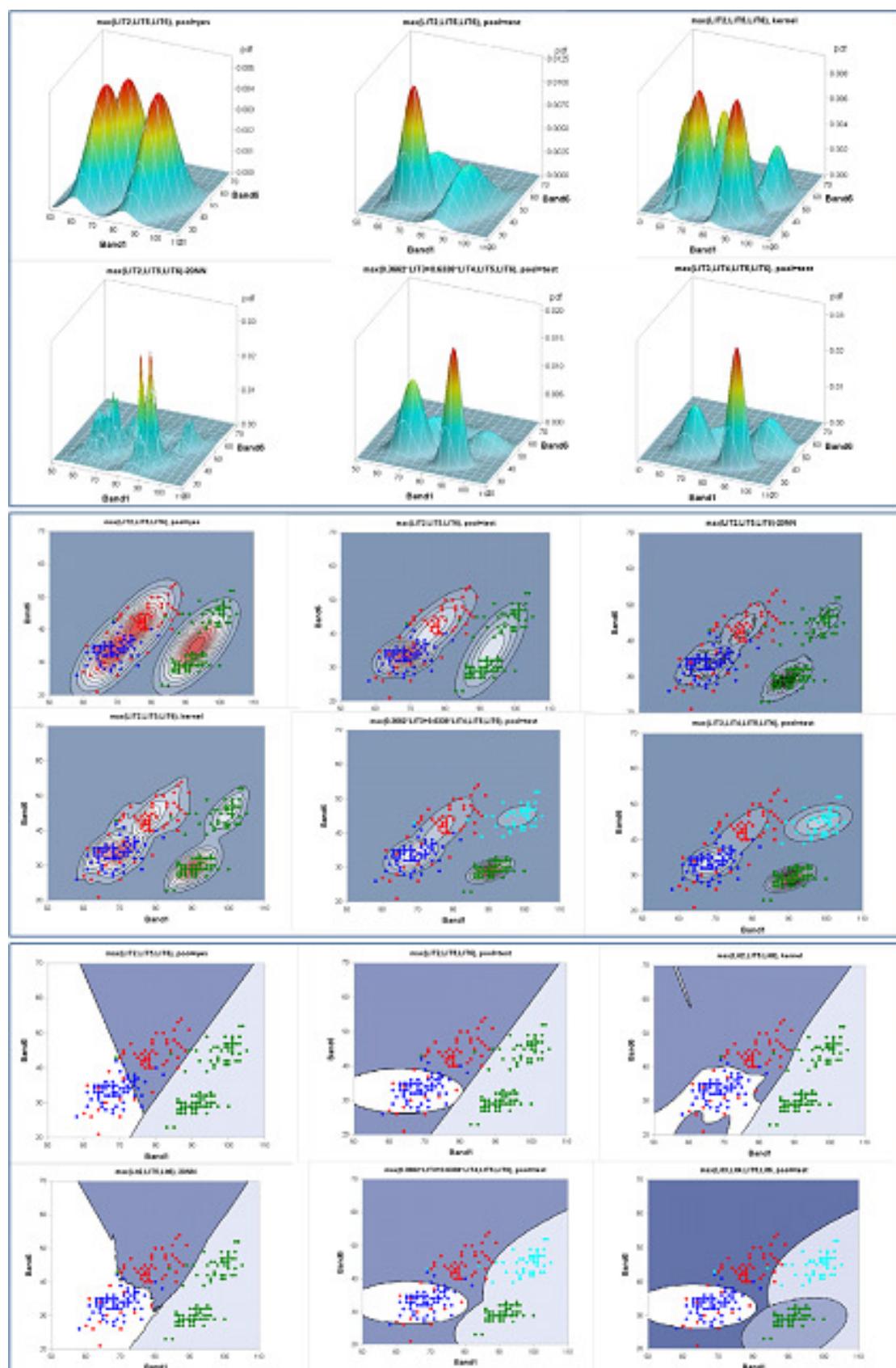


Figure 5.6: The classification regions corresponding to the different types of estimate of the pdf's.

	Number of observations from	Classified as							Sum
		π_1	...	π_i	...	π_j	...	π_k	
Nature	π_1	N_{11}	...	N_{1i}	...	N_{1j}	...	N_{1k}	$N_{1..}$
	\vdots	\vdots		\vdots		\vdots		\vdots	\vdots
	π_i	N_{i1}	...	N_{ii}	...	N_{ij}	...	N_{ik}	$N_{i..}$
	\vdots	\vdots		\vdots		\vdots		\vdots	\vdots
	π_j	N_{j1}	...	N_{ji}	...	N_{jj}	...	N_{jk}	$N_{j..}$
	\vdots	\vdots		\vdots		\vdots		\vdots	\vdots
	π_k	N_{k1}	...	N_{ki}	...	N_{kj}	...	N_{kk}	$N_{k..}$
Sum		$N_{..1}$...	$N_{..i}$...	$N_{..j}$...	$N_{..k}$	$N_{..}$

Table 5.2: Confusion matrix showing the result of classifying $N_{..}$ observations. N_{ij} is the number of observations from population π_i classified as coming from population π_j .

5.3 Evaluation

5.3.1 Some performance measures for a classifier

The evaluation of the quality of a classifier is most commonly based on how well observations from a test data set with known classes are classified. Let us more specifically assume that we have $N_{..}$ observations that are classified yielding the results in table 6.x. Such a table is often called a **confusion matrix**.

The confusion matrix can be based on different (types of) test data. Three popular choices are

1. **Resubstitution** classification of the training data set: Each observation in the training data set is classified using the estimated discriminant function.
2. **Cross Validation** of the training data set: Each observation in the training data set is classified using a discriminant function computed from the other observations in the training data set excluding the observation being classified.
3. **Set Aside** classification: Divide the input data set randomly into two data sets, the training and the test data set. Estimate the discriminant functions on the training data set and test them on the test data set.

Since the classifier is trained to fit the training data set, the resubstitution method will normally be overoptimistic with respect to future performance. The cross validation method will reduce the bias considerably, and in “well behaved” cases give error estimates with a small bias. If the set aside dataset is independent of the training data set, this method will provide unbiased estimators of the error rates.

A more elaborate way of splitting the data is to divide it into k parts – **k -fold cross validation**. Using the terminology from the SAS Procedure GLMSelect we may consider

1. Block(k) where the k parts are made of blocks of $\text{int}(n/k)$ or $\text{int}(n/k) + 1$ successive observations, where n is the number of observations.
2. Split(k) where the parts consist of observations $\{1, k + 1, 2k + 1, 3k + 1, \dots\}, \{2, k + 2, 2k + 2, 3k + 2, \dots\}, \dots, \{k, 2k, 3k, \dots\}$.
3. Random(k) where the data are partitioned in random subsets each with roughly $\text{int}(n/k)$ observations.
4. Variable where we use the formatted value of an input data set variable to define the parts in cases where one needs to exercise extra control over how the data are partitioned by taking into account factors such as important but rare observations that should be “spread out” across the various parts.

In table 5.3 we present some of the most commonly used error rates based on the confusion matrix. These error rates do not take the uncertainty in the classification into account. An observation will contribute with either a zero or a one in the expressions. But some observations may be classified based on a posterior probability close to one, others will be classified, maybe based on a posterior probability equal to 0.2 where the second largest may be 0.19. Therefore the uncertainty of the assessment of class membership may vary very much again creating a relatively large variance on the values in the confusion matrix and thus on the estimation of the error rates. This variance may be reduced by using the posterior

probabilities directly in the estimation of error rates. We present the estimates used in the SAS procedure DISCRIM in table 5.4. In this we use the definitions

$$R_j = \{\text{the observations classified as } \pi_j\}$$

$$R_{ij} = \{\text{the observations from } \pi_i \text{ classified as } \pi_j\}$$

We shall not go into details with respect to deriving those formulas, but merely state that e.g. the global error rate is equal to 1 minus the average value over all

Measure	Formula	Probabilistic interpretation
Global accuracy (ACC)	$\frac{1}{N_{..}} \sum_{i=1}^k N_{ii}$	Probability of correct classification
Global error rate, misclassification rate (MIS)	$1 - \frac{1}{N_{..}} \sum_{i=1}^k N_{ii} = \frac{1}{N_{..}} \sum_{i \neq j} N_{ij}$	Probability of misclassification
Class accuracy	$\frac{N_{ii}}{N_{i..}}$	Conditional probability of being classified as π_i given true class is π_i .
Class error rate, misclassification rate	$1 - \frac{N_{ii}}{N_{i..}} = \frac{N_{i..} - N_{ii}}{N_{i..}}$	Conditional probability of not being classified as π_i given true class is π_i

Table 5.3: Error rates estimated from the confusion matrix

Measure	Formula
Error rate for population π_i	$\hat{e}_i = 1 - \frac{1}{N_{..} p_i} \sum_{x \in R_i} k(t x)$
Stratified error rate for population π_i	$\hat{e}_i^{(s)} = 1 - \frac{1}{p_i} \sum_{j=1}^k p_j \left(\frac{1}{n_j} \sum_{x \in R_{ji}} k(t x) \right)$
Global error rate	$\hat{e} = \sum_{v=1}^k p_v \hat{e}_v = 1 - \frac{1}{N_{..}} \sum_{v=1}^k \sum_{x \in R_v} k(t x)$
Global stratified error rate	$\hat{e}^{(s)} = \sum_{v=1}^k p_v \hat{e}_v^{(s)}$

Table 5.4: Error rates estimated from the posterior probabilities.

observations in the test set of the maximal value of the posterior probabilities. If this value is very small then (most of) the maximal posterior probabilities must be close to 1 indicating a low uncertainty on the classification.

5.3.2 Terminology from non-statistical communities.

In many areas substantial parts of the terminology is based on binary classification, typically between populations that are not considered of equal importance: π_1 is called *positive* and π_2 *negative*. In statistical test theory, the choice between what should be called the hypothesis and what should be the alternative is by convention that the hypothesis H_0 corresponds to the ‘normal’ state (e.g. absence of a disease ~ negative class) and the alternative H_1 corresponds to the non-normal state (presence of the disease ~ positive class). The test theoretical formulation of the decision problem is therefore that we test the hypothesis $H_0 : \text{the class is negative}(\pi_2)$ versus the alternative $H_1 : \text{the class is positive}(\pi_1)$. Selecting the positive class will therefore correspond to rejecting the hypothesis.

Number of observations from		Classified as		Sum
		pos	neg	
Nature	pos	$tp = N_{11}$ = # true positive	$fn = N_{12}$ = # false negative	$P = N_{11} + N_{12}$ = # of pos
	neg	$fp = N_{21}$ = # false positive	$tn = N_{22}$ = # true negative	$NN = N_{21} + N_{22}$ = # of neg
Sum		$CP = N_{11} + N_{21}$ = # clas. as pos	$CN = N_{12} + N_{22}$ = # clas. as neg	$TN = N_{11} + N_{12} + N_{21} + N_{22}$ = total # classified

Table 5.5: The binary confusion matrix.

The two types of error we may commit in this situation are thus

1. Type I error: Reject a true hypothesis, i.e. conclude condition is positive (the patient has the disease) when it really is negative (the patient does not have the disease). This is called a false positive
2. Type II error: Accept a false hypothesis, i.e. conclude condition is negative (the patient is healthy) when it really is positive (the patient has the disease). This is called a false negative.

This gives a confusion matrix with a special terminology that is shown in table 5.5. The uncertainty measures derived from this are shown in table 5.6

A common procedure for generalizing these values to the case with multiclass classification is to consider averages of values for each class based on k different binary confusion matrices, each obtained by ‘collapsing’ the other $k - 1$ classes. The matrix needed to give class specific values for class π_i is given in table 5.7 and the average values for the k -class classification problem are presented in table 5.8.

5.3.3 Comparing classifiers: The ROC curve and McNemar’s test.

In order to compare classifiers it is necessary to define a relevant descriptor that summarizes much of the performance. One such way of describing a binary classifier is the **Receiver Operating Characteristic (ROC)** or **ROC curve** is a plot showing the simultaneous values (FPR, TPR) of the false positive rate and the true positive rate for different values of the discrimination threshold that determines the border between deciding that the condition is positive or negative.

Measure	Formula	Probabilistic (Bayesian) interpretation
Accuracy (ACC)	$ACC = \frac{tp+tn}{tp+fn+fp+tn} = \frac{tp+tn}{TN} = 1 - MIS$	Probability of correct classification
Error rate, misclassification rate (MIS)	$MIS = \frac{fp+fn}{tp+fn+fp+tn} = \frac{fp+fn}{TN} = 1 - ACC$	Probability of misclassification
Sensitivity, true positive rate (TPR), recall	$TPR = \frac{tp}{tp+fn} = \frac{tp}{P}$	Conditional probability of being classified positive given true class is positive.
False negative rate (FNR), miss rate	$FNR = \frac{fn}{tp+fn} = \frac{fn}{P}$	Conditional probability of being classified negative given true class is positive.
False positive rate (FPR), fall-out	$FPR = \frac{fp}{fp+tn} = \frac{fp}{NN} = 1 - SPC$	Conditional probability of being classified positive given true class is negative.
Specificity (SPC), true negative rate (TNR)	$SPC = TNR = \frac{tn}{fp+tn} = \frac{tn}{NN}$	Conditional probability of being classified negative given true class is negative.
Precision, positive predictive value (PPV)	$PPV = \frac{tp}{tp+fp} = \frac{tp}{CP} = 1 - FDR$	Posterior probability that the true class is positive given observation is classified positive.
False discovery rate (FDR)	$FDR = \frac{fp}{fp+tp} = \frac{fp}{CP} = 1 - PPV$	Posterior probability that the true class is negative given observation is classified positive.
Negative predictive value (NPV)	$NPV = \frac{tn}{fn+tn} = \frac{tn}{CN}$	Posterior probability that the true class is negative given observation is classified negative.

Table 5.6: Some uncertainty measures for the binary confusion matrix.

	Classified as		Sum
	π_i	not π_i	
Nature π_i	$tp_i = N_{ii}$ = # true positive	$fn_i = N_{..} - N_{ii}$ = # false negative	$P_i = N_{..}$ = # from π_i
	$fp_i = N_{..} - N_{ii}$ = # false positive	$tn_i = N_{..} - N_{..} - N_{..} + N_{ii}$ = # true negative	$NN_i = N_{..} - N_{..}$ = # in all classes but π_i
Sum	$CP_i = N_{..}$ = # clas. as π_i	$CN_i = N_{..} - N_{..}$ = # clas. as not π_i	$TN = N_{..}$ = total # classified

Table 5.7: The binary confusion matrix for class π_i based on the $k \times k$ confusion matrix.

Measure	Formula
Average class accuracy	$\frac{1}{k} \sum_{i=1}^k \frac{tp_i + tn_i}{TN} = \frac{\frac{2}{k} \sum_{i=1}^k N_{ii} + \frac{(k-2)}{k}}{\frac{2}{k} \sum_{i=1}^k N_{..}} = 1 - \frac{\frac{2}{k} \frac{1}{N_{..}} \sum_{i \neq j} N_{ij}}{\frac{2}{k} \sum_{i=1}^k N_{..}}$
Average class error rate, misclassification rate	$\frac{1}{k} \sum_{i=1}^k \frac{fp_i + fn_i}{TN} = \frac{\frac{2}{k} \frac{1}{N_{..}} \sum_{i=1}^k N_{ii}}{\frac{2}{k} \sum_{i=1}^k N_{..}} = \frac{2}{k} - \frac{\frac{2}{k} \frac{1}{N_{..}} \sum_{i \neq j} N_{ij}}{\frac{2}{k} \sum_{i=1}^k N_{..}}$
Average class precision	$\frac{1}{k} \sum_{i=1}^k \frac{tp_i}{tp_i + fp_i} = \frac{1}{k} \sum_{i=1}^k \frac{N_{ii}}{N_{..}}$
Average class sensitivity	$\frac{1}{k} \sum_{i=1}^k \frac{tp_i}{tp_i + fn_i} = \frac{1}{k} \sum_{i=1}^k \frac{N_{ii}}{N_{..}}$
Average class specificity	$\frac{1}{k} \sum_{i=1}^k \frac{tn_i}{fp_i + tn_i} = \frac{1}{k} \sum_{i=1}^k \frac{N_{..} - N_{..} - (N_{..} - N_{ii})}{N_{..} - N_{..}}$

Table 5.8: The average uncertainties for the k -class classification problem

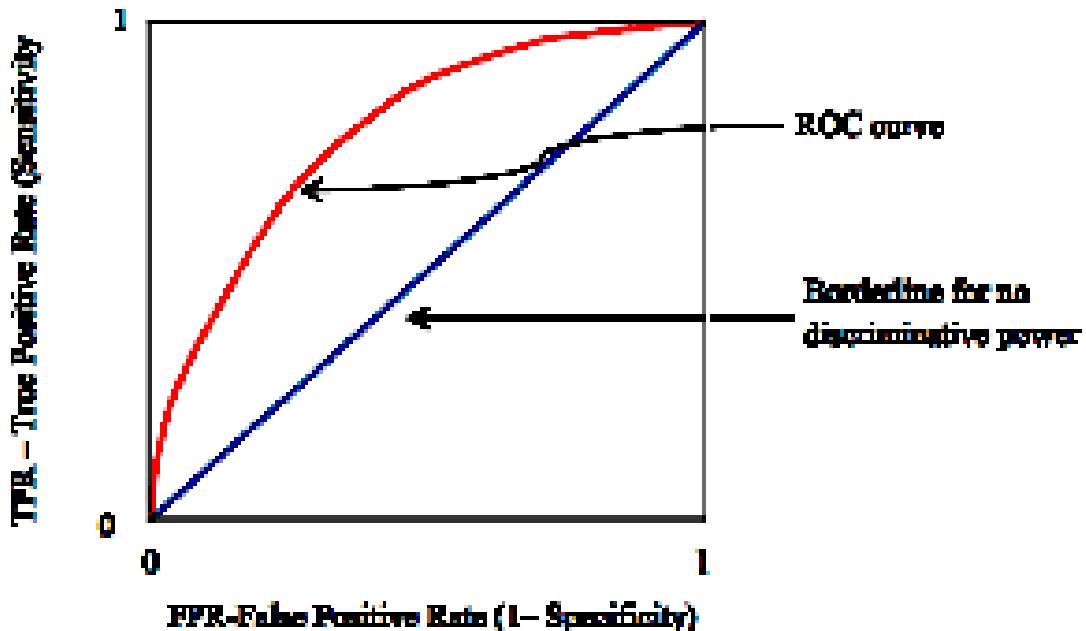


Figure 5.7: Reciever Operator Curve (ROC)

If a classifier (for a given threshold) has the same true and false positive rates it means that the conditional probability of being classified positive is independent of whether the true class is positive or negative, i.e. the classifier is equivalent to a random selection (e.g. by using a coin) of class. In general the ROC curve shows the tradeoff between FPR and TPR. The optimal ROC curve connects (0,0) to (0,1) to (1,1).

If one classifier – say A – has a ROC curve totally above the curve of another classifier – say B – then A is obviously better than B. If the curves intersect, an unambiguous answer cannot be given.

Despite this, it is customary to summarize the performance of a classifier even further into a single parameter, the **Area Under the Curve, AUC** or **AUROC**. The AUC has an interesting statistical interpretation. Let us consider a random selection of pairs of individuals, one from the positive and one from the negative group. Then the AUC is equivalent to the probability that the positive individual will be ranked before the negative. This is again closely related to the Wilcoxon rank sum test statistic. We shall not go into further details regarding this, only refer to the literature, e.g. Fawcett (2006) and Cortes et Mohri (2006). Also the AUC is often used in comparing classifiers.

If we want to compare two classifiers A and B, another approach is to consider the outcome from classifying a test set using both classifiers and the form a contingency table that summarizes how often the classifiers agree, and how often they disagree.

	B succeeded	B failed
A succeeded	N_{ss}	N_{sf}
A failed	N_{fs}	N_{ff}

Table 5.9: Contingency table summarizing successes and failures of two classifiers A and B.

We may now test whether the two classifiers perform equally well by using **McNemar's test statistic**

$$M = \frac{(|N_{fs} - N_{sf}| - 1)^2}{N_{fs} + N_{sf}}.$$

M is approximately Chi-square distributed with 1 degree of freedom if the two classifiers perform equally well, and large values are critical, i.e. we reject the hypothesis if the observed value

$$m > \chi^2(1)_{1-\alpha}$$

when using the significance level α .

5.4 Feature selection and extraction.

5.4.1 Test for further information

Given one has obtained measurements of a number of variables for some individuals with the objective of determining a discriminant function. Often the question arises if it is really necessary with all the measurements, or if one can do with fewer variables in order to separate the populations from each other.

For $i = 1, \dots, k$ we consider populations

$$\pi_i \leftrightarrow N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \text{ with prior probabilities } p_i.$$

and observations

$$X_{i1}, \dots, X_{in_i}$$

We define the within groups, between groups and total sums of squares matrices

$$W = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)^T$$

$$B = \sum_{i=1}^k n_i (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})^T$$

$$\mathbf{T} = \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}})(\mathbf{X}_{ij} - \bar{\mathbf{X}})^T.$$

A fundamental equation is

$$\mathbf{T} = \mathbf{B} + \mathbf{W}.$$

With $N = \sum n_i$ we have the following estimates of means and the dispersion matrix:

$$\begin{aligned}\hat{\boldsymbol{\mu}}_i &= \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{X}_{ij} \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{N-k} \mathbf{W}\end{aligned}$$

Without loss of generality we now want to investigate whether the last $p - q$ variables contain relevant information on population differences given that we already are using the first q . We partition the observations and parameters accordingly:

$$\mathbf{X}_{ij} = \begin{bmatrix} X_{ij,1} \\ \vdots \\ X_{ij,q} \\ X_{ij,q+1} \\ \vdots \\ X_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{ij}^{(1)} \\ \mathbf{X}_{ij}^{(2)} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_i^{(1)} \\ \boldsymbol{\mu}_i^{(2)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Thus \mathbf{X}_{ij}^1 corresponds to the first q coordinates in \mathbf{X}_{ij} and \mathbf{X}_{ij}^2 to the last $p - q$ coordinates. The index i corresponds to the population π_i where the observation is taken.

The conditional means and dispersion of \mathbf{X}_{ij}^2 given \mathbf{X}_{ij}^1 are

$$\begin{aligned}E\left(\mathbf{X}_{ij}^{(2)} \mid \mathbf{X}_{ij}^{(1)} = \mathbf{x}^{(1)}\right) &= \boldsymbol{\mu}_i^{(2)} + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}^{(1)} - \boldsymbol{\mu}_i^{(1)}) \\ &= \boldsymbol{\mu}_i^{(2)} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\mu}_i^{(1)} + \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \mathbf{x}^{(1)} \\ &= \boldsymbol{\mu}_i^{(2|1)}\end{aligned}$$

$$D\left(\mathbf{X}_{ij}^{(2)} \mid \mathbf{X}_{ij}^{(1)} = \mathbf{x}^{(1)}\right) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{2|1}$$

If all differences (for $i \neq j$) between the conditional means given the first q variables are equal to $\mathbf{0}$, i.e.

$$\boldsymbol{\mu}_i^{(2|1)} - \boldsymbol{\mu}_j^{(2|1)} = \boldsymbol{\mu}_i^{(2)} - \boldsymbol{\mu}_j^{(2)} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_i^{(1)} - \boldsymbol{\mu}_j^{(1)}) = \mathbf{0}$$

Here $\Sigma_{2|1}$ is the Schur complement Σ/Σ_{11} of Σ with respect to Σ_{11} . and therefore

$$|\Sigma| = |\Sigma_{11}| |\Sigma_{2|1}|$$

We furthermore introduce the same partitioning of \mathbf{W} , \mathbf{B} , and \mathbf{T} as we have for:

$$\begin{bmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}$$

and have

$$\mathbf{W}_{2|1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$$

$$\mathbf{T}_{2|1} = \mathbf{T}_{22} - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{12}$$

|||| Theorem 5.16

The hypothesis that the last $p - q$ variables provide no additional information to the discrimination between the populations π_1, \dots, π_k given that we are using the first q variables may be tested using the test statistic

$$\Lambda_{2|1} = \frac{|\mathbf{W}_{2|1}|}{|\mathbf{T}_{2|1}|} = \frac{|\mathbf{T}_{11}|}{|\mathbf{T}|} \times \frac{|\mathbf{W}|}{|\mathbf{W}_{11}|} = \frac{\Lambda_p}{\Lambda_q}$$

where Λ_p and Λ_q are values the test statistic Wilks' Lambda for the case with all p variables and for the case using the first q variables. Under the hypothesis the statistic follows a $U(p, k - 1, N - k - q)$ distribution.

|||| Proof omitted

See e.g. Rao (1973) or McLachlan(1992). The result on the last equalities follow directly from properties of determinants of portioned matrices. More specifically we have that $\mathbf{W}_{2|1}$ is the Schur complement $\mathbf{W}/\mathbf{W}_{11}$ of \mathbf{W} with respect to \mathbf{W}_{11} and therefore

$$|\mathbf{W}| = |\mathbf{W}_{11}| |\mathbf{W}_{2|1}|$$

and similarly for the matrix \mathbf{T} .

|||| **Theorem 5.17**

If $q = p - 1$, i.e. we investigate one variable at a time, we have

$$\frac{N - k - p + 1}{k - 1} \times \frac{1 - \Lambda_{2|1}}{\Lambda_{2|1}} \sim F(k - 1, N - k - p + 1)$$

|||| **Proof**

Follows from the general formulas on the relation between the U- and the F-distribution.

■

We now consider the case $k = 2$, i.e. $i = 1, 2$. The Mahalanobis distance between the two conditional distributions given the first q variables is

$$\Delta_{(2|1)}^2 = (\boldsymbol{\mu}_1^{(2|1)} - \boldsymbol{\mu}_2^{(2|1)})^T \boldsymbol{\Sigma}_{2|1}^{-1} (\boldsymbol{\mu}_1^{(2|1)} - \boldsymbol{\mu}_2^{(2|1)})$$

Using all variables and the first $p-q$ variables we get the unconditional Mahalanobis distances

$$\begin{aligned} \Delta^2 &= \left[(\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)})^T \quad (\boldsymbol{\mu}_1^{(2)} - \boldsymbol{\mu}_2^{(2)})^T \right] \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right]^{-1} \left[\begin{array}{c} \boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)} \\ \boldsymbol{\mu}_1^{(2)} - \boldsymbol{\mu}_2^{(2)} \end{array} \right] \\ \Delta_1^2 &= (\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)})^T \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)}). \end{aligned}$$

Since

$$\left[\begin{array}{cc} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right]^{-1} = \left[\begin{array}{cc} \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} & -\boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \\ -\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{2|1}^{-1} & \boldsymbol{\Sigma}_{2|1}^{-1} \end{array} \right]$$

we see that the conditional Mahalanobis distance is equal to the difference between the unconditional distances

$$\Delta_{(2|1)}^2 = \Delta^2 - \Delta_1^2$$

Using the estimates given in section 5.1.3, we get the corresponding empirical measures

$$D_{(2|1)}^2 = (\widehat{\boldsymbol{\mu}}_1^{(2|1)} - \widehat{\boldsymbol{\mu}}_2^{(2|1)})^T \widehat{\boldsymbol{\Sigma}}_{2|1}^{-1} (\widehat{\boldsymbol{\mu}}_1^{(2|1)} - \widehat{\boldsymbol{\mu}}_2^{(2|1)})$$

$$D^2 = (\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2)$$

$$D_1^2 = (\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)})^T \hat{\Sigma}_{11}^{-1} (\hat{\mu}_1^{(1)} - \hat{\mu}_2^{(1)})$$

and thus

$$D_{(2|1)}^2 = D^2 - D_1^2$$

In parallel to the unconditional case, a reasonable test for the hypothesis that $\Delta_{(2|1)}^2 = 0$ could be based on $D_{(2|1)}^2$. However, a simpler distribution is obtained if we use the form given in

|||| Theorem 5.18

The critical region for testing the hypothesis that the last $p - q$ variables do not contribute to the discrimination between the populations π_1 and π_2 , i.e. the hypothesis that $\Delta_{(2|1)}^2 = 0$ against all alternatives is

$$\left\{ \mathbf{x}_{11}, \dots, \mathbf{x}_{2n_2} \mid \frac{n_1+n_2-p-1}{p-q} \frac{d^2-d_1^2}{(n_1+n_2)(n_1+n_2-2)/(n_1n_2)+d_1^2} > F(p-q, n_1+n_2-p-1)_{1-\alpha} \right\}$$

Here d^2 and d_1^2 are the observed values of D^2 and D_1^2 .

|||| Proof omitted

May be found in Rao (1973).

5.4.2 Principal Component Analysis

In classification tasks one often encounters the problem that the number of features available is too large to enable a proper estimation of a classifier, but at the same time it may not be a satisfactory solution to discard some of the variables using a feature selection algorithm. In such cases it may be a solution to compute the principal components of the variables and then use so many components that a reasonable fraction of the variation in the training set is described. We want the principal components to retain the differences between the groups so we use

$$\frac{1}{N-1} \mathbf{T} = \frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}})(\mathbf{X}_{ij} - \bar{\mathbf{X}})^T$$

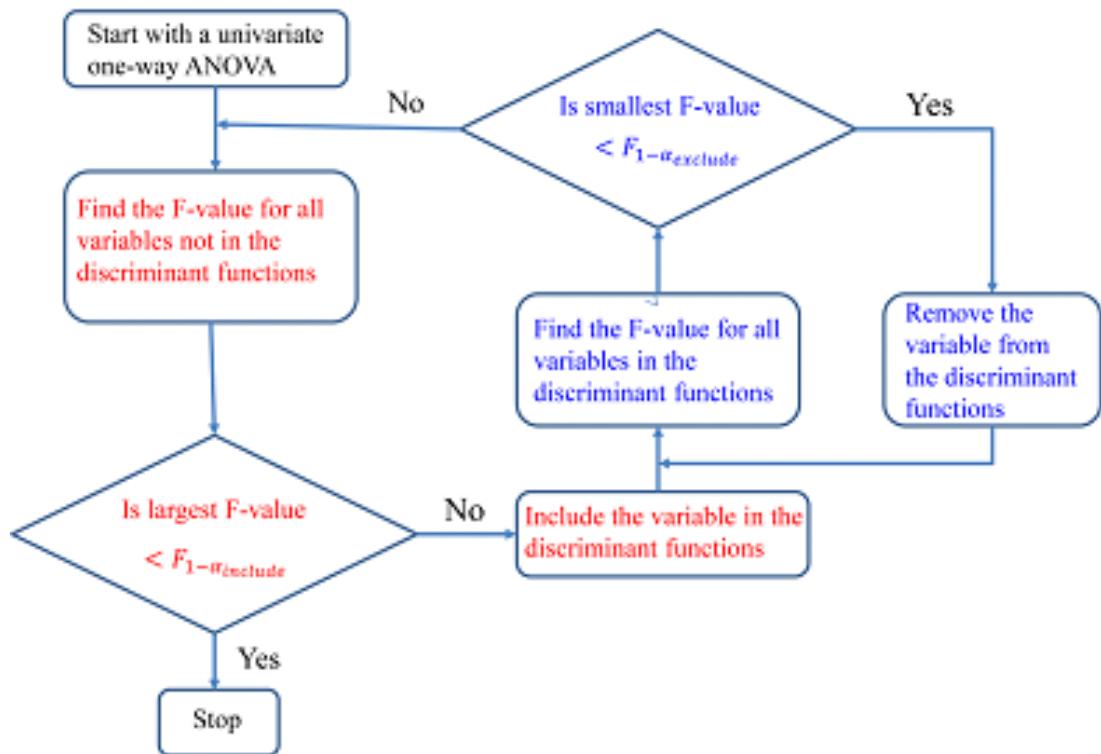


Figure 5.8: Flow diagram for a stepwise discriminant analysis procedure. The red part basically corresponds to a forward selection method and the blue part to a backward elimination method. The F-values are computed as shown in theorem 5.18

as an estimated dispersion matrix. The eigenvalues and eigenvectors of this matrix are collected in two matrices

$$\boldsymbol{\Lambda}_T = \text{diag} [\lambda_{T1} \ \cdots \ \lambda_{Tp}]$$

$$\boldsymbol{P}_T = [\boldsymbol{p}_{T1} \ \cdots \ \boldsymbol{p}_{Tp}]$$

If we retain m principal components the fraction of total variation that is described is

$$\frac{\lambda_{T1} + \cdots + \lambda_{Tm}}{\lambda_{T1} + \cdots + \lambda_{Tm} + \cdots + \lambda_{Tp}}$$

The number m may for instance be determined so that we describe say 50% of the total variation. Having determined m , we then compute the principal components as the projections on the first m eigenvectors

$$\mathbf{Y}_{ij} = \begin{bmatrix} Y_{ij1} \\ \vdots \\ Y_{ijm} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{p}_{T1})^T \\ \vdots \\ (\boldsymbol{p}_{Tm})^T \end{bmatrix} \mathbf{X}_{ij}$$

The analysis may now be performed on the \mathbf{Y}_{ij} instead of the \mathbf{X}_{ij} .

5.4.3 Canonical Discriminant Analysis

Obviously the analysis in the preceding section neglected the group structure in our training set. In this section we shall present another transformation of the data that takes this into account. This method will at the same time generalize theorem 5.6.

We still consider the situation described in the previous sections and (implicitly) assume that the hypothesis $H_0 : \boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k$ is true. We look for a (best) discriminator function that maximizes the ratio between the variation between groups and variation within groups. i.e. we seek a function $\mathbf{y} = \mathbf{d}^T \mathbf{x}$ so

$$\varphi(\mathbf{d}) = \frac{\mathbf{d}^T \mathbf{B} \mathbf{d}}{\mathbf{d}^T \mathbf{W} \mathbf{d}}$$

is maximized. We note from theorem 1.23 that the maximum value is the largest eigenvalue λ_1 and the corresponding eigenvector \mathbf{d}_1 to

$$|\mathbf{B} - \lambda \mathbf{W}| = 0$$

or

$$|\mathbf{W}^{-1} \mathbf{B} - \lambda \mathbf{I}| = 0$$

where we choose d_1 so that

$$d_1^T W d_1 = 1$$

We then seek a new discriminant function d_2 so

$$\varphi(d_2) = \frac{d_2^T B d_2}{d_2^T W d_2}$$

is maximised under the constraint that

$$d_2^T W d_1 = 0 \quad \text{and} \quad d_2^T W d_2 = 1$$

This corresponds to the second largest eigenvalue for $W^{-1}B$ and the corresponding eigenvector.

In this way one can continue until one gets an eigenvalue for $W^{-1}B$ which is 0 (or until $W^{-1}B$ is exhausted).

A plot of the values

$$\begin{bmatrix} d_r^T(x_{ij} - \bar{x}) \\ d_s^T(x_{ij} - \bar{x}) \end{bmatrix}$$

is a very useful way of visualizing the data. These plots separates the data best in the sense described above as maximizing the difference between groups with respect to the variation within groups.

Another useful plot consists of the vectors

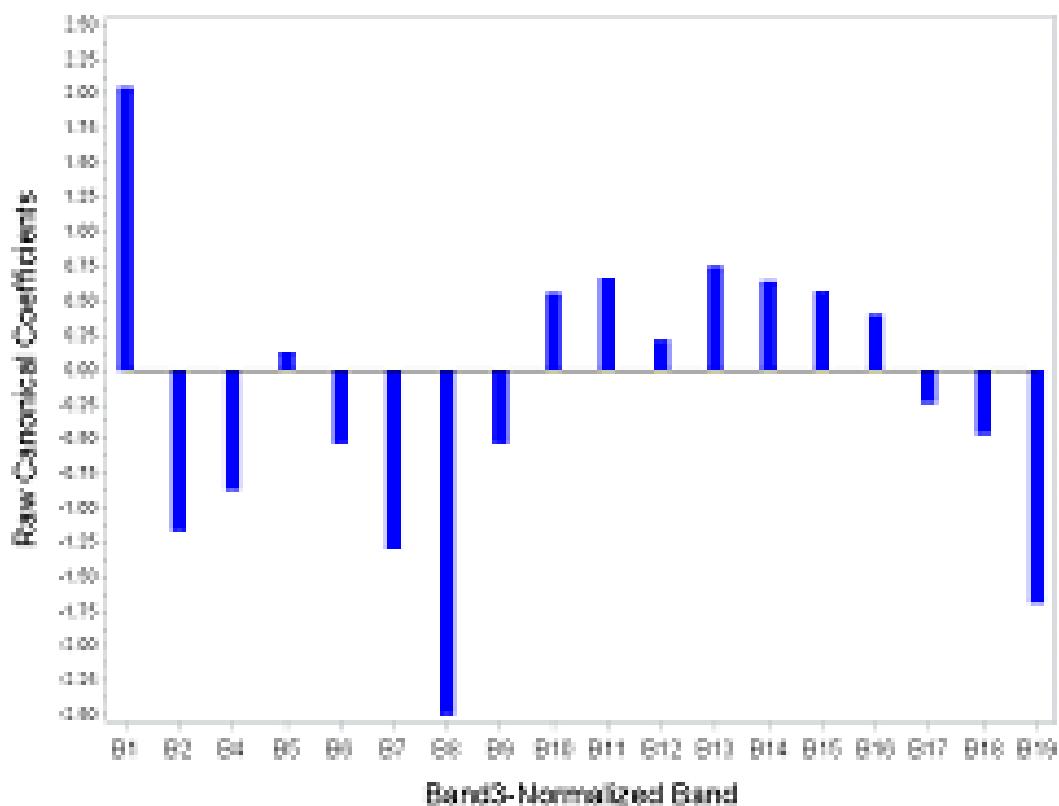
$$\begin{bmatrix} d_{11} \\ d_{21} \end{bmatrix}, \dots, \begin{bmatrix} d_{1p} \\ d_{2p} \end{bmatrix}.$$

These show with which weight the value of each single variable contributes to the plot on the $(d_1; d_2)$ -plane.

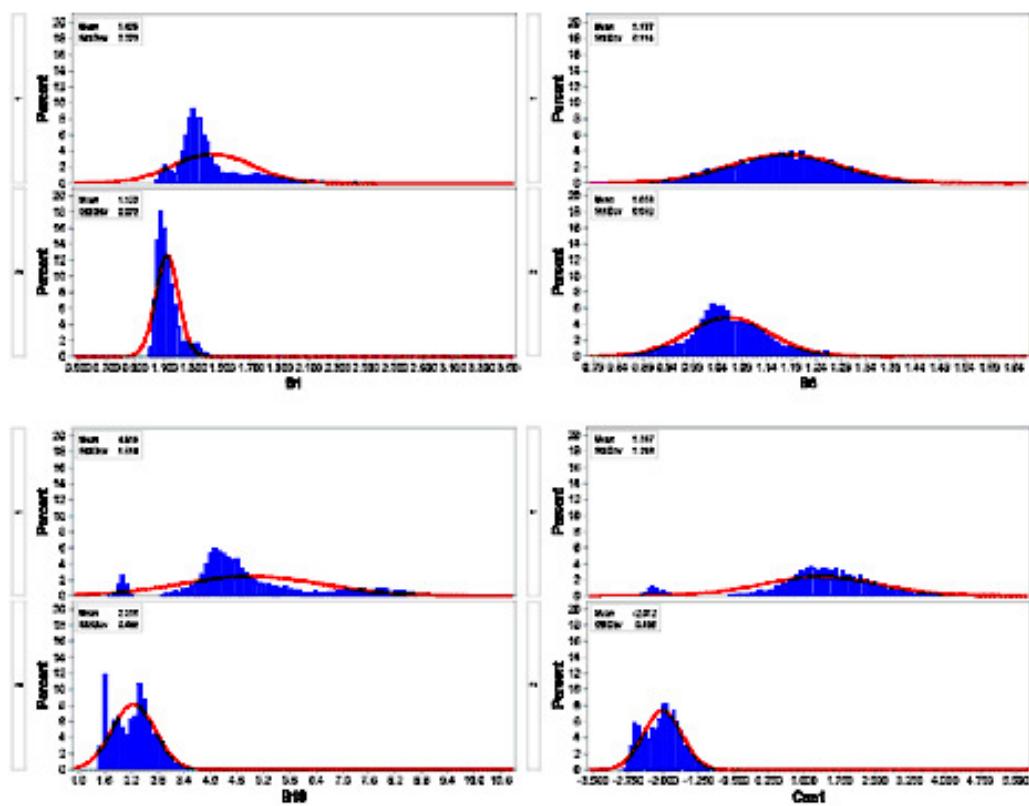
The functions $d_1^T x$ are called **Canonical Discriminant Functions (CDF)** and the type of analysis a **Canonical Discriminant Analysis (CDA)**.

||| Example 5.19

We consider the salami data. The project aimed at investigate changes in appearance under a fermentation process. During the fermentation process there is a change in color especially for the meat parts of the salami. In order to capture the variation in color between the samples we define a meat color scale based on an standard Canonical Discriminant analysis Analysis, using training areas from samples at days 2 and 42. First we however investigate how we may use a CDA in distinguishing between meat and fat. The result of the statistical analysis is shown below.

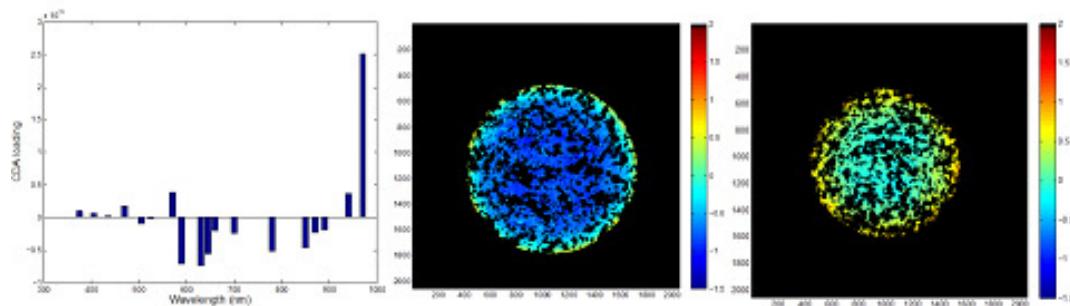


Loadings for computing the canonical discriminant variable in the salami case.



Histograms for the two salami classes fat and meat for selected bands (normalized) and for the canonical discriminant scores in the salami case.

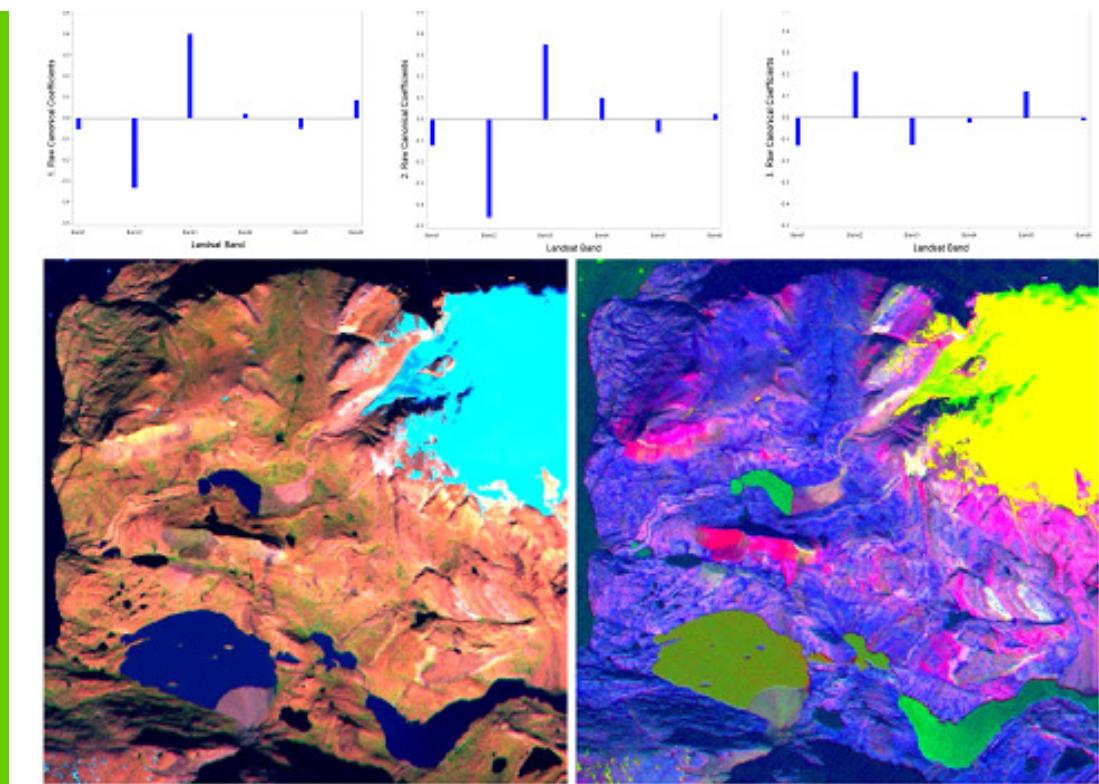
Based on the meat parts from days 2 and 42 we now make another CDA in order to distinguish between meat at those two days. In this way we obtain a statistical color scale shown in the figure below.



Leftmost graph: CDA loadings for the statistical meat color scale. Each loading refers to a spectral band. Two images to the right: CDA meat color scale. The darker blue is fresh meat, whereas yellow and orange represent darker red, fermented meat.

||| Example 5.20

Below we show the loadings for computing the three first canonical discriminant scores based on a CDA with 20 different lithological units.



Landsat false color composite and a false color composite based on the first three canonical discriminant scores used as red, green, and blue.

Furthermore the canonical discriminant scores are used to generate an image showing “maximum” difference between the geological units.

|||| Chapter 6

Principal components, canonical variables and correlations, and factor analysis

In this chapter we will give a first overview of some of the methods which can be used to show the underlying structure in a multidimensional data material.

Principal components simply correspond to the results of an eigenvalue analysis of the variance covariance matrix for a multi-dimensional stochastic variable. The method has its origin from around the turn of the century (Karl Pearson), but it was not until the thirties it got its precise formulation by Harold Hotelling.

Factor analysis was originally developed by psychologists - Spearman (1904) and Thurstone at the beginning of the previous century. Because of this the terminology has unfortunately largely been determined by the terminology of the psychologists. Around 1940 Lawley developed the maximum likelihood solutions to the problems in factor analysis - developments which later have been refined by Jöreskog and who in this period introduced factor analysis as a "statistical method".

The canonical variables and correlations also date back to Harold Hotelling. The concept resembles principal components a lot, however, we are now considering the correlation between two sets of variables instead of just transforming a single one.

6.1 Principal components

6.1.1 Definition and simple characteristics

We consider a multi-dimensional, stochastic variable

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix},$$

which has the variance-covariance (dispersion-) matrix

$$\mathbf{D}(X) = \boldsymbol{\Sigma},$$

and without loss of generality we can assume it has the mean value $\mathbf{0}$.

We will sort the eigenvalues in $\boldsymbol{\Sigma}$ descending order and will denote them

$$\lambda_1 \geq \cdots \geq \lambda_k.$$

The corresponding orthonormal eigenvectors are denoted

$$\mathbf{p}_1, \dots, \mathbf{p}_k,$$

and we define the orthogonal matrix \mathbf{P} by

$$\mathbf{P} = (\mathbf{p}_1 \cdots \mathbf{p}_k).$$

We then have the following

|||| Definition 6.1

By the i 'th principal axis of \mathbf{X} we mean the direction of the eigenvector \mathbf{p}_i corresponding to the i 'th largest eigenvalue.

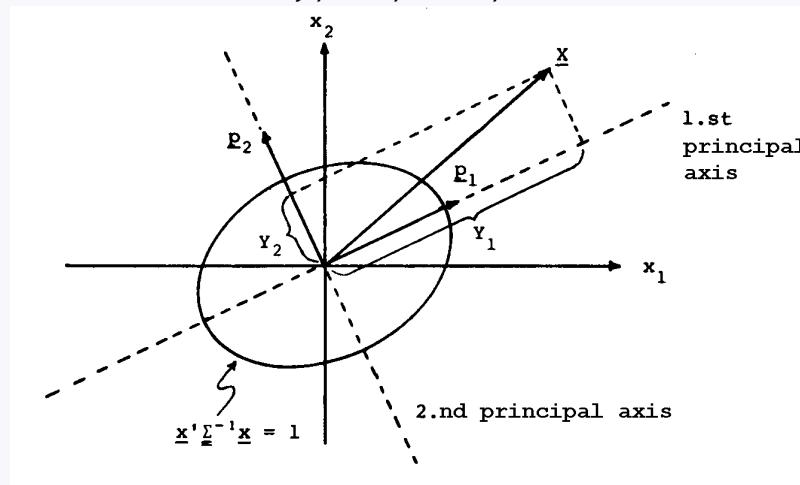
||| Definition 6.2

By the i 'th principal component of X we will understand X 's projection $Y_i = p_i'X$ on the i 'th principal axis.

The vector

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} = P'X$$

is called the *vector of principal components*.



The situation has been sketched geometrically in the figure above where we have drawn the unit ellipsoid corresponding to the variance-covariance structure i.e. the ellipsoid with the equation

$$\underline{x}'\Sigma^{-1}\underline{x} = 1.$$

It is seen that the principal axes are the main axes in this ellipsoid.

A number of theorems hold about the characteristics of the principal components. Most of these theorems are statistical reformulations of a number of the results corresponding to symmetrical positive semidefinite matrices which are given in chapter A.

||| Theorem 6.3

The principal components are uncorrelated and the variance of the i 'th component is λ_i i.e. the i 'th largest eigenvalue.

|||| **Proof**

From the theorems 1.6 (p. 6) and A.23 (p. 361) we have

$$\begin{aligned} D(\mathbf{Y}) &= D(\mathbf{P}' \mathbf{X}) = \mathbf{P}' \boldsymbol{\Sigma} \mathbf{P} = \boldsymbol{\Lambda} = \\ &\left(\begin{array}{ccc} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{array} \right), \end{aligned}$$

and the result follows readily.

■

Further we have

|||| **Theorem 6.4**

The generalised variance of the principal components is equal to the generalised variance of the original observations.

|||| **Proof**

From the definition p. 56 we have

$$GV(\mathbf{X}) = \det \boldsymbol{\Sigma}$$

and

$$GV(\mathbf{Y}) = \det \boldsymbol{\Lambda} = \lambda_1 \cdots \lambda_k,$$

■

A similar result is the following

||| Theorem 6.5

The total variance i.e. the sum of variance of the original variables is equal to the sum of the variance of the principal components i.e.

$$\sum_i V(X_i) = \sum_i V(Y_i)$$

||| Proof

Since

$$\sum V(X_i) = \text{tr } \Sigma$$

and

$$\sum V(Y_i) = \text{tr } \Lambda$$

the result follows from the note above. ■

Finally we have

||| Theorem 6.6

The first principal component is the linear combination (with normed coefficients) of the original variables which has the largest variance. The m 'th principal components is the linear combination (with normed coefficients) of the original variables which is uncorrelated with the first $m - 1$ principal components and then has the largest variance. Formally expressed:

$$\sup_{\|b\|=1} V(b'X) = \lambda_1,$$

and the supremum is given when $b = p_1$. Further we have

$$\begin{aligned} & \sup_{\substack{b \perp p_1, \dots, p_{m-1} \\ \|b\|=1}} V(b'X) = \lambda_m, \end{aligned}$$

and the supremum is given by $b = p_m$

|||| Proof

Since

$$\text{V}(\mathbf{b}' \mathbf{X}) = \mathbf{b}' \boldsymbol{\Sigma} \mathbf{b},$$

and

$$\begin{aligned}\text{Cov}(Y_i, \mathbf{b}' \mathbf{X}) &= \text{Cov}(\mathbf{p}_i' \mathbf{X}, \mathbf{b}' \mathbf{X}) = \mathbf{p}_i' \boldsymbol{\Sigma} \mathbf{b} \\ &= \lambda_i \mathbf{p}_i' \mathbf{b},\end{aligned}$$

so that

$$\text{Cov}(Y_i, \mathbf{b}' \mathbf{X}) = 0 \Leftrightarrow \mathbf{p}_i \perp \mathbf{b},$$

the theorem is just a reformulation of theorem A.30 p. 367.

■

|||| **Remark 6.7**

From the theorem we have that if we seek the linear combination of the original variables which explains most of the variation in these, then the first principal component is the solution. If we seek the m variables which explain most of the original variation, then the solution is the m first principal components. A measure of how well these describe the original variation is found by means of theorems 6.3 and 6.5 which show that the m first principal components describe the fraction

$$\frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_m + \cdots + \lambda_k}$$

of the original variation.

A better and more qualified measure of how good the “recreation ability” is, is found by trying to *reconstruct the original X from the vector*

$$Y^* = (Y_1, \dots, Y_m, 0, \dots, 0)'.$$

Since

$$Y = P'X \Leftrightarrow X = PY,$$

It is tempting to try with

$$X^* = PY^*.$$

We find

$$\begin{aligned} D(X^*) &= PD(Y^*)P' \\ &= (p_1 \cdots p_k) \begin{pmatrix} \lambda_1 & & \cdots & 0 \\ & \ddots & & \\ \vdots & & \lambda_m & \vdots \\ 0 & & \cdots & 0 \end{pmatrix} \begin{pmatrix} p'_1 \\ \vdots \\ p'_k \end{pmatrix} \\ &= \lambda_1 p_1 p'_1 + \cdots + \lambda_m p_m p'_m. \end{aligned}$$

The spectral decomposition of Σ is (p. 362)

$$\Sigma = \lambda_1 p_1 p'_1 + \cdots + \lambda_m p_m p'_m + \lambda_{m+1} p_{m+1} p'_{m+1} + \cdots + \lambda_k p_k p'_k,$$

which means that

$$\Sigma - D(X^*) = \lambda_{m+1} p_{m+1} p'_{m+1} + \cdots + \lambda_k p_k p'_k.$$

If there is a large difference between the eigenvalues then the smallest ones will be negligible and the difference between the original variance-covariance matrix and the one “reconstructed” from the first m principal components is therefore small.

6.1.2 Estimation and Testing

If the variance covariance matrix is unknown but is estimated on the basis of n observations, then one estimates the principal components and their variances simply by using the estimated variance covariance matrix as if it were known. If all the eigenvalues in Σ are different it can be shown that the eigenvalue and eigenvectors we get in this way are maximum likelihood estimates of the true parameters (see e.g. [2]).

There is, however, a very common problem here since it can be shown that the principal components are dependent of the scales of measurements our original variables have been measured in. Therefore one often chooses only to consider the normed (standardised) variables i.e.

$$Y_{\ell i} = \frac{X_{\ell i} - \bar{X}_{\ell}}{\sqrt{\sum_i (\bar{X}_{\ell i} - \bar{X}_{\ell})^2 / (n - 1)}},$$

where

$$X_i = \begin{pmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{pmatrix}, \quad i = 1, \dots, n.$$

This transformation corresponds to analysing the empirical correlation matrix instead of analysing the empirical variance covariance matrix.

If one decides to use only some of the principal components in the further analysis one could e.g. choose a strategy such as to retain as many of the components needed to account for at least e.g. 90% of the total variation.

Another criterion would be to test a hypothesis like

$$H_0 : \lambda_1 \geq \dots \geq \lambda_m \geq \lambda_{m+1} = \dots = \lambda_k$$

against the alternative that we have a distinct "greater than" ($>$) among the $k - m$ last eigenvalues.

If we are using the estimated variance covariance matrix $\hat{\Sigma}$, the test statistic becomes

$$Z_1 = -n' \ln \frac{\det \hat{\Sigma}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n' \ln \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}^{k-m}},$$

where

$$n' = n - m - \frac{1}{6}(2(k - m) + 1 + \frac{2}{k - m}),$$

and

$$\hat{\lambda} = (\text{tr } \hat{\Sigma} - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m) / (k - m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k) / (k - m).$$

The critical region using a test at level α is approximately

$$\{(x_1, \dots, x_n) | z_1 > \chi^2(\frac{1}{2}(k - m + 2)(k - m - 1))_{1-\alpha}\}.$$

If we instead are using the estimated **correlation matrix** \hat{R} we get the criterion

$$Z_2 = -n \ln \frac{\det \hat{R}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n \ln \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}^{k-m}},$$

where

$$\hat{\lambda} = (k - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m) / (k - m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k) / (k - m).$$

The critical region for a test at level α becomes approximately equal to

$$\{x_1, \dots, x_n | z_2 > \chi^2(\frac{1}{2}(k - m + 2)(k - m - 1))_{1-\alpha}\}.$$

However, it should be noted that this approximation is far worse than the corresponding approximation for the variance covariance matrix.

A discussion of the above mentioned tests can be found in [13].

We now give an example.

||| Example 6.8

The example is based on an example from [14] p. 486. The background material is measurements of seven variables on 25 boxes with randomly generated sides. The seven variables are

- X_1 : longest side
- X_2 : second longest side
- X_3 : smallest side
- X_4 : longest diagonal
- X_5 : radius in the circumscribed sphere divided by radius in the inscribed sphere
- X_6 : (longest side + second longest side)/shortest side
- X_7 : surface area/volume.

In the following table we have shown some of the observations of the seven variables.

Box	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	3.760	3.660	0.540	5.275	9.768	13.741	4.782
2	8.590	4.990	1.340	10.022	7.500	10.162	2.130
:	:	:	:	:	:	:	:
24	8.210	3.080	2.420	9.097	3.753	4.657	1.719
25	9.410	6.440	5.110	12.495	2.446	3.103	0.914

We will now consider the question: Which things about a box determine how we perceive its size?

In order to answer this question we will perform a principal component analysis of the above mentioned data. By such an analysis we hope to find out if the above mentioned 7 variables, which all in one way or another are related to "size" or "form" vary freely in the 7 dimensional space or if they are more or less concentrated in some subspaces.

We first give the empirical-variance covariance matrix for the variables. It is

$$\hat{\Sigma} = \begin{bmatrix} 5.400 & 3.260 & 0.779 & 6.391 & 2.155 & 3.035 & -1.996 \\ 3.260 & 5.846 & 1.465 & 6.083 & 1.312 & 2.877 & -2.370 \\ 0.779 & 1.465 & 2.774 & 2.204 & -3.839 & -5.167 & -1.740 \\ 6.391 & 6.083 & 2.204 & 9.107 & 1.610 & 2.782 & -3.283 \\ 2.155 & 1.312 & -3.839 & 1.610 & 10.710 & 14.770 & 2.252 \\ 3.035 & 2.877 & -5.167 & 2.782 & 14.770 & 20.780 & 2.622 \\ -1.996 & -2.370 & -1.740 & -3.283 & 2.252 & 2.622 & 2.594 \end{bmatrix}$$

Then we determine the eigenvectors and eigenvalues for $\hat{\Sigma}$. The eigenvectors are given in descending order together with the fraction and the cumulated fraction of the total variance that the eigenvalues contribute:

Eigenvalue $\hat{\lambda}_i, i = 1, \dots, 7$	Percentage of total variance	Cumulated percent- age of total variance
34.490	60.290	60.290
19.000	33.210	93.500
2.540	4.440	97.940
0.810	1.410	99.350
0.340	0.600	99.950
0.033	0.060	100.010
0.003	0.004	100.014

Computational errors in the determination of the eigenvalues lead to deviations like the cumulated sum being more than 100%.

The corresponding coordinates of the eigenvectors are shown in the following table.

Variable	\hat{p}_1	\hat{p}_2	\hat{p}_3	\hat{p}_4	\hat{p}_5	\hat{p}_6	\hat{p}_7
X_1	0.164	0.422	0.645	-0.090	0.225	0.415	-0.385
X_2	0.142	0.447	-0.713	-0.050	0.395	0.066	-0.329
X_3	-0.173	0.257	-0.130	0.629	-0.607	0.280	-0.211
X_4	0.170	0.650	0.146	0.212	0.033	-0.403	0.565
X_5	0.546	-0.135	0.105	0.165	-0.161	-0.596	-0.513
X_6	0.768	-0.133	-0.149	-0.062	-0.207	0.465	0.327
X_7	0.073	-0.313	0.065	0.719	0.596	0.107	0.092

It is seen that the first eigenvector is the direction which corresponds to more than 60% of the total variation, has especially numerically large 5th and 6th coordinates. This means that the first principal component

$$Y_1 = 0.164X_1 + \dots + 0.546X_5 + 0.768X_6 + 0.073X_7$$

is especially sensitive to variations in X_5 and X_6 . These two variables: The ratio between the radius in the circumscribed sphere and the radius in the inscribed sphere and the ratio between the sum of the two longest sides and the shortest side both have something to do with how “flat” a box is. The larger these two variables, the flatter the box. Therefore, the first principal component measures the difference in “flatness” of the boxes. The second eigenvector has large positive coordinates for the first 4 variables and a fairly large negative coordinate for the last variable. If the second principle component

$$Y_2 = 0.422X_1 + 0.447X_2 + 0.257X_3 + 0.650X_4 + \dots - 0.313X_7,$$

is large then one or more of the variables X_1, \dots, X_4 must be large while X_7 is small. Now we know that a cube is the box which for a given volume has the smallest surface. Therefore we also know that if a box deviates a lot from a cube then it will have a large X_7 - value, and this corresponds to a very strong reduction of Y_2 . A large Y_2 - value therefore indicates that most of the sides are large - and furthermore - more or less equal. We therefore conclude that Y_2 measures a more general perception of size.

In the following figure we have depicted the boxes in a coordinate system where the axes are the first two principal axes. The coordinates for a single box then become the values of the first and the second principal component for that specific box.

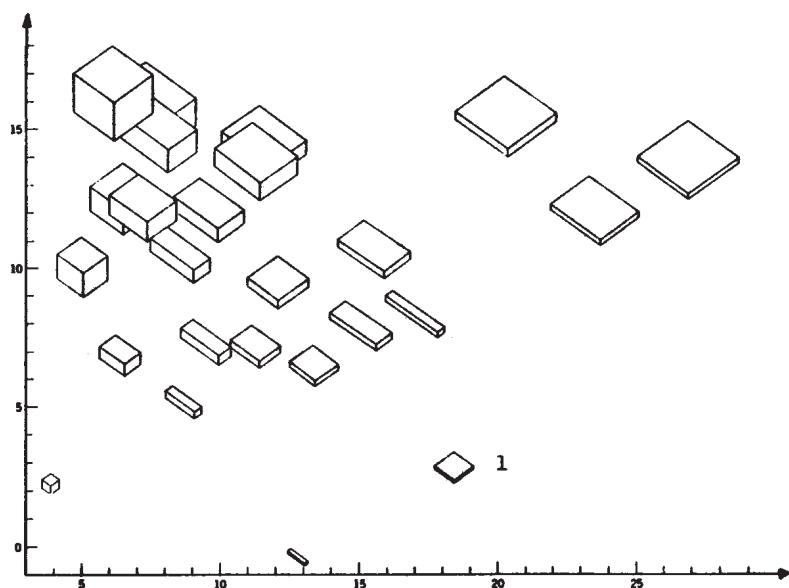


Figure 6.8.

For the first box we e.g. find

$$\begin{aligned} Y_1 &= 0.164 \cdot 3.760 + \dots + 0.073 \cdot 4.782 = 18.18 \\ Y_2 &= 0.422 \cdot 3.760 + \dots - 0.313 \cdot 4.782 = 2.15. \end{aligned}$$

At the coordinate (18.18, 2.15) we have then drawn a picture of box No. 1, etc..

From this graph we also very clearly see the interpretation we have given the principal components. To the left in the graph corresponding to small values of component No. 1 we have shown the “fattest” boxes and to the right the “flattest”. At the top of the graph corresponding to big values of component No. 2 we have the big boxes and at the bottom we have the small ones.

On the other hand we do not seem to have any precise discrimination between the oblong boxes and the more flat boxes. This discrimination is first seen when we also consider the third principal component. It is

$$Y_3 = 0.645X_1 - 0.713X_2 + \dots + 0.065X_7.$$

This component puts a large positive weight on variable No. 1 the length of the largest side and a large negative weight on the length of the second largest side. An oblong box will have $X_1 \gg X_2$ and therefore Y_3 will be relatively large for such a box. If the base of the box corresponding to the two largest sizes is close to a square then Y_3 will be close to 0 for the respective box.

The three first principal components then take care of about 98% of the total variation and by means of these we can partition a box’s “size characteristics” in three uncorrelated components: one corresponding to the flatness of the box (Y_1), one which corresponds to a more general concept of size (Y_2), and one which corresponds to “the degree of oblong-ness” (Y_3). Now the initial question of: What is

“the size of a box” should at least be partly illustrated.

The next example is based on some investigations by Agterberg et al. (see [15] p. 128).

||| Example 6.9

The Mount Albert peridotit intrusion is part of the Appalachic ultramafic belt in the Quebec province. A number of mineral samples were collected and the values of the 4 following variables were determined:

- X_1 : mol% forsterit (= Mg-olivin)
- X_2 : mol% enstatit (= Mg-ortopyroxen)
- X_3 : dimension of unit-cell of chrome-spinel
- X_4 : specific density of mineral sample.

Using between 99 and 156 observations the following correlation matrix between the variables was estimated:

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.00 & 0.32 & 0.41 & -0.31 \\ 0.32 & 1.00 & 0.68 & -0.38 \\ 0.41 & 0.68 & 1.00 & -0.36 \\ -0.31 & -0.38 & -0.36 & 1.00 \end{bmatrix}.$$

It is quite obvious that we should analyse the correlation matrix rather than the variance-covariance matrix. Because we are analysing variables which are measured in non-comparable units we must standardise the numbers.

The eigenvalues and the corresponding eigenvectors are

$$\begin{aligned} \hat{\lambda}_1 &= 2.25; \quad \hat{\mathbf{p}}_1 = \begin{bmatrix} 0.43 \\ 0.55 \\ 0.57 \\ -0.44 \end{bmatrix} \\ \hat{\lambda}_2 &= 0.74; \quad \hat{\mathbf{p}}_2 = \begin{bmatrix} -0.66 \\ 0.49 \\ 0.37 \\ 0.44 \end{bmatrix} \\ \hat{\lambda}_3 &= 0.70; \quad \hat{\mathbf{p}}_3 = \begin{bmatrix} 0.60 \\ -0.02 \\ 0.16 \\ 0.78 \end{bmatrix} \\ \hat{\lambda}_4 &= 0.31; \quad \hat{\mathbf{p}}_4 = \begin{bmatrix} -0.14 \\ -0.68 \\ 0.72 \\ -0.06 \end{bmatrix} \end{aligned}$$

All the eigenvectors have fairly large coordinates in most places so there does not seem to be any obvious possibility of giving an intuitive interpretation of the principal components.

The first principal component corresponds to $2.25/4 = 56.25\%$ of the total variation.

It would be interesting to know if the three smallest eigenvectors of the correlation matrix can be considered as being of the same magnitude.

The test statistic we will use is

$$Z = -n \ln \frac{0.74 \cdot 0.70 \cdot 0.31}{[(0.74 + 0.70 + 0.31)/3]^3} = 0.2120n,$$

where n is the number of observations on which we have based the correlation matrix on. Since this number is not the same for all the different correlation coefficients the theoretical background for the test disappears so to speak. However, if we disregard that problem, then the number of degrees of freedom in the χ^2 -distribution with which to compare the test statistic becomes

$$f = \frac{1}{2}(4 - 1 + 2)(4 - 1 - 1) = 5.$$

Since

$$\chi^2(5)_{0.995} = 16.7,$$

and since $0.21n$ for n approximately equal to 100 is quite a lot larger than this value it would be reasonable to conclude that the three smallest eigenvectors in the (true) correlation matrix are not of the same order of magnitude.

6.2 Canonical variables and correlations

Below we show two satellite images taken over the same area in India, one in March and the other in May. The data are observations from the Landsat Thematic Mapper programme - a series of satellite-borne instruments. Each observation consists of values of reflected light from six spectral bands shown in the table below. The pixel size is $30 \text{ m} \times 30 \text{ m}$.

The images are co-registered so that a given ground location corresponds to the same pixel in the two images. We may therefore organize the observations as 12-dimensional variables, i.e. we for pixel no i have:

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{Y}_i \\ \mathbf{X}_i \end{bmatrix}, \quad \mathbf{Y}_i = \begin{bmatrix} b1_{May,i} \\ \vdots \\ b6_{May,i} \end{bmatrix}, \quad \mathbf{X}_i = \begin{bmatrix} b1_{March,i} \\ \vdots \\ b6_{March,i} \end{bmatrix}$$

Spectral band	Wavelength (in μm)	Description
b1	0.45 – 0.52	visible blue
b2	0.52 – 0.60	visible green
b3	0.63 – 0.69	visible red
b4	0.76 – 0.90	near infrared
b5	1.55 – 1.75	near infrared
b6	2.08 – 2.35	near infrared

Table 6.1: Wavelengths for the spectral bands of the Landsat Thematic Mapper Earth observation satellite

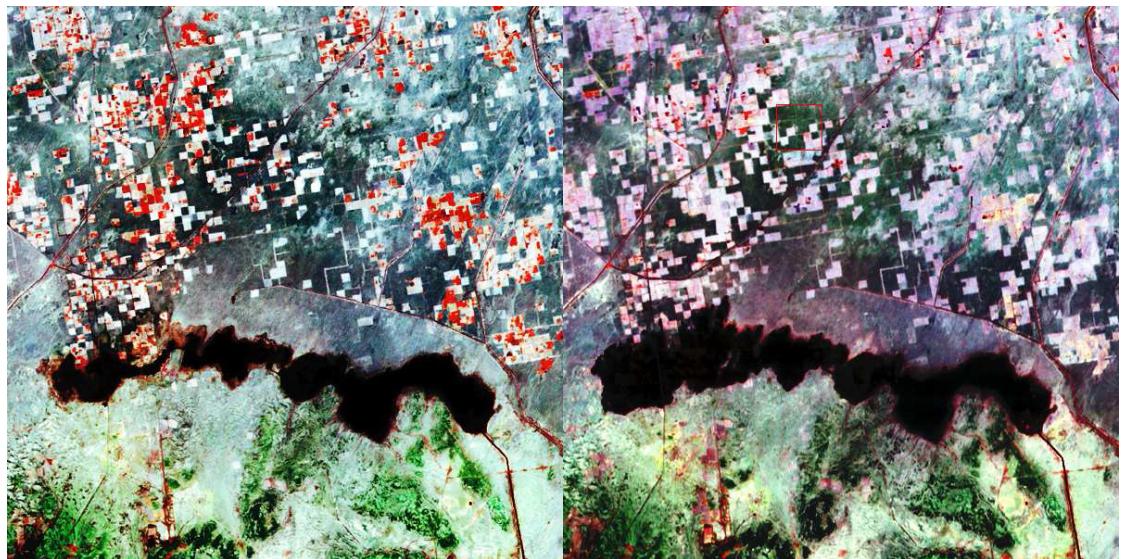


Figure 6.1: Landsat images from India observed in May and in March.

It is now of interest to investigate the relationship between the two images, which obviously amounts to comparing the \mathbf{Y} and \mathbf{X} variables. This will be done by finding suitable linear combinations $\mathbf{a}^T \mathbf{Y}$ and $\mathbf{b}^T \mathbf{X}$ of the two sets of variables determined so that the first set of linear combinations mapped as images make the two images as similar as possible. Then we shall determine two other combinations that i) are independent of the first combinations, and ii) under that constraint makes the images as similar as possible. We shall continue this operation as long as it is possible. As similarity measure we use the correlation coefficient. In the following we shall formalize the concepts indicated in the introductory remarks above.

6.2.1 Definition and properties

We consider a random variable

$$\mathbf{Z} \sim N_{p+q}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $p \leq q$ and \mathbf{Z} and the parameters have been partitioned as follows:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{X} \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}.$$

||| Definition 6.10

Consider \mathbf{Z} as above. Then the **first pair of canonical variables** is the pair of linear combinations

$$V_1 = \mathbf{a}_1^T \mathbf{Y} \text{ and } W_1 = \mathbf{b}_1^T \mathbf{X}$$

each having variance 1 that maximize the correlation $\rho(\mathbf{a}^T \mathbf{Y}, \mathbf{b}^T \mathbf{X})$ for all (\mathbf{a}, \mathbf{b}) . The maximum correlation ϱ_1 is the **first canonical correlation**. For $r \leq p$ we define the **r 'th pair of canonical variables** as the pair of linear combinations

$$V_r = \mathbf{a}_r^T \mathbf{Y} \text{ and } W_r = \mathbf{b}_r^T \mathbf{X},$$

which each has the variance 1, which are uncorrelated with the previous $r - 1$ pairs of canonical variables, and which maximizes the correlation $\rho(\mathbf{a}^T \mathbf{Y}, \mathbf{b}^T \mathbf{X})$ under those constraints. The maximum correlation ϱ_r is the **r 'th canonical correlation**.

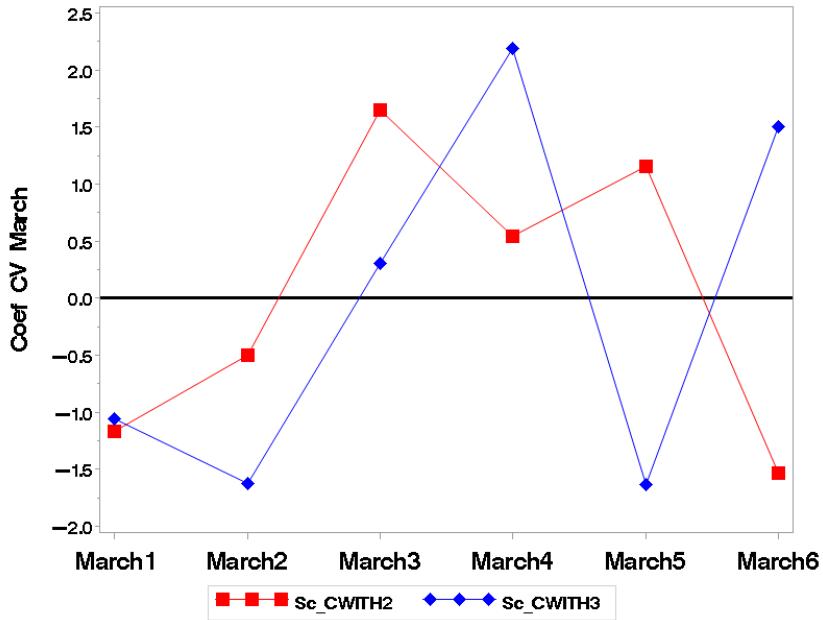


Figure 6.2: The standardized canonical coefficients for canonical variables 2 and 3 of the March variables (the WITH variables).

We have

$$\mathbf{V} = \begin{bmatrix} V_1 \\ \vdots \\ V_p \end{bmatrix} = [\mathbf{a}_1 \ \cdots \ \mathbf{a}_p]^T \mathbf{Y} = \mathbf{A}^T \mathbf{Y}$$

$$\mathbf{W} = \begin{bmatrix} W_1 \\ \vdots \\ W_q \end{bmatrix} = [\mathbf{b}_1 \ \cdots \ \mathbf{b}_p]^T \mathbf{X} = \mathbf{B}^T \mathbf{X}$$

The matrices \mathbf{A} and \mathbf{B} contain the coefficients for computing the canonical variables. If we introduce the two diagonal matrices of standard deviations $\sigma(\mathbf{Y}) = \text{diag}(\sigma(Y_1) \ \cdots \ \sigma(Y_p))$ and $\sigma(\mathbf{X}) = \text{diag}(\sigma(X_1) \ \cdots \ \sigma(X_q))$, the matrices $\mathbf{A}_{std} = \sigma(\mathbf{Y}) \mathbf{A}$ and $\mathbf{B}_{std} = \sigma(\mathbf{X}) \mathbf{B}$ contain the coefficients for computing the canonical variables based on the standardized coefficients, or shortly the **standardized canonical coefficients**.

These coefficients are important in the interpretation of the canonical variables. One should however, be aware of the fact that in cases where the correlation matrix of one (or both) set(s) of variables is (near) singular we are facing a similar problem as mentioned under multicollinearity in the section on regression analysis. In this case we can not use the coefficients, but must rely on correlations between the variables and the canonical variables. We have

$$\text{Cor} \begin{bmatrix} Y \\ X \\ V \\ W \end{bmatrix} = \begin{bmatrix} R_{yy} & R_{yx} & R_{yv} & R_{yw} \\ R_{xy} & R_{xx} & R_{xv} & R_{xw} \\ R_{vy} & R_{vx} & I_p & R_{vw} \\ R_{wy} & R_{wx} & R_{vv} & I_p \end{bmatrix},$$

where

$$R_{vw} = R_{vv} = \begin{bmatrix} \varrho_1 & & 0 \\ & \ddots & \\ 0 & & \varrho_p \end{bmatrix}$$

is diagonal with the canonical correlations in the diagonal, and

R_{yv} are the correlations between the y-variables and their canonical variables, or – in SAS terminology - the correlations between the VAR variables and their canonical variables.

R_{xv} are the correlations between the x-variables and the canonical variables of the y-variables, or the correlations between the WITH variables and the canonical variables of the VAR variables.

R_{yw} are the correlations between the y-variables and the canonical variables of the x-variables, or the correlations between the VAR variables and the canonical variables of the WITH variables.

R_{xw} are the correlations between the x-variables and their canonical variables, or the correlations between the WITH variables and their canonical variables.

In the interpretation of the canonical variables it may be useful to map these correlations in different ways as indicated in fig. 6.3

We shall now show a relation between the canonical correlations and independence between the two set of variables.

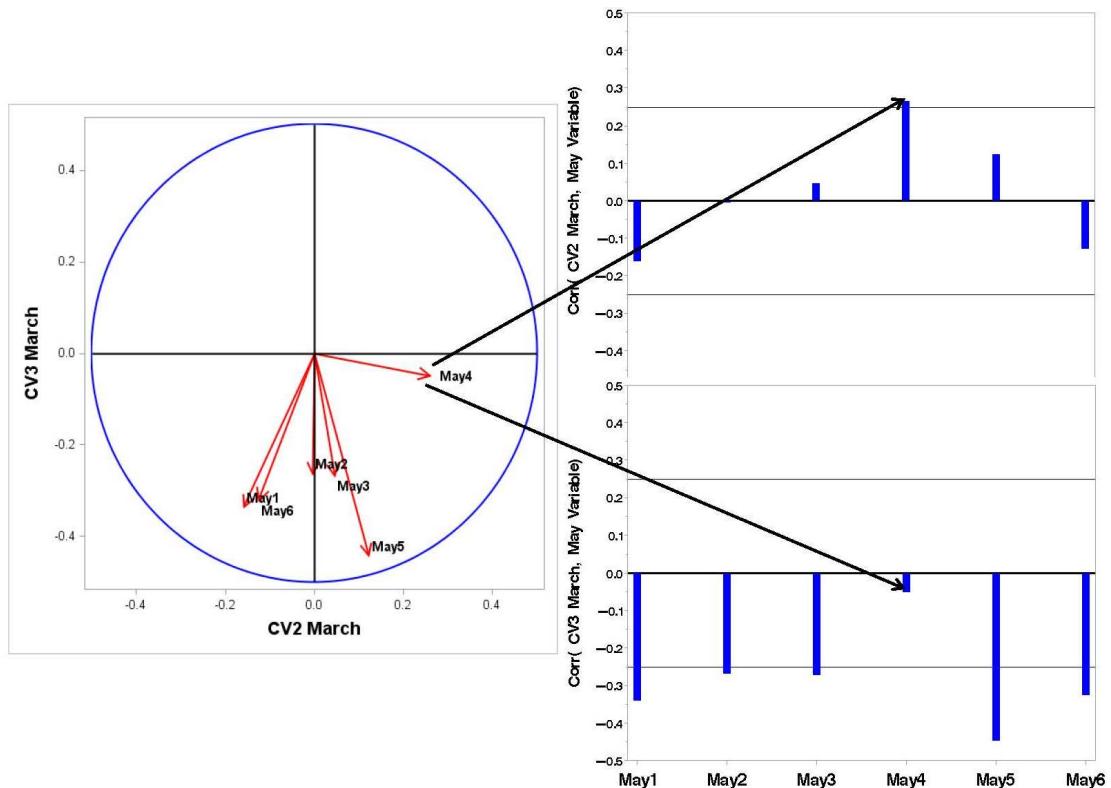


Figure 6.3: On the right hand side we show the correlations between the May values and the second and the third canonical correlation of the March variables, each mapped against the May variable name. On the left hand side we have shown the same correlations mapped as vectors, one for each May variable.

|||| **Theorem 6.11**

Let the situation be given in the above mentioned definition and let $D(\mathbf{Z}) = \Sigma$ be partitioned analogously

$$\Sigma = \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix}$$

Then the r 'th canonical correlation is equal to the r 'th largest root ρ_r of

$$\det \begin{bmatrix} -\rho \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & -\rho \Sigma_{xx} \end{bmatrix} = 0$$

and the coefficients in the r 'th pair of canonical variables satisfies

$$(i) \quad \begin{bmatrix} -\rho_r \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & -\rho_r \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \mathbf{a}_r \\ \mathbf{b}_r \end{bmatrix} = \mathbf{0}$$

$$(ii) \quad \mathbf{a}_r^T \Sigma_{yy} \mathbf{a}_r = 1$$

$$(iii) \quad \mathbf{b}_r^T \Sigma_{xx} \mathbf{b}_r = 1$$

|||| **Proof**

We have a maximization problem with constraints and one can solve the problem by using a Lagrange multiplier technique, see e.g. ?? p. 289.

■

One can also determine the correlations and the coefficients by solving an eigenvalue problem since we have

|||| Theorem 6.12

Let the situation be as in the previous theorem. Then we have

$$\begin{aligned} (\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \varrho_r^2\Sigma_{yy})\mathbf{a}_r &= \mathbf{0} \\ |(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \varrho_r^2\Sigma_{yy})| &= 0 \end{aligned}$$

respectively

$$\begin{aligned} (\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \varrho_r^2\Sigma_{xx})\mathbf{b}_r &= \mathbf{0} \\ |(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \varrho_r^2\Sigma_{xx})| &= 0 \end{aligned}$$

|||| Proof omitted

Omitted, see e.g. ??.

|||| Theorem 6.13

Let the situation be as above. Then \mathbf{Y} and \mathbf{X} are independent iff the first canonical correlation coefficient between \mathbf{Y} and \mathbf{X} is zero.

|||| Proof

We consider two one-dimensional variables V and W given by

$$V = \mathbf{a}^T \mathbf{Y} \quad \text{and} \quad W = \mathbf{b}^T \mathbf{X}$$

Then we have

$$D \begin{bmatrix} V \\ W \end{bmatrix} = \begin{bmatrix} \mathbf{a}^T \\ \mathbf{b}^T \end{bmatrix} \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \mathbf{a} & \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{a}^T \Sigma_{yy} \mathbf{a} & \mathbf{a}^T \Sigma_{yx} \mathbf{b} \\ \mathbf{b}^T \Sigma_{xy} \mathbf{a} & \mathbf{b}^T \Sigma_{xx} \mathbf{b} \end{bmatrix}.$$

Assuming that the variances (the diagonal elements) are greater than 0, the correlation between V and W is well defined and become

$$\rho(V, W) = \frac{\mathbf{a}^T \Sigma_{yx} \mathbf{b}}{\sqrt{\mathbf{a}^T \Sigma_{yy} \mathbf{a} \mathbf{b}^T \Sigma_{xx} \mathbf{b}}}.$$

We now have

$$\begin{aligned}\Sigma_{yx} = \mathbf{0} &\Leftrightarrow \forall \mathbf{a}, \mathbf{b} : \rho(\mathbf{a}^T Y, \mathbf{b}^T X) = 0 \\ &\Leftrightarrow \forall \mathbf{a}, \mathbf{b} : \rho^2(\mathbf{a}^T Y, \mathbf{b}^T X) = 0 \\ &\Leftrightarrow \max_{\mathbf{a}, \mathbf{b}} : \rho^2(\mathbf{a}^T Y, \mathbf{b}^T X) = 0,\end{aligned}$$

which concludes the proof. ■

|||| Remark 6.14

onsequently, we may test whether Y and X are independent by testing whether the first canonical correlation is 0. This may of course be done directly without the detour around the canonical correlations. If we estimate the dispersion parameters on the basis of n observations of Z , the test could be performed - as shown in section 4.4 - by investigating

$$\frac{|S|}{|S_{yy}| |S_{xx}|}$$

which is $U(p, q, n - 1 - q)$ distributed under the hypothesis of independence.

6.2.2 Estimation and testing

If the parameters are unknown they may be estimated from observations. If we insert the maximum likelihood estimates for Σ in the previous theorems we get the maximum likelihood estimates of the parameters. Most often one will probably insert the unbiased estimate S and one then gets what one can call the empirical values for the parameters involved. More specifically will we assume that we have n independent observations of Z organized in a data matrix

$$[\mathbf{Z}] = [\mathbf{Y} \quad \mathbf{X}] = \begin{bmatrix} \mathbf{Y}_1^T & \mathbf{X}_1^T \\ \vdots & \vdots \\ \mathbf{Y}_n^T & \mathbf{X}_n^T \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} & X_{11} & \cdots & X_{1q} \\ \vdots & & \vdots & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{np} & X_{n1} & \cdots & X_{nq} \end{bmatrix}$$

and we assume that the mean has been subtracted from the variables. The matrices \mathbf{Y} and \mathbf{X} should not be confused with the index less general notation for the random variables used in section 6.2.1 Then we have the unbiased estimator $\widehat{\Sigma}$ given by

$$(n - 1) \widehat{\Sigma} = [\mathbf{Y} \quad \mathbf{X}]^T [\mathbf{Y} \quad \mathbf{X}] = \begin{bmatrix} \mathbf{Y}^T \mathbf{Y} & \mathbf{Y}^T \mathbf{X} \\ \mathbf{X}^T \mathbf{Y} & \mathbf{X}^T \mathbf{X} \end{bmatrix}$$

Based on this matrix we can then obtain estimates of canonical correlations and variables by using the formulas in the preceding section.

In order to test whether the canonical correlations are 0, we set up matrices similar to what was done in the multivariate linear model for the case where the \mathbf{Y} 's are predicted by means of the \mathbf{X} 's. Thus

$$\begin{aligned} \mathbf{T} &= \mathbf{Y}^T \mathbf{Y} = (n - 1) \widehat{\Sigma}_{yy} \\ \mathbf{H} &= \mathbf{Y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} = (n - 1) \widehat{\Sigma}_{yx} \widehat{\Sigma}_{xx}^{-1} \widehat{\Sigma}_{xy} \\ \mathbf{E} &= \mathbf{T} - \mathbf{H} = (n - 1) (\widehat{\Sigma}_{yy} - \widehat{\Sigma}_{yx} \widehat{\Sigma}_{xx}^{-1} \widehat{\Sigma}_{xy}) \end{aligned}$$

We see that \mathbf{T} corresponds to the total variation and \mathbf{E} to the residual variation after having predicted \mathbf{Y} by means of \mathbf{X} .

|||| Theorem 6.15

Testing whether the canonical correlations are zero is equivalent to test whether the eigenvalues of $\mathbf{E}^{-1} \mathbf{H}$ are zero.

|||| **Proof**

The eigenvalues of $T^{-1}H$ are solutions λ_r to

$$(T^{-1}H - \lambda I)\mathbf{a} = \mathbf{0}$$

or

$$(H - \lambda T)\mathbf{b} = \mathbf{0}$$

i.e.

$$(\hat{\Sigma}_{yx}\hat{\Sigma}_{xx}^{-1}\hat{\Sigma}_{xy} - \lambda\hat{\Sigma}_{yy})\mathbf{a} = \mathbf{0}$$

The r 'th solution $\lambda_r = \varrho_r^2$ is equal to the r 'th squared canonical correlation according to Theorem 7.6. Next we find the eigenvalues of $E^{-1}H$, i.e. we must solve

$$(E^{-1}H - \gamma I)\mathbf{a} = \mathbf{0}$$

or

$$(H - \gamma(T - H))\mathbf{a} = \mathbf{0}$$

which gives

$$(T^{-1}H - \gamma(I - T^{-1}H))\mathbf{a} = \mathbf{0}$$

and

$$\left(T^{-1}H - \frac{\gamma}{1+\gamma}I\right)\mathbf{a} = \mathbf{0}.$$

Therefore

$$\frac{\gamma_r}{1+\gamma_r} = \varrho_r^2 \quad \text{and} \quad \gamma_r = \frac{\varrho_r^2}{1-\varrho_r^2}.$$

It now follows that the eigenvalues γ_r and the canonical correlations ϱ_r are zero at the same time, and the theorem follows. ■

|||| **Remark 6.16**

Here we may refer back to the tests presented in section 4.2, REMARK 4.23., and obtain the possibilities Wilks' Lambda (or Anderson's U), Pillai's Trace, Hotelling-Lawley's Trace, and Roy's maximum root.

In addition to the above tests SAS also provides output enabling a thorough analysis of how well different sets of variables are explained by other sets of variables.

||| Example 6.17 The Portland cement case

Portland cement is manufactured by mixing and grinding raw materials as limestone, clay minerals, and small amounts of other materials. This may be done in wet or dry processes. The mixture is heated in a long cylindrical kiln sloping downwards and rotating slowly. The material undergoes different processes and finally reaches a temperature of 1400°C - 1500°C and the clinker pellets are formed. Around 2-5% of gypsum is added to the clinker before it is ground to a fine powder forming cement. When water is added to the cement powder a series of not totally understood chemical reactions take place. These processes are called hydration of the cement. After a few hours, the cement starts setting and hardens over a period of weeks.

The main constituents of Portland clinker and the added gypsum in Portland cement are the cement minerals given in the table below. The ranges in weight percent for typical Danish cement are given in the last column (taken from: Portlandcementer, BETON-TEKNIK nr. 1/01/1983, rev. 1999, found at http://www aalborgportland.dk/media/pdf_filer/portlandcementer_web.pdf). The mineral names, chemical names and different forms of notation are likewise presented.

Mineral name	Chemical name	Cement notation	Oxide formula	Range (wt%)
Alite	Tricalcium silicate	C3S,	$(\text{CaO})_3\text{SiO}_2$	52-63
Belite	Dicalcium silicate	C2S,	$(\text{CaO})_2\text{SiO}_2$	11-24
Aluminate	Tricalcium aluminate	C3A,	$(\text{CaO})_3\text{Al}_2\text{O}_3$	4-8
Ferrite	Tetracalcium aluminoferrite	C4AF,	$(\text{CaO})_4\text{Al}_2\text{O}_3\text{Fe}_2\text{O}_3$	1-11
Gypsum	Calciumsulfatedihydrate	CSH ₂	$(\text{CaO})\text{SO}_3 \cdot 2\text{H}_2\text{O}$	2-5

Table 1: The major cement minerals.

In the cement minerals, we see a considerable substitution between elements. For instance, a stable compound with any composition between C2A and C2F may be formed, and C4AF is an approximation representing the midpoint in the series. It is not easy to determine the relative amounts of the cement minerals of a Portland cement. By direct chemical analysis, the relative amounts of the different elements are obtained. They are normally converted to a weight fraction in oxide form. Having determined those, a simple estimate of the amount of the different clinker minerals may be found using a simple set of equations known as Bogue's formulas. These may be found in "The Science of Concrete", <http://www.iti.northwestern.edu/cement/>, authored by Dr. Jeff Thomas and Dr. Hamlin Jennings, Assistant Research Professor and Professor respectively at Northwestern University, Evanston, IL. This reference is also a good source on cement technology.

The same source presents the oxide composition of a Portland cement, partly the ranges, partly the average values obtained from measurements at 125 laboratories for a specific cement. The rows giving values for the oxides are self-explanatory. For the remaining rows we quote: "The loss on ignition (Danish: glødetab)... is the weight lost when the cement was heated to 1000°C. At this temperature any water or CO₂ present in the cement specimen is driven off. The insoluble residue is the mass of material that is not dissolved by acid. The free CaO (often called the "free lime") is the amount of calcium oxide present as CaO, that is, not bound into the cement minerals. The free CaO is already counted in the weight fraction of CaO, so the total amount of cement mass accounted for ... is the sum of all the rows except for the last, which is 99.9%. The remaining 0.1% is likely in the form of other trace elements that were not tested for." (In the present study we use the term FRICAO for the free CaO).

Oxide		Range (wt%)	Cement #135 (wt%)
Calcium oxide	CaO	60.2 – 66.3	63.81
Silicon dioxide	SiO ₂	18.6 – 23.4	21.45
Aluminium oxide	Al ₂ O ₃	2.4 – 6.3	4.45
Ferric oxide	Fe ₂ O ₃	1.3 – 6.1	3.07
Sulfur trioxide	SO ₃	1.7 – 4.6	2.46
Magnesium oxide	MgO	0.6 – 4.8	2.42
Sodium oxide	Na ₂ O	0.05 – 1.20	0.20
Potassium oxide	K ₂ O	(Na ₂ O eq.)	0.83
Phosphorus pentoxide	P ₂ O ₅	-	0.11
Titanium oxide	TiO ₂	-	0.22
Loss on Ignition		-	0.81
Insoluble residue		-	0.16
Free CaO		-	0.64

Table 2: Oxide composition of Portland cement.

The properties of the different cement minerals will of course have a major influence on the quality of the final concrete. C3S has a rapid development of the strength, and it has a high final strength. C2S has a slow strength development, but also a high final strength. C3A has a very rapid strength development that is normally slowed down by adding gypsum to the clinkers before grinding them. Adequate amounts of added gypsum (XSO₃) has a positive influence on the early strength development. The contribution of C4AF to the strength is insignificant. Curves showing relations between strength development, hydration and fineness are presented in Example 1.33.

Other chemical parameters based on the oxide composition are the silica modulus

$$MS = \frac{SiO_2}{Al_2O_3 + Fe_2O_3},$$

the alumina modulus

$$MS = \frac{Al_2O_3}{Fe_2O_3},$$

and the lime saturation factor

$$LSF = \frac{\text{CaO} - 0.7 \times \text{SO}_3}{2.8 \times \text{SiO}_2 + 1.2 \times \text{Al}_2\text{O}_3 + 0.65 \times \text{Fe}_2\text{O}_3}.$$

These parameters are used in describing differences in the cement mineral composition. Thus as an example, a higher LSF indicates a higher proportion of C3S to C2S.

A different factor that is important for the strength development is the fineness of the cement powder because the hydration rate (obviously) will depend on this. The fineness may be determined in different ways. The Blaine number is a measure of the specific surface of the cement obtained by an air permeability test. An alternative is an assessment of the particle size distribution, e.g. characterized by a single number: the fraction of cement particles larger than a given threshold. This value may be found by a sedimentation process. In the Danish literature this measure is sometimes denoted LSR (LuftSlemmeRest)

The data from the present study are reported in Steen Tokkesdal Pedersen and Poul Skjøth: "Statistical Analysis of Data from Manufacturing". IMSOR, Technical University of Denmark, 1978, 178 pp. We are analyzing data on 198 samples of Portland cement. Means, standard deviations, and correlations are presented in Table ???. The variables STRGTH3, STRGTH7, and STRGTH28 give the strength after 3, 7, and 28 days of hardening.

In order to investigate the relation between the 3 strength measurements and the 20 chemical and physical variables we make two canonical correlation analyses: one using all 20 chemical and physical variables, and one, where we partial (condition) on the fineness measures Blaine and LSR. We present less extensive output from the latter case. From Table 5 follows that the canonical correlations are quite substantial, the largest around 0.89 implying that 80% of the variation in the first canonical strength variable is explained by the first canonical clinker variable. The corresponding correlations in the partial case are 0.657903, 0.591532, and 0.393646.

For the strength variables we get

$$\begin{bmatrix} \text{CV1} \\ \text{CV2} \\ \text{CV3} \end{bmatrix} = \begin{bmatrix} 0.8805 & 0.1341 & -0.0038 \\ -1.9852 & 2.0113 & 0.2086 \\ 0.7397 & -1.8451 & 1.5175 \end{bmatrix} \begin{bmatrix} \text{STRGTH3} \\ \text{STRGTH7} \\ \text{STRGTH28} \end{bmatrix}.$$

We see that the canonical variables largely are proportional to
the *initial strength at day 3*, to
the *increase in strength from day 3 to day 7*, and to
the *increase in strength from day 7 to day 28*.

This result also holds in the case with the partial correlations – and it gives a very nice and simple interpretation of the strength canonical variables. This is also evident from the graphs in Figure ??.

NAME	MEAN	STDDEV	Correlation with		
			STRGTH3	STRGTH7	STRGTH28
STRGTH3	263.89	36.883	1.00000	0.89557	0.60826
STRGTH7	382.29	36.705	0.89557	1.00000	0.74561
STRGTH28	515.28	32.671	0.60826	0.74561	1.00000
C3S	56.77	3.589	-0.04747	-0.00682	0.09445
C2S	21.33	3.091	-0.22076	-0.24479	-0.19806
C3A	9.63	1.131	-0.11860	-0.07761	-0.09927
C4AF	7.67	1.122	0.58759	0.56758	0.36644
SIO2	22.38	0.489	-0.57806	-0.55253	-0.25412
AL2O3	5.24	0.349	0.25156	0.28825	0.12588
FE2O3	2.52	0.369	0.58759	0.56758	0.36644
MGO	0.43	0.618	0.37851	0.37317	0.17588
CAO	65.77	0.941	-0.38782	-0.33542	-0.07893
SO3	0.73	0.280	-0.08018	-0.22494	-0.23080
TOTK2O	0.61	0.145	0.15851	0.13232	-0.04818
TOTNA2O	0.32	0.062	-0.09045	-0.05896	-0.10339
GLTAB	0.99	0.197	0.16759	0.05982	-0.14028
FRICAO	1.01	0.282	0.28937	0.28832	0.05928
XSO3	1.94	0.254	0.47398	0.37344	0.27931
MS	2.90	0.238	-0.64556	-0.64680	-0.35488
MA	2.12	0.329	-0.44018	-0.41991	-0.31494
LSF	0.93	0.014	0.14945	0.15647	0.12265
BLAINE	3095.72	234.225	0.80042	0.73643	0.51413
LSR	40.29	8.554	-0.17769	-0.07175	-0.08501

Table 3: Basic statistics and correlations between cement variables based on 198 samples of Portland cement.

	C3S	C2S	C3A	C4	Si	Al2O3	Fe2O3	MgO	CaO	SO3	TOT	K2O	TOT	GL	FRI	X	MS	MA	LSF	BLA	LSR
C3S	1.00	-.89	-.11	-.27	-.03	-.31	-.27	-.33	.64	-.19	-.19	-.01	-.07	-.27	-.04	0.30	0.09	0.85	-.08	0.02	
C2S	-.89	1.00	0.09	-.04	0.48	0.08	-.04	0.02	-.29	0.16	-.01	-.05	-.13	-.03	-.04	0.12	0.09	-.95	-.13	-.02	
C3A	-.11	0.09	1.00	-.58	-.02	0.83	-.58	-.57	0.35	0.16	-.02	-.13	-.15	-.19	-.16	-.13	0.85	0.16	-.15	0.09	
C4AF	-.27	-.04	-.58	1.00	-.60	-.03	1.00	0.86	-.78	-.30	0.21	0.05	0.23	0.50	0.38	-.72	-.91	-.19	0.52	-.13	
SiO2	-.03	0.48	-.02	-.60	1.00	-.43	-.60	-.59	0.59	-.02	-.39	-.13	-.41	-.58	-.16	0.85	0.37	-.44	-.44	0.00	
Al2O3	-.31	0.08	0.83	-.03	-.43	1.00	-.03	-.11	-.10	0.00	0.12	-.13	-.03	0.11	0.06	-.72	0.42	0.08	0.16	0.03	
FE2O3	-.27	-.04	-.58	1.00	-.60	-.03	1.00	0.86	-.78	-.30	0.21	0.05	0.23	0.50	0.38	-.72	-.91	-.19	0.52	-.13	
MgO	-.33	0.02	-.57	0.86	-.59	-.11	0.86	1.00	-.88	-.23	0.29	0.23	0.29	0.58	0.28	-.58	-.80	-.24	0.34	-.11	
CAO	0.64	-.29	0.35	-.78	0.59	-.10	-.78	-.88	1.00	-.03	-.44	-.22	-.42	-.71	-.20	0.65	0.64	0.42	-.36	0.07	
SO3	-.19	0.16	0.16	-.30	-.02	0.00	-.30	-.23	-.03	1.00	0.07	-.04	0.15	-.11	-.45	0.17	0.29	0.07	-.12	0.22	
TOTK2O	-.19	-.01	-.02	0.21	-.39	0.12	0.21	0.29	-.44	0.07	1.00	0.79	0.03	0.27	0.02	-.29	-.13	-.01	0.01	0.02	
TOTNA2O	-.01	-.05	-.13	0.05	-.13	-.13	0.05	0.23	-.22	-.04	0.79	1.00	-.10	0.06	0.05	0.01	-.09	-.00	-.16	-.01	
GLTAB	-.07	-.13	-.15	0.23	-.41	-.03	0.23	0.29	-.42	0.15	0.03	-.10	1.00	0.39	0.08	-.23	-.19	0.09	0.26	-.09	
FRICAO	-.27	-.03	-.19	0.50	-.58	0.11	0.50	0.58	-.71	-.11	0.27	0.06	0.39	1.00	0.14	-.50	-.40	-.08	0.29	-.03	
XSO3	-.04	-.04	-.16	0.38	-.16	0.06	0.38	0.28	-.20	-.45	0.02	0.05	0.08	0.14	1.00	-.30	-.33	-.09	0.49	-.52	
MS	0.30	0.12	-.13	-.72	0.85	-.72	-.72	-.58	0.65	0.17	-.29	0.01	-.23	-.50	-.30	1.00	0.39	-.07	-.51	0.06	
MA	0.09	0.09	0.85	-.91	0.37	0.42	-.91	-.80	0.64	0.29	-.13	-.09	-.19	-.40	-.33	0.39	1.00	0.18	-.40	0.12	
LSF	0.85	-.95	0.16	-.19	-.44	0.08	-.19	-.24	0.42	0.07	-.01	-.00	0.09	-.08	-.09	-.07	0.18	1.00	0.04	0.08	
BLAINE	-.08	-.13	-.15	0.52	-.44	0.16	0.52	0.34	-.36	-.12	0.01	-.16	0.26	0.29	0.49	-.51	-.40	0.04	1.00	-.20	
LSR	0.02	-.02	0.09	-.13	0.00	0.03	-.13	-.11	0.07	0.22	0.02	-.01	-.09	-.03	-.52	0.06	0.12	0.08	-.20	1.00	

Table 4: Correlations between cement variables based on 198 samples of Portland cement.

	Canonical Correlation	Squared Canonical Correlation	Eigenvalues of $\text{Inv}(E)^*H = \text{CanRsq}/(1-\text{CanRsq})$			Test of H0: The canonical correlations in the current row and all that follow are zero			
			Eigen-value	Proportion	Cumulative	Approximate F Value	Num DF	Den DF	Pr > F
1	0.892985	0.797136	3.9294	0.8302	0.8302	14.54	42	537.7	<.0001
2	0.610819	0.373100	0.5952	0.1257	0.9560	5.44	26	364	<.0001
3	0.415177	0.172372	0.2083	0.0440	1.0000	3.18	12	183	0.0004

	Standardized Canonical Coefficients for the Strength Variables			Standardized Partial Canonical Coefficients for the Strength Variables			Correlations Between the Strength Variables and their Canonical Variables		
	CV1	CV2	CV3	PCV1	PCV2	PCV3	CV1	CV2	CV3
STRGTH3	0.8805	-1.9852	0.7397	0.8408	-1.1794	0.6360	0.9983	-0.0571	0.0103
STRGTH7	0.1341	2.0113	-1.8451	0.2261	1.2588	-1.3912	0.9199	0.3889	-0.0512
STRGTH28	-0.0038	0.2086	1.5175	-0.0554	0.2959	1.2861	0.6318	0.5007	0.5917

	Standardized Canonical Coefficients for the Clinker Variables			Correlations Between the Clinker Variables and Their Canonical Variables			Correlations Between the Clinker Variables and Their Partial Canonical Variables		
	CW1	CW2	CW3	CW1	CW2	CW3	PCW1	PCW2	PCW3
C3S	-2.3964	9.3453	-7.8384	-0.0482	0.1641	0.2910	0.0525	0.1882	0.2903
C2S	-1.6198	6.7173	-6.3596	-0.2536	-0.1562	-0.0294	-0.3240	-0.1482	-0.0484
C3A	-1.0914	5.5817	1.5192	-0.1282	0.0960	-0.2292	0.0331	0.0666	-0.2303
C4AF	-2.4072	0.9962	0.9511	0.6632	0.0843	-0.1361	0.5326	0.0934	-0.1639
SIO2	0	0	0	-0.6521	-0.0272	0.4968	-0.6767	0.0430	0.5049
AL2O3	0	0	0	0.3996	0.0831	-0.6033	0.6074	0.0206	-0.5836
FE2O3	0	0	0	0.6632	0.0843	-0.1361	0.5326	0.0934	-0.1639
MGO	0.2359	-0.1130	-0.0244	0.4286	0.0586	-0.3412	0.3044	0.0512	-0.3946
CAO	1.4229	-4.0297	3.2468	-0.4325	0.1290	0.5112	-0.2866	0.1836	0.5390
SO3	0	0	0	-0.1117	-0.5648	0.0101	-0.0151	-0.6710	0.1963
TOTK2O	0.3397	-0.6100	-0.2594	0.1764	-0.0959	-0.4817	0.3991	-0.1427	-0.4457
TOTNA2O	-0.2576	0.5570	0.1387	-0.0976	0.0645	-0.2770	0.1209	0.0625	-0.3078
GLTAB	-0.0788	-0.3001	-0.2271	0.1749	-0.3956	-0.4800	-0.1334	-0.4439	-0.4952
FRICAO	0	0	0	0.3272	0.0381	-0.5453	0.1726	-0.0049	-0.5949
XSO3	0.1627	-0.7219	0.1199	0.5224	-0.2154	0.2058	0.2285	-0.1062	0.1215
MS	-1.5520	1.3196	3.3027	-0.7323	-0.1528	0.4272	-0.7331	-0.1264	0.4855
MA	-1.2126	-3.5639	-2.5231	-0.4958	-0.0596	-0.0692	-0.3368	-0.0815	-0.0705
LSF	0	0	0	0.1871	0.0094	0.0787	0.3357	-0.0193	0.1504
BLAINE	0.6095	0.3694	0.1510	0.8978	-0.0010	0.0325			
LSR	-0.0167	0.1389	-0.3012	-0.1857	0.3122	-0.3084			

Table 5: The canonical correlations, test statistics and coefficients for computing canonical variables. Correlations between canonical variables and other variables.

	Correlations Between the Strength Variables and the Canonical Variables of the Clinker Variables			Correlations Between the Strength Variables and the Partial Canonical Variables of the Clinker Variables		
	CW1	CW2	CW3	PCW1	PCW2	PCW3
STRGTH3	0.8913	-0.0349	0.0043	0.6528	-0.0618	0.0263
STRGTH7	0.8213	0.2376	-0.0212	0.5482	0.3228	-0.0353
STRGTH28	0.5641	0.3058	0.2457	0.2702	0.3797	0.2549

	Correlations Between the Clinker Variables and the Canonical Variables of the StrengthVariables			Correlations Between the Clinker Variables and the Partial Canonical Variables of the StrengthVariables		
	CV1	CV2	CV3	PCV1	PCV2	PCV3
C3S	-0.0431	0.1002	0.1208	0.0345	0.1113	0.1143
C2S	-0.2265	-0.0954	-0.0122	-0.2131	-0.0877	-0.0191
C3A	-0.1145	0.0586	-0.0952	0.0218	0.0394	-0.0907
C4AF	0.5921	0.0515	-0.0565	0.3504	0.0553	-0.0645
SIO2	-0.5821	-0.0167	0.2063	-0.4451	0.0253	0.1987
AL2O3	0.2597	0.1066	-0.1547	0.2322	0.0807	-0.1489
FE2O3	0.5921	0.0515	-0.0565	0.3504	0.0553	-0.0645
MGO	0.3827	0.0358	-0.1417	0.2002	0.0303	-0.1553
CAO	-0.3862	0.0788	0.2122	-0.1885	0.1086	0.2122
SO3	-0.0999	-0.3414	0.0055	-0.0103	-0.3930	0.0782
TOTK2O	0.1575	-0.0586	-0.2000	0.2626	-0.0844	-0.1755
TOTNA2O	-0.0872	0.0394	-0.1150	0.0795	0.0369	-0.1212
GLTAB	0.1561	-0.2416	-0.1993	-0.0878	-0.2626	-0.1949
FRICAO	0.2932	0.0178	-0.2280	0.1155	-0.0089	-0.2350
XSO3	0.4664	-0.1316	0.0854	0.1503	-0.0628	0.0478
MS	-0.6538	-0.0934	0.1774	-0.4823	-0.0748	0.1911
MA	-0.4427	-0.0364	-0.0288	-0.2216	-0.0482	-0.0277
LSF	0.1521	0.0436	0.0080	0.1975	0.0253	0.0283
BLAINE	0.8016	-0.0006	0.0135			
LSR	-0.1658	0.1907	-0.1281			

Table 6: Correlations between one set variables and the canonical variables of the opposite set of variables.

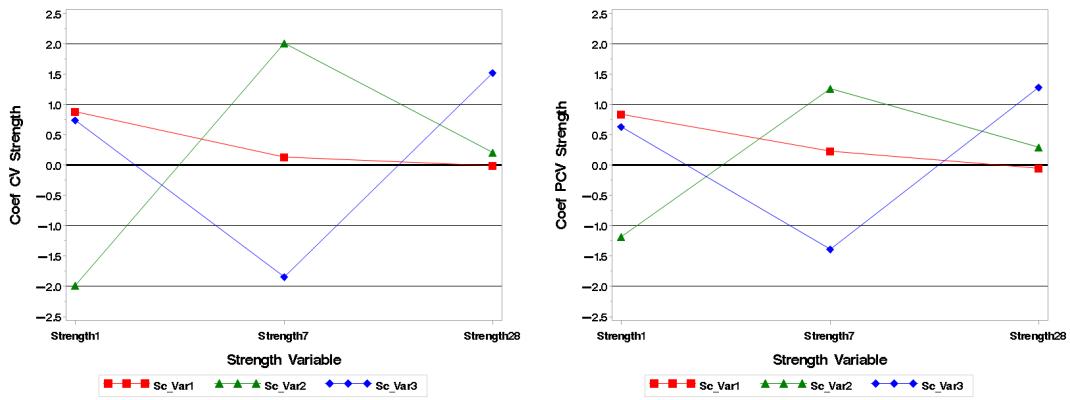


Figure 1: Standardized coefficients for computing the strength canonical variables in the full variable case and in the partial case.

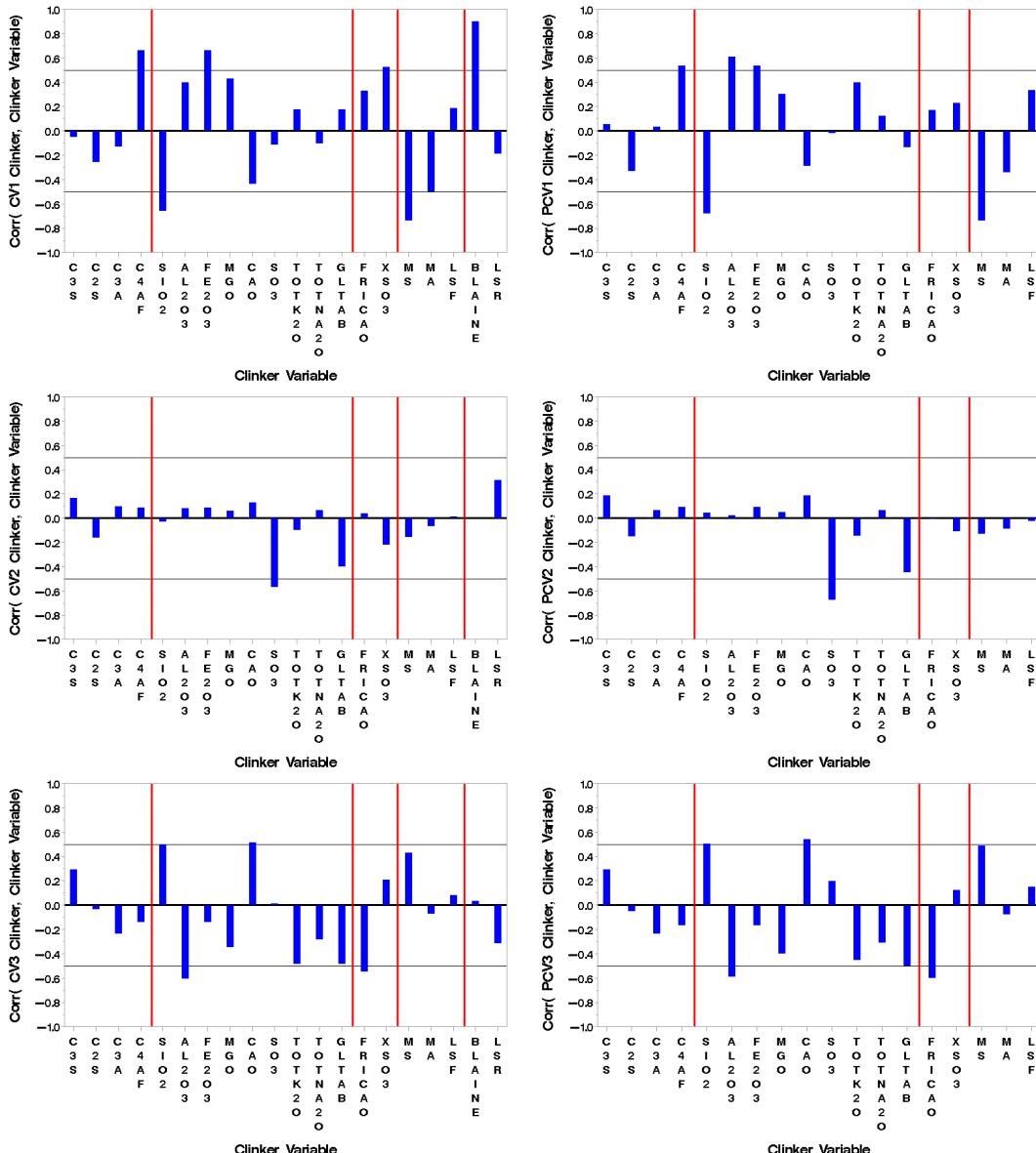


Figure 2: Correlations between the canonical variables for the clinkers and the

clinker variables. Left: Ordinary analysis, right: Analysis partialled on Blaine and LSR.

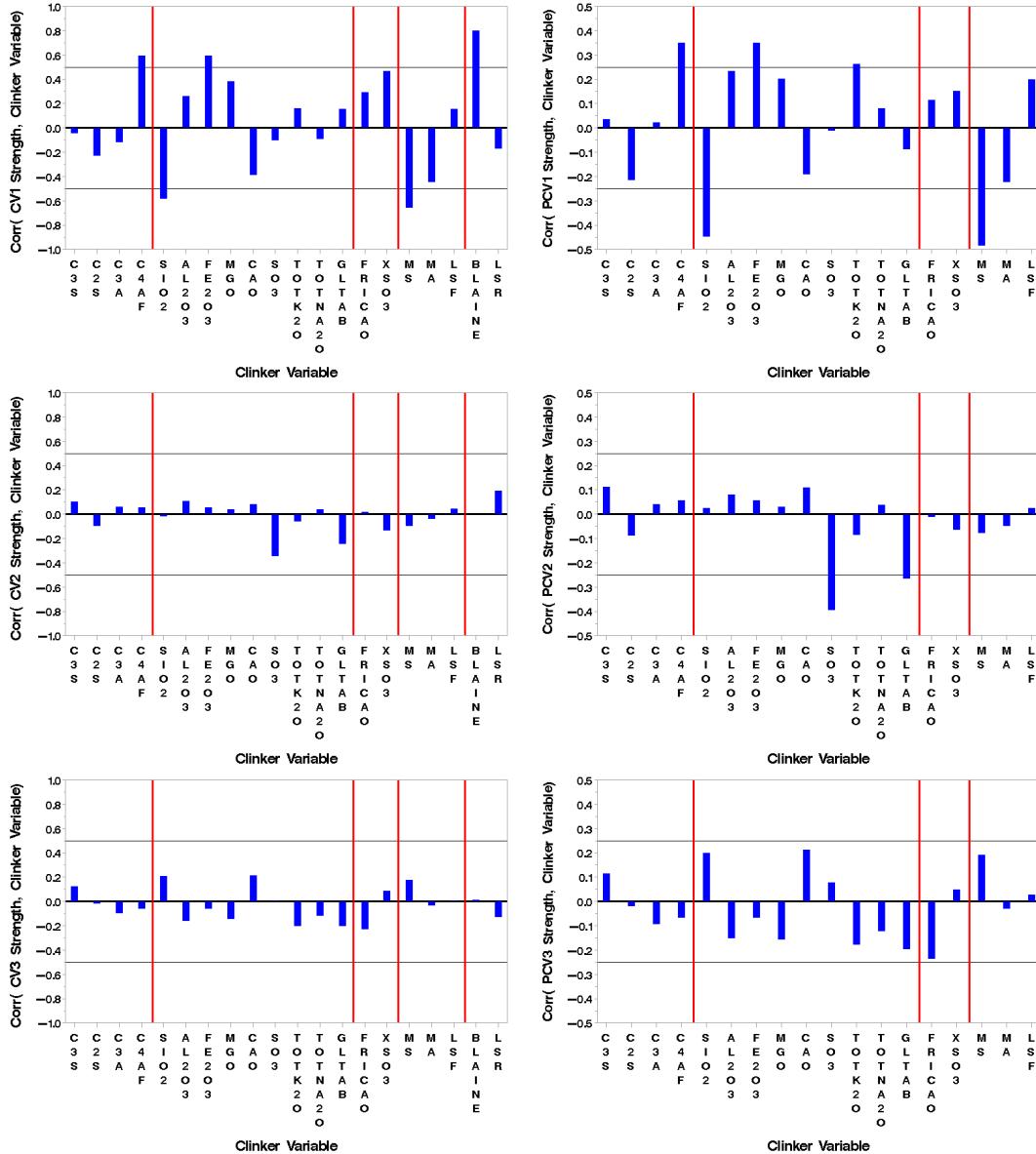


Figure 3: Correlations between the canonical variables for the strength measurements and the clinker variables. Left: Ordinary analysis, right: Analysis partialled on Blaine and LSR.

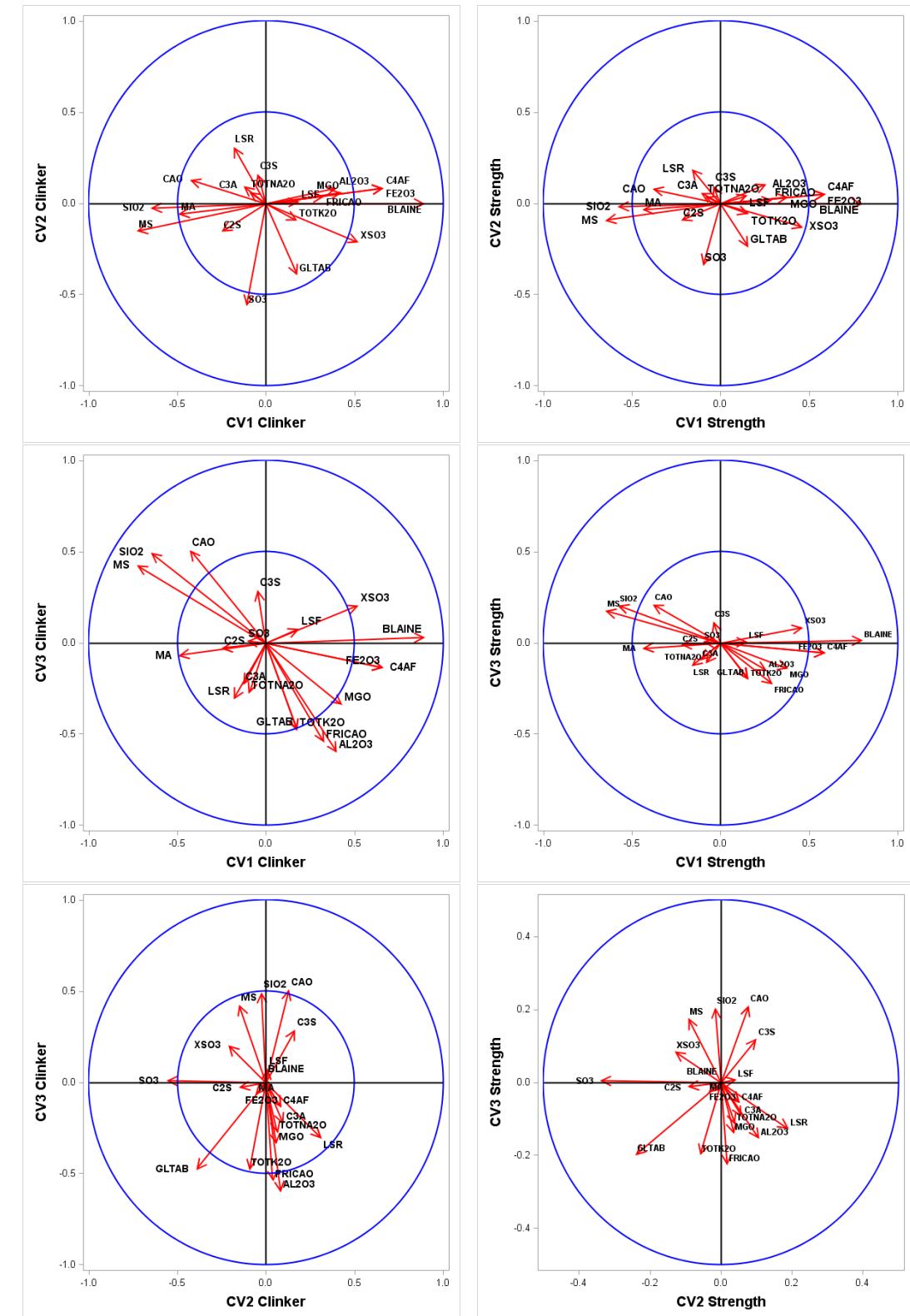


Figure 4: The correlations from Figures 2 and 3 (left part) presented as vector plots.

We now consider the clinker variables (more correctly, we should use the term cement canonical variables since we also are considering non-clinker variables as e.g.

XSO3 in the analysis). Those variables are the WITH variables in SAS terminology. Their correlation matrix does not have full rank. Therefore some canonical coefficients will be zero. If we change the order of the WITH variables, the coefficients will change. Thus the linear combinations giving the canonical clinker variables are not unique and therefore the interpretation of those canonical variables cannot be made using the standardized coefficients. (For the strength variables (the VAR variables in SAS terminology) the correlation matrix has full rank, and therefore there are no problems in using the coefficients in the interpretation.).

Instead we describe the clinker canonical variables by looking at the correlations with the clinker variables as presented in Table 5. Here we e.g. see that the correlation between SIO2 and CW1 is -0.6519 despite the fact that SIO2 does not enter the computation of CW1. Furthermore we have illustrated graphically in figures 2-4.

Groups of variables showing similar behavior are:

1. BLAINE, XSO3 C4AF/FE2O3, high values for CV1, low on others
2. MS, SIO2, CAO, large negative on CV1, positive on CV3
3. TOTK2O,FRICAO, AL2O3, GLTAB, moderate on CV1, small on CV2, large negative on CV3.

This pattern is also seen in the partial correlation analysis.

Secondly, let us consider the cross correlations between the strength canonical variables and the clinker and cement variables. The most striking feature is the dominant correlation between the first strength CV and the BLAINE value, namely 0.80, in good accordance with the fact that a very fine cement is developing strength fast due to the faster hydration of the cement. The other variables in the first group above show a similar behavior. Also the variables in the second group - MS, SIO2, CAO – display the same pattern regarding correlations with the strength CV as was the case with the clinker CV. Let us specifically look at the cement minerals. The dominant mineral C3S accounts for 58 wt% of the cement. The ‘raw’ correlations with the 3 and 7 day’s strength are negative despite the fact that C3S develops strength fast. It is seen that it has a positive correlation with the strength development from day 3 to day 7 (~ CV2 strength), and if we condition on the fineness it is also positively correlated with the development from day 7 to day 28 (~ PCV3 strength). The strength canonical variables CV1, CV2, CV3 show increasing correlations with C2S in accordance with the fact that S2C develops strength slowly and eventually becoming very strong. The same development is seen for SIO2 and the silica modulus MS.

It is generally assumed that the cement mineral C4AF is not contributing substantially to the development of strength of the cement. However, the ‘raw’ correlations with the original strength variables are considerable (0.59, 0.57, and 0.37). It still has a large correlation with the first strength canonical variable (0.59), but the correlations with the next two are small (0.05 and -0.06). If we condition on the fineness of

the cement (Blaine and LSR) also the first correlation drops considerably indicating that the higher initial value might be due to that clinkers with high C4AF content are finer (ground easier?). An alternative explanation might be that we see the effect of interaction between C4AF and other constituents. A very similar pattern is seen for the added gypsum, XSO₃.

It is beyond the scope of this exposition to go into very detailed assessments on the relation between clinker and cement chemistry and physical parameters, but the above illustrates how a careful canonical correlation analysis may give valuable contributions to the understanding of cement behavior!

6.3 Factor analysis

Once again we will consider the analysis of the correlation structure for a single multidimensional variable but contrary to the case in the section on principal components we here assume an underlying model of the structure.

6.3.1 Model and assumptions

It is assumed that we have an observation

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix},$$

which - considering the situation historically - can be thought of as a single person's scores in e.g. k different types of intelligence tests or the reactions of a person to k different stimuli.

One then has a model for how one thinks that these reactions (scores) depend on some underlying factors or more specifically that

$$\mathbf{X} = \mathbf{A} \mathbf{F} + \mathbf{G},$$

or in more detail

$$\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} G_1 \\ \vdots \\ G_k \end{bmatrix}.$$

Here we call \mathbf{F} the vector of *common factors*, they are also called *factor scores*. These are not observable. Examples of these are characteristics like three dimensional intelligence, verbal intelligence etc.

The elements of the \mathbf{A} matrix are called *factor loadings* and they give the weights of how the single factors effects the different variables. Let us e.g. assume that F_1 describes geometric intelligence and F_m verbal intelligence, and furthermore that X_1 is the result of a geometric test and X_k the result of a reading test. Then we will obviously have that a_{11} is large and that a_{1m} is small, and similarly that a_{k1} is small and a_{km} is large.

In factor analysis an important task is to interpret the unobservable factors F_j based on observations X_i and estimated values of the a_{ij} -values. Let us - e.g. - assume that we know nothing about the F 's, but still have that X_1 is the result of a geometric test and X_k the result of a reading test, and assume that a_{11} is large and that a_{1m} is small. Thus we have

$$(outcome \text{ of } geometric \text{ test}) \sim (large \ a_{11}) \times (unknown \ F_1) + \dots + (small \ a_{1m}) \times (unknown \ F_m) \quad (6-1)$$

Then we may obviously conclude that a large value of the unknown factor F_1 will give high scores in geometric tests whereas the outcome of F_m is of no importance for the geometric test score. Therefore F_1 is related to geometric tests and we may tentatively say that it describes a person's ability to solve geometric problems, i.e. the factor F_1 is describing geometric intelligence.

The vector \mathbf{G} is called the vector of *unique factors* and can be thought of as composed of some specific factors i.e. factors which are special for these specific tests and of errors i.e. non-describable deviations. Obviously these factors are not observable either.

Here we must emphasize that both \mathbf{X} and \mathbf{F} and \mathbf{G} are assumed to be stochastic. Therefore we are not considering a general linear model with the parameters F_1, \dots, F_m .

In order to make this difference quite clear we therefore give the model in the case where we have several observations X_1, \dots, X_n . We then have the n models

$$\begin{bmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \begin{bmatrix} F_{1i} \\ \vdots \\ F_{mi} \end{bmatrix} + \begin{bmatrix} G_{1i} \\ \vdots \\ G_{ki} \end{bmatrix},$$

Here we note that F_i and G_i change value when the observations X_i change value. We can aggregate the above models into

$$\begin{bmatrix} X_{11} \cdots X_{1n} \\ \vdots \\ X_{k1} \cdots X_{kn} \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \\ a_{k1} \cdots a_{km} \end{bmatrix} \begin{bmatrix} F_{11} \cdots F_{1n} \\ \vdots \\ F_{m1} \cdots F_{mn} \end{bmatrix} + \begin{bmatrix} G_{11} \cdots G_{1n} \\ \vdots \\ G_{k1} \cdots G_{kn} \end{bmatrix}.$$

It is assumed that F and G are uncorrelated and that

$$D(F) = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = I = I_m,$$

and

$$D(G) = \begin{pmatrix} \delta_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \delta_k \end{pmatrix} = \Delta.$$

Furthermore, we assume that the observations are standardised in such a way that $V(X_i) = 1, \forall i$ i.e. that the variance-covariance matrix for X is equal to its correlation matrix which is denoted

$$D(X) = R = \begin{pmatrix} 1 & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & 1 \end{pmatrix}.$$

From the original factor equation we find by means of theorem 1.6 p. 6, that

$$R = AA' + \Delta.$$

From this we especially find that for $j = 1, \dots, k$ we have

$$V(X_j) = a_{j1}^2 + \cdots + a_{jm}^2 + \delta_j = 1.$$

Here we introduce the notation

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2, \quad j = 1, \dots, k.$$

These quantities are called *communalities* and h_j^2 describes how large a proportion of X_j 's variance is due to the m common factors. Correspondingly δ_j gives the *uniqueness* in X_j 's variance. i.e. the proportion of X_j 's variance which is not due to the m common factors.

Finally the (i, j) 'th factor weight gives the correlation between the i 'th variable and the j 'th factor i.e.

$$\text{Cov}(X_i, F_j) = \text{Cov}\left(\sum_v a_{iv} F_v + G_i, F_j\right) = a_{ij}.$$

It can be shown [16] that

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2 \geq r_{j|1\dots k}^2,$$

i.e. that the j 'th communality is always larger than or equal to the square of the multiple correlation coefficient between X_j and the rest of the variables. This is not strange when remembering that this quantity exactly equals the proportion of X_j 's variance which is described by the variance in the other X_i 's.

6.3.2 Estimation of factor loadings

We now turn to the more basic problem of estimating the factors. What we are interested in determining is \mathbf{A} . We find

$$\mathbf{A} \mathbf{A}' = \mathbf{R} - \Delta.$$

The diagonal elements in this matrix are

$$1 - \delta_j = h_j^2, \quad j = 1, \dots, k.$$

We do not know these but we could estimate them e.g. by inserting the squares of the multiple correlation coefficient. If we insert these we get a matrix

$$\mathbf{V} = \begin{bmatrix} r_{1|2\dots k}^2 & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & r_{k|1\dots k-1}^2 \end{bmatrix},$$

in which the elements outside the diagonal are equal to the original correlation matrix \mathbf{R} 's elements. This matrix is still symmetric but not necessarily positive semidefinite. However, since it is still an estimate of one, we will (silently) assume that it still is positive semidefinite.

Independently of how the communalities have been estimated the resulting "correlation matrix" is called \mathbf{V} . \mathbf{V} could e.g. be the above mentioned.

We will call the eigenvalues of \mathbf{V} and the corresponding normed orthogonal eigenvectors respectively

$$\lambda_1 \geq \cdots \geq \lambda_k,$$

and

$$\mathbf{p}_1, \dots, \mathbf{p}_k.$$

If we let

$$\mathbf{P} = (\mathbf{p}_1, \dots, \mathbf{p}_k),$$

we then have from theorem A.23 p. 361, that

$$\mathbf{P}' \mathbf{V} \mathbf{P} = \Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}.$$

Since \mathbf{P} is orthogonal (as a consequence of being orthonormal) we get

$$\mathbf{V} = \mathbf{P} \Lambda \mathbf{P}' = (\mathbf{P} \Lambda^{\frac{1}{2}})(\mathbf{P} \Lambda^{\frac{1}{2}})',$$

where

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda}_k \end{pmatrix}.$$

We now define

$$\Lambda_*^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ & & \vdots \\ \vdots & & \sqrt{\lambda}_m \\ & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

i.e. $\Lambda_*^{\frac{1}{2}}$ consists of the first m columns in $\Lambda^{\frac{1}{2}}$ corresponding to the m largest eigenvalues. We then see that

$$\begin{aligned} (\mathbf{P} \Lambda_*^{\frac{1}{2}})(\mathbf{P} \Lambda_*^{\frac{1}{2}})' &= \mathbf{P} \Lambda_*^{\frac{1}{2}} \Lambda_*^{\frac{1}{2}}' \mathbf{P}' \\ &= \mathbf{P} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \lambda_m & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \mathbf{P}' \\ &\simeq \mathbf{V}, \end{aligned}$$

cf. the analogous considerations p. 276.

Since \mathbf{V} is an estimate of $\mathbf{A} \mathbf{A}'$, we then have

$$\mathbf{A} \mathbf{A}' \simeq (\mathbf{P} \Lambda_*^{\frac{1}{2}})(\mathbf{P} \Lambda_*^{\frac{1}{2}})',$$

so it would be natural to choose $\mathbf{P} \Lambda_*^{\frac{1}{2}}$ as an estimate of \mathbf{A} . This solution is called the principle factor solution for our estimation problem.

We will summarize our considerations in the following

|||| **Theorem 6.18**

We consider the factor model $\mathbf{X} = \mathbf{A}\mathbf{F} + \mathbf{G}$ where \mathbf{X} is k -dimensional and \mathbf{F} m -dimensional. The correlation matrix of \mathbf{X} is denoted \mathbf{R} , and \mathbf{V} is the matrix which we find by substituting the ones in the diagonal of \mathbf{R} with estimates of the communalities. These should be chosen in the interval $[r^2, 1]$ where r^2 is the multiple correlation coefficient between the relevant variable and the rest of the variables. Usually one chooses either r^2 or 1. The *principle factor solution* to the estimation problem is then

$$\mathbf{P}\Lambda_*^{\frac{1}{2}} = (\sqrt{\lambda_1}\mathbf{p}_1, \dots, \sqrt{\lambda_m}\mathbf{p}_m),$$

where $\lambda_i, i = 1, \dots, m$ are the m largest eigenvalues of \mathbf{V} and where $\mathbf{p}_i, i = 1, \dots, m$ are the corresponding normed eigenvectors.

|||| **Remark 6.19**

In the theorem we assume that the number of factors m is known. If this is not the case it is common to retain those which correspond to eigenvalues larger than 1. Other authors recommend that one retains one, two or three because that will usually be the upper limit to how many factors one can give a reasonable interpretation.

6.3.3 Factor rotation

Once again we consider the expression

$$\mathbf{A}\mathbf{A}' \simeq (\mathbf{P}\Lambda_*^{\frac{1}{2}})(\mathbf{P}\Lambda_*^{\frac{1}{2}})'$$

If \mathbf{Q} is an arbitrary $m \times m$ orthonormal matrix i.e. $\mathbf{Q}\mathbf{Q}' = \mathbf{I}$ then we have

$$\begin{aligned} (\mathbf{P}\Lambda_*^{\frac{1}{2}}\mathbf{Q})(\mathbf{P}\Lambda_*^{\frac{1}{2}}\mathbf{Q})' &= (\mathbf{P}\Lambda_*^{\frac{1}{2}})\mathbf{Q}\mathbf{Q}'(\mathbf{P}\Lambda_*^{\frac{1}{2}})' \\ &= (\mathbf{P}\Lambda_*^{\frac{1}{2}})(\mathbf{P}\Lambda_*^{\frac{1}{2}})' \\ &= \mathbf{A}\mathbf{A}'. \end{aligned}$$

This means that we can have as many estimates of the \mathbf{A} -matrix as we want by multiplying the principle factor solution by an orthonormal matrix.

The problem is then how to choose the \mathbf{Q} -matrix in a reasonable way. The main principle is that one wants the \mathbf{A} -matrix to become “simple” (without explaining what this means).

One of the most often used criterions is the one introduced by Kaiser, the *Varimax* criterion. It says that we must choose \mathbf{Q} in such a way that the quantity

$$\sum_j m \left\{ \sum_i \left(\frac{a_{ij}^2}{h_i^2} \right)^2 - \frac{1}{m} \left[\sum_i \left(\frac{a_{ij}^2}{h_i^2} \right) \right]^2 \right\}$$

is maximised. It is seen that the expression is the empirical variance of the terms a_{ij}^2/h_i^2 . The maximisation will therefore mean that many of the a_{ij} 's become 0 (approximately) and many become large (close to ± 1). This corresponds to a simple structure which will be easy to interpret.

Another rotation principle is the so-called *quartimax principle*. Here we try to make the rows in the factor matrix simple so that the single variables have a simple relation with the factors.

Contrary to this the Varimax criterion tries to make the columns simple corresponding to easily interpretable factors.

Before we continue with the theory we give an example.

||| Example 6.20

We will now perform a factor analysis on the data given in example 6.8.

First we determine the correlation matrix. From the estimate of the variance-covariance matrix p. 279 we find

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.000 & 0.580 & 0.201 & 0.911 & 0.283 & 0.287 & -0.533 \\ 0.580 & 1.000 & 0.364 & 0.834 & 0.166 & 0.261 & -0.609 \\ 0.201 & 0.364 & 1.000 & 0.439 & -0.704 & -0.681 & -0.649 \\ 0.911 & 0.834 & 0.439 & 1.000 & 0.163 & 0.202 & -0.676 \\ 0.283 & 0.166 & -0.704 & 0.163 & 1.000 & 0.990 & 0.427 \\ 0.287 & 0.261 & -0.681 & 0.202 & 0.990 & 1.000 & 0.357 \\ -0.533 & -0.609 & -0.649 & -0.676 & 0.427 & 0.357 & 1.000 \end{bmatrix}$$

Completely analogously with the procedure in example 6.8 we then determine the eigenvalues and vectors for $\hat{\mathbf{R}}$ (note that in this case our choice of \mathbf{V} is simply $\hat{\mathbf{R}}$). We find

Eigenvalue $\hat{\lambda}_i, 1, \dots, 7$	Percentage of total variance	Cumulated percent- age of total variance
3.3946	48.495	48.495
2.8055	40.078	88.573
0.4373	6.247	94.820
0.2779	3.971	98.791
0.0810	1.157	99.948
0.0034	0.049	99.996
0.0003	0.004	100.000

The coordinates of the corresponding eigenvectors are shown in the following table.

Variable	\hat{p}_1	\hat{p}_2	\hat{p}_3	\hat{p}_4	\hat{p}_5	\hat{p}_6	\hat{p}_7
X_1	0.405	0.293	-0.667	0.089	-0.227	0.410	-0.278
X_2	0.432	0.222	0.698	-0.034	-0.437	0.144	-0.254
X_3	0.385	-0.356	0.148	0.628	0.512	0.188	-0.108
X_4	0.494	0.232	-0.119	0.210	-0.105	-0.588	5.536
X_5	-0.128	0.575	0.209	0.111	0.389	-0.423	-0.556
X_6	-0.097	0.580	0.174	-0.006	0.355	0.500	0.498
X_7	-0.481	0.130	0.018	0.735	-0.455	0.033	0.049

We now assume that the number of factors is 2 (the assumption is not based on any deep consideration of the structure of the problem. The number 2 is chosen because there are only two eigenvalues larger than 1).

From theorem 6.18 the estimated principal factor solution to the problem is $(\sqrt{\hat{\lambda}_1} \hat{p}_1, \sqrt{\hat{\lambda}_2} \hat{p}_2)$, where

$$\begin{pmatrix} \sqrt{\hat{\lambda}_1} \hat{p}'_1 \\ \sqrt{\hat{\lambda}_2} \hat{p}'_2 \end{pmatrix} = \begin{pmatrix} 0.747 & 0.795 & 0.710 & 0.910 & -0.235 & -0.178 & -0.886 \\ 0.491 & 0.373 & -0.596 & 0.389 & 0.963 & 0.971 & 0.218 \end{pmatrix}.$$

E.g. we find

$$\hat{h}_7^2 = (-0.886)^2 + 0.218^2 = 0.833$$

The vector of estimated communalities is

$$\hat{h}^{2'} = [0.798 \ 0.771 \ 0.860 \ 0.979 \ 0.983 \ 0.976 \ 0.833],$$

and we see that e.g. the variation in variable 4 (the length of the longest diagonal) is described by the variation of the two factors by a proportion of 97.9%.

On the other hand the quantities $\delta_j = 1 - \hat{h}_j^2$ give the uniqueness value i.e. the fraction of the variance of X_j 's which is not explained by the two common factors but which is assigned to the j 'th unique factor (cf. p. 306). We find

$$\delta' = [0.202 \ 0.229 \ 0.140 \ 0.021 \ 0.017 \ 0.024 \ 0.167].$$

A more qualified measure of the ability to describe the variation in the data material of the two factors is found by recomputing the correlation matrix only from the factors.

We therefore compute the so-called residual correlation matrix

$$\hat{\mathbf{Z}} = \hat{\mathbf{R}} - \hat{\mathbf{A}}\hat{\mathbf{A}}',$$

as a more detailed measure of the factors ability to describe the original variability in the material. We get

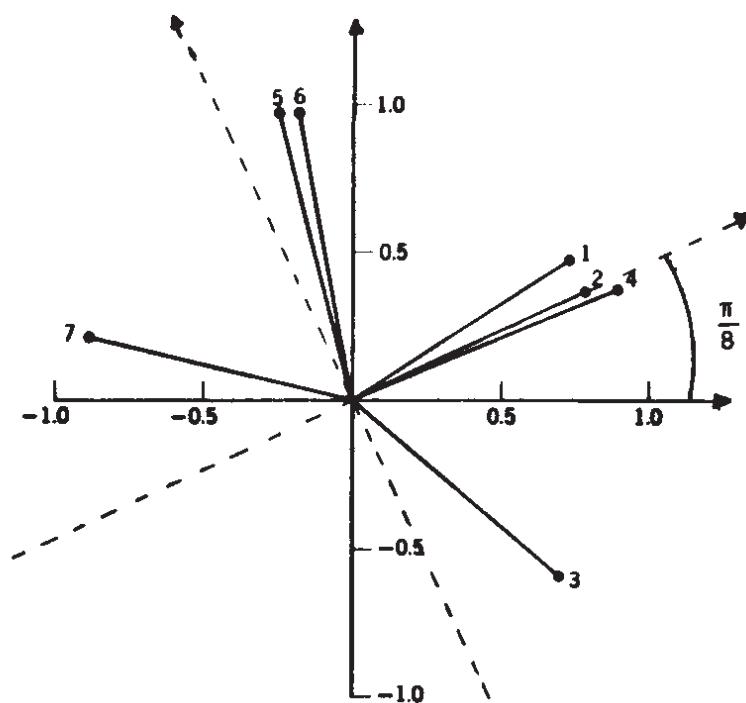
$$\hat{\mathbf{Z}} = \begin{bmatrix} 0.202 & -0.196 & -0.037 & 0.041 & -0.914 & -0.057 & 0.021 \\ -0.196 & 0.229 & 0.071 & -0.035 & -0.006 & 0.041 & 0.015 \\ -0.037 & 0.021 & 0.140 & 0.024 & 0.037 & 0.025 & 0.111 \\ 0.041 & -0.035 & 0.024 & 0.021 & 0.002 & -0.013 & 0.046 \\ -0.014 & -0.006 & 0.037 & 0.002 & 0.017 & 0.012 & 0.009 \\ -0.057 & 0.041 & 0.025 & -0.013 & 0.012 & 0.024 & -0.013 \\ 0.021 & 0.015 & 0.111 & 0.046 & 0.009 & -0.013 & 0.167 \end{bmatrix}.$$

The more $\hat{\mathbf{Z}}$ deviates from the $\mathbf{0}$ -matrix the poorer the factors describe the original material.

Apart from using the variance-covariance matrix in example 6.8 while we use the correlation matrix here, then the biggest difference in the analysis is that we have multiplied the factors by the square root of the eigenvalues corresponding to each factor. In this way the length of each factor becomes proportional to the proportion of the total variance which it explains.

We will now see if we can obtain factors which are easier to interpret by rotating the factors.

First we depict the factor weights (given on p. 312) \hat{a}_{ij} in a two-dimensional coordinate system. We find



It is noted that most of the variables have large first and second coordinates.

It seems to be possible to obtain a simple structure by rotating the coordinate system about $\frac{\pi}{8}$ ($= 22\frac{1}{2}^\circ$) anti-clockwise (dashed coordinate system).

This corresponds to multiplication by the matrix

$$\begin{pmatrix} \cos \frac{\pi}{8} & -\sin \frac{\pi}{8} \\ \sin \frac{\pi}{8} & \cos \frac{\pi}{8} \end{pmatrix} = \begin{pmatrix} 0.9239 & -0.3827 \\ 0.3827 & 0.9239 \end{pmatrix},$$

cf. section A.4.1.

The new factors or rather factor weights then become

$$\begin{bmatrix} 0.747 & 0.491 \\ 0.795 & 0.373 \\ 0.710 & -0.596 \\ 0.910 & 0.389 \\ -0.235 & 0.963 \\ -0.178 & 0.971 \\ -0.886 & 0.218 \end{bmatrix} \begin{bmatrix} 0.9239 & -0.3827 \\ 0.3827 & 0.9239 \end{bmatrix} = \begin{bmatrix} 0.878 & 0.168 \\ 0.877 & 0.040 \\ 0.428 & -0.822 \\ 0.990 & 0.011 \\ 0.151 & 0.980 \\ 0.207 & 0.965 \\ -0.735 & 0.540 \end{bmatrix}.$$

These new factor weights are simpler than the original ones in the sense that we have more values close to ± 1 and close to 0. Later we will see that this solution found visually is quite close to the Varimax-solution.

Apart from the Varimax-principle there are as mentioned a large number of

other methods for orthogonal rotation of factors which are not within the scope of this description. The interested reader is referred to the literature (e.g. [17] and [18]).

There also exists a number of rotation methods which allow relaxation of the assumption of orthogonality. These rotation methods are called "oblique rotations". The philosophy behind these is that the factors are not necessarily independent but may be correlated. Use of these methods demands thorough knowledge of the subject. We again refer to [17] and [18].

6.3.4 Computation of the factor scores

If we in the above mentioned example 6.20 wish to make a diagram analogous to the one mentioned on p. 281 then we must compute the factor scores for the single boxes. This is a bit more complicated than it was when we did the principal component analysis. Then we just had to compute the values of the principal components on the different axes. The reason that we cannot just perform the analogue operation is the existence of the specific factors.

We have the model (cf p. 305)

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \mathbf{G},$$

where

$$\begin{aligned} \mathbf{D}(\mathbf{F}) &= \mathbf{I} \\ \mathbf{D}(\mathbf{G}) &= \Delta, \end{aligned}$$

and where \mathbf{F} and \mathbf{G} are uncorrelated.

Therefore we have

$$\mathbf{D} \begin{pmatrix} \mathbf{X} \\ \mathbf{F} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\mathbf{A}' + \Delta & \mathbf{A} \\ \mathbf{A}' & \mathbf{I} \end{pmatrix}.$$

As previously mentioned, since we have that

$$\text{Cov}(X_i, F_j) = a_{ij},$$

we now have that the matrices outside the diagonal are the \mathbf{A} -matrix and its transposed respectively.

The estimate of this variance-covariance matrix is

$$\begin{bmatrix} \hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Delta} & \hat{\mathbf{A}} \\ \hat{\mathbf{A}}' & \mathbf{I} \end{bmatrix}.$$

Assuming that the underlying distributions are normal, the conditional distribution of \mathbf{F} given \mathbf{X} has the mean value

$$\boldsymbol{\mu}_F + \mathbf{A}'(\mathbf{A}\mathbf{A}' + \Delta)^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)$$

(cf. section 1.2.3).

Since our computations are performed on the standardised x -values it is reasonable to assume that $\boldsymbol{\mu}_x = \mathbf{0}$. The level for the factor scale is arbitrary but it is usually set equal to 0 so that we have the expression

$$\mathbf{A}'(\mathbf{A}\mathbf{A}' + \Delta)^{-1}\mathbf{x}$$

for the conditional mean value of F .

As an estimate of the i 'th observation of the factor score of X_i we then have

$$\hat{F}_i = \hat{\mathbf{A}}'(\hat{\mathbf{A}}\hat{\mathbf{A}}' + \hat{\Delta})^{-1}\mathbf{X}_i. \quad (6-2)$$

Now the \mathbf{A} -matrix will often have a large number of rows which means we have to invert a fairly large matrix. This can be circumvented by the following identity

$$(\mathbf{A}\mathbf{A}' + \Delta)^{-1}\mathbf{A} = \Delta^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\Delta^{-1}\mathbf{A})^{-1},$$

which gives

$$\hat{F}_i = (\mathbf{I} + \hat{\mathbf{A}}'\hat{\Delta}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\Delta}^{-1}\mathbf{X}_i. \quad (6-3)$$

The validity of the identity is found by the following relationships

$$\begin{aligned} \Leftrightarrow (\mathbf{A}\mathbf{A}' + \Delta)^{-1}\mathbf{A} &= \Delta^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\Delta^{-1}\mathbf{A})^{-1} \\ \mathbf{A} &= (\mathbf{A}\mathbf{A}' + \Delta)\Delta^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\Delta^{-1}\mathbf{A})^{-1} \\ &= \mathbf{A}(\mathbf{A}'\Delta^{-1}\mathbf{A} + \mathbf{I})(\mathbf{I} + \mathbf{A}'\Delta^{-1}\mathbf{A})^{-1}, \end{aligned}$$

and the last relationship is trivially fulfilled.

Now $\mathbf{I} + \mathbf{A}'\Delta^{-1}\mathbf{A}$ is an $m \times m$ matrix where m is the number of factors i.e. often not more than 2-3-4 so the inversion problem is not overwhelming. On the other hand as mentioned $(\mathbf{A}\mathbf{A}' + \Delta)$ is a $k \times k$ matrix where k is the number of variables i.e. often far larger than m .

If k is only of moderate size we can use the first expression for

F_i directly. Here one should utilise that

$$\mathbf{R} = \mathbf{AA}' + \Delta$$

(cf. p. 307). This gives the expression which is equivalent to (6-2)

$$\hat{\mathbf{F}}_i = \hat{\mathbf{A}}'\hat{\mathbf{R}}^{-1}\mathbf{X}_i \quad (6-4)$$

Finally we must emphasize that there are a number of other methods of determining the factor scores see e.g. [17] or [19]. It must also be noted that the problem is treated rather weakly in the main part of the literature. The main reason is probably that this problem does not have great interest for psychologists and sociologists who for many years have been the main users of factor analysis. However in a number of technical/natural science (and sociological) uses one is often interested in classifying single measurements by the size of the factor scores.

We will now illustrate the computation of factor scores on our box example.

||| Example 6.21

In example 6.20, p. 311 we found a rotated factor solution with two factors. The rotated factor weights were

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.878 & 0.168 \\ 0.877 & 0.040 \\ 0.428 & -0.828 \\ 0.990 & 0.011 \\ 0.151 & 0.980 \\ 0.207 & 0.965 \\ -0.735 & 0.540 \end{bmatrix}.$$

In order to determine the factor scores for the single boxes we must first find the communalities and the uniqueness values. We find

j	1	2	3	4	5	6	7
\hat{h}_j^2	0.7991	0.7707	0.8589	0.9802	0.9832	0.9741	0.8318
$\hat{\delta}_j$	0.2009	0.2293	0.1411	0.0198	0.0168	0.0259	0.1682
$1/\hat{\delta}_j$	4.9776	4.3611	7.0872	50.5051	59.5238	38.6100	5.9453

Here we have (cf. p. 307)

$$\hat{h}_j^2 = \hat{a}_{j1}^2 + \hat{a}_{j2}^2 = 1 - \hat{\delta}_j.$$

We note that the given communalities are equal to those we found on p. 312 for the unrotated factors. This always holds and can be used as a check in the computation of the rotated factors.

Since we have

$$\hat{\Delta} = \text{diag}(\hat{\delta}_j),$$

i.e.

$$\hat{\Delta}^{-1} = \text{diag}\left(\frac{1}{\hat{\delta}_j}\right),$$

we then have

$$(\mathbf{I} + \hat{\mathbf{A}}'\hat{\Delta}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\Delta}^{-1} =$$

$$\begin{bmatrix} 0.0669 & 0.0597 & 0.0593 & 0.7839 & 0.0244 & 0.0510 & -0.0750 \\ -0.0002 & -0.0059 & 0.0655 & -0.0943 & 0.5770 & 0.3641 & 0.0415 \end{bmatrix}$$

Equation (6-3) assumes that the variables X are standardised. We must therefore first determine the mean value and the standard deviation for each of the 7 variables. These are

j	1	2	3	4	5	6	7
\bar{X}_j	7.1000	4.7730	2.3488	9.1338	5.4582	7.1674	2.3462
s_j	2.3238	2.4178	1.6656	3.0178	3.2733	4.5581	1.6105

The standardised values for e.g. the first box becomes

$$\mathbf{z} = (-1.4373 \quad -0.4603 \quad -1.0860 \quad -1.2787 \quad 1.3167 \quad 1.4422 \quad 1.5124)',$$

where e.g. the second value is found as

$$z_2 = \frac{3.660 - 4.773}{2.4178} = -0.4603.$$

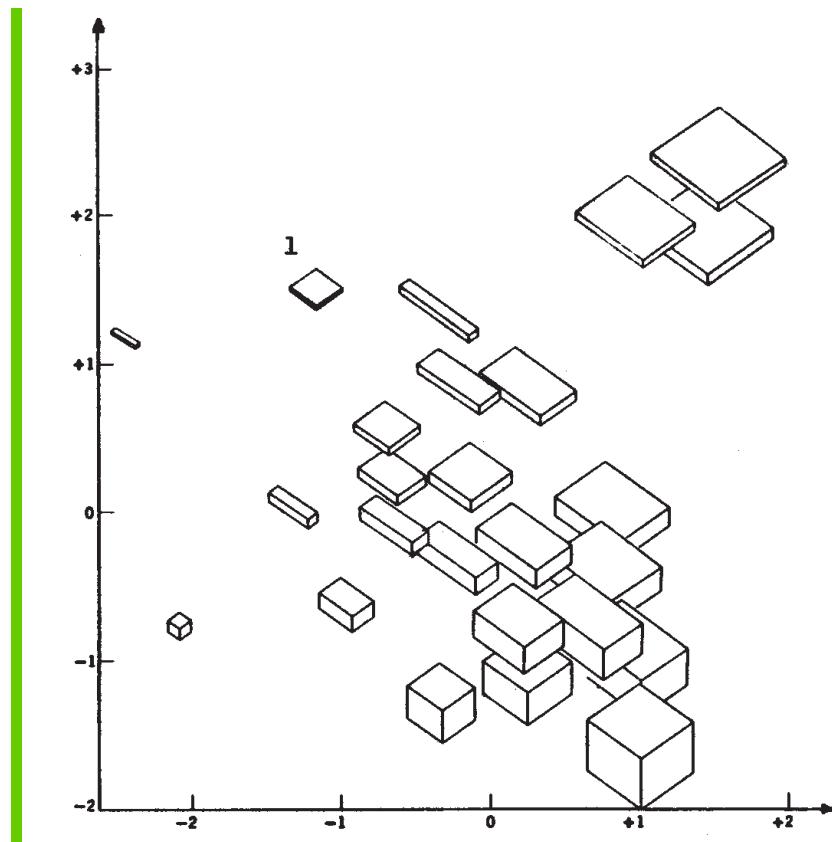
We now easily find the factor scores corresponding to the first box as

$$\hat{\mathbf{f}}_1 = (\mathbf{I} + \hat{\mathbf{A}}'\hat{\Delta}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\Delta}^{-1}\mathbf{z} = \begin{pmatrix} -1.20 \\ 1.40 \end{pmatrix}.$$

The others are found analogously.

In the following figure we have shown the 25 boxes in a 2-dimensional coordinate system so that each box is placed at the coordinates corresponding to its factor scores (cf. p. 281).

We note (cf. example 6.20) that the two factors describe "thickness" and "size". However, we also note that the "importance" of the two concepts has been switched compared to example 6.8.



6.3.5 Briefly on maximum likelihood factor analysis

From a statistical point of view the maximum likelihood method is somewhat more satisfactory than e.g. the principal factor method. Furthermore, the maximum likelihood solution has a scale-invariance property which is very satisfactory.

We will not concern ourselves with the important numerical and technical problems in determining the maximum likelihood solution but more consider the scale-invariance.

We denote the empirical variance-covariance matrix \mathbf{S} and if we assume normality of the observations we have that \mathbf{S} is Wishart distributed with the parameters $(n - 1, \frac{1}{n-1}\Sigma)$ where Σ equals $D(X_i)$. Thus the density is

$$c_1(\det \mathbf{S})^{\frac{1}{2}(n-k-2)} (\det \Sigma)^{-\frac{1}{2}(n-1)} \exp(-\frac{1}{2}(n-1) \text{tr}(\mathbf{S} \Sigma^{-1})),$$

where c_1 is an integration constant which only depends on n and k . The logarithm of the likelihood function is therefore (disregarding the terms which do

not depend on Σ):

$$\ln L(\Sigma) = -\frac{1}{2}(n-1)\ln(\det \Sigma) - \frac{1}{2}(n-1)\text{tr}(\mathbf{S}\Sigma^{-1}).$$

Here we now introduce the usual m factor model

$$\mathbf{D}(\mathbf{X}) = \Sigma = \mathbf{A}\mathbf{A}' + \Delta,$$

where \mathbf{A} and Δ are as in section 6.3.4. Note that we are not assuming that Σ has ones on the diagonal. This gives

$$\ln L(\mathbf{A}, \Delta) = -\frac{1}{2}(n-1)\ln(\det(\mathbf{A}\mathbf{A}' + \Delta)) - \frac{1}{2}(n-1)\text{tr}(\mathbf{S}(\mathbf{A}\mathbf{A}' + \Delta)^{-1}).$$

Maximisation of this function with respect to \mathbf{A} and Δ gives the ML-solution to our factor analysis. Concerning the technical problems which remain, we refer to [20].

By partial differentiation of the logarithm of the likelihood function, and after long and tedious algebraic manipulations, one obtains the equation:

$$\hat{\mathbf{A}} = (\hat{\Delta} + \hat{\mathbf{A}}\hat{\mathbf{A}}')\mathbf{S}^{-1}\hat{\mathbf{A}}, \quad (6-5)$$

see e.g. [19].

If we perform a scale-transformation of the X 's i.e. we introduce

$$\mathbf{Z}_i = \mathbf{C} \mathbf{X}_i,$$

we then have

$$\mathbf{S}_z = \mathbf{C} \mathbf{S}_x \mathbf{C}'$$

where z and x as subscripts shows whether the different quantities have been computed on the base of the \mathbf{Z}_i 's or the \mathbf{X}_i 's. With the same convention of notation we then have

$$\hat{\mathbf{A}}_z = (\hat{\Delta}_z + \hat{\mathbf{A}}_z\hat{\mathbf{A}}'_z)\mathbf{C}'^{-1}\mathbf{S}_x^{-1}\mathbf{C}^{-1}\hat{\mathbf{A}}_z.$$

If we pre-multiply by \mathbf{C}^{-1} we get

$$\mathbf{C}^{-1}\hat{\mathbf{A}}_z = [\mathbf{C}^{-1}\hat{\Delta}_z\mathbf{C}'^{-1} + \mathbf{C}^{-1}\hat{\mathbf{A}}_z(\mathbf{C}^{-1}\hat{\mathbf{A}}_z)']\mathbf{S}_x^{-1}\mathbf{C}^{-1}\hat{\mathbf{A}}_z. \quad (6-6)$$

By comparing (6-5) and (6-6) we find that if \mathbf{A} is a solution to (6-5) then

$$\mathbf{A}_z = \mathbf{C}^{-1}\mathbf{A}$$

will be a solution to (6-6). This means that a scaling of the X (the observations) with the matrix \mathbf{C} implies that the factor weights are scaled by \mathbf{C}^{-1} .

If we retain the assumption of normality we can test if the factor model is valid i.e. test

$$H_0 : \Sigma = \Delta + \mathbf{A}\mathbf{A}' \quad \text{against} \quad H_1 : \Sigma \text{ arbitrary.}$$

The ratio test will then be equivalent to the test given by the test statistic

$$Z = (n - 1 - \frac{1}{6}(2k + 5) - \frac{2}{3}m) \ln \frac{|\hat{\Delta} + \hat{\mathbf{A}}\hat{\mathbf{A}}'|}{|\mathbf{S}|}$$

and we will reject for

$$Z > \chi^2(\frac{1}{2}\{(k - m)^2 - k - m\}).$$

||| Example 6.22

In the following table we have shown the result of a principle factor solution (PCA), and a maximum likelihood solution (ML) and finally a little Jiffy solution (see [21]).

The data consists of 198 samples of Portland cement where each sample is analysed for 15 variables (contents of different cement minerals, fine grainedness etc.). The 15 variables have only been given by their respective numbers because we do not consider the interpretation here but only the comparison of the three methods. In the table, weights, which are numerically less than 0.25, have been set equal to 0 to ease the interpretation.

We note that the three methods give remarkably similar results. For factor three we note that the PCA solution differs somewhat from the ML and the LJIF solutions.

Variable	Factor1			Factor2			Factor3		
	PCA	ML	LJIF	PCA	ML	LJIF	PCA	ML	LJIF
1	-0.26	0	0	0.95	0.91	0.95	0	0.36	0
2	0	0	0	-0.98	-1.00	-0.99	0	0	0
3	-0.50	0.93	1.08	0	0	0	-0.40	-0.34	-0.72
4	0.94	-0.78	-0.80	0	0	0	0	-0.62	-0.32
5	0	0.29	0.34	0	0	0	-0.48	0	0
6	0	0	0	0	0	0	0	0	-0.25
7	0	0	0	0	0	0	0	0	0
8	0.53	-0.32	-0.32	0	0	0	0.27	-0.31	0
9	0.90	-0.72	-0.76	0	0	0	0	-0.45	0
10	0	0	0	0	0	0	0.72	0	0
11	0	-0.28	-0.31	0	0	0	0.82	0	0
12	0	0	0	0	0	0	-0.78	0	0
13	-0.73	0	0	0	0	0	0	0.98	0.95
14	-0.86	0.97	1.05	0	0	0	-0.31	0	0
15	0	0.25	0	0.93	0.93	0.92	0	0	-0.35

6.3.6 Q-mode analysis

In the form of factor analysis we have regarded up till now - the so-called R-modus analysis - one investigates the correlations between the different variables. The samples of the individuals etc. are used as repetitions and these are used to estimate the different correlations. If we call the observations X_1, \dots, X_n and let

$$\mathbf{X}' = \begin{bmatrix} X_{11} & \dots & X_{1n} \\ \vdots & & \vdots \\ X_{k1} & \dots & X_{kn} \end{bmatrix},$$

where the rows corresponds to the single variables and the columns to the individuals. If we assume that the observations have been normalised so they have mean value 0 and variance 1 we get the correlation matrix as

$$\mathbf{R} = \mathbf{X}'\mathbf{X},$$

cf. theorem 1.30. In a **dual** way we could of course define

$$\mathbf{Q} = \mathbf{XX}',$$

and then interpret it as an expression for the correlation between individuals and then perform a factor analysis on these. The results of such a procedure will be a classification of individuals into groups which are close to each other.

We give a small example which comes from [22].

||| Example 6.23

We consider 12 stream sediment samples collected in Jameson Land in East Greenland. They are analysed for 7 elements which are Cu, Ni, V, Pb, Zr, Ca and Ba. An ordinary R-modus analysis showed that the two first factors described $42\% + 37\% = 79\%$ of the variation. In the following figure we have shown the rotated factor weights.

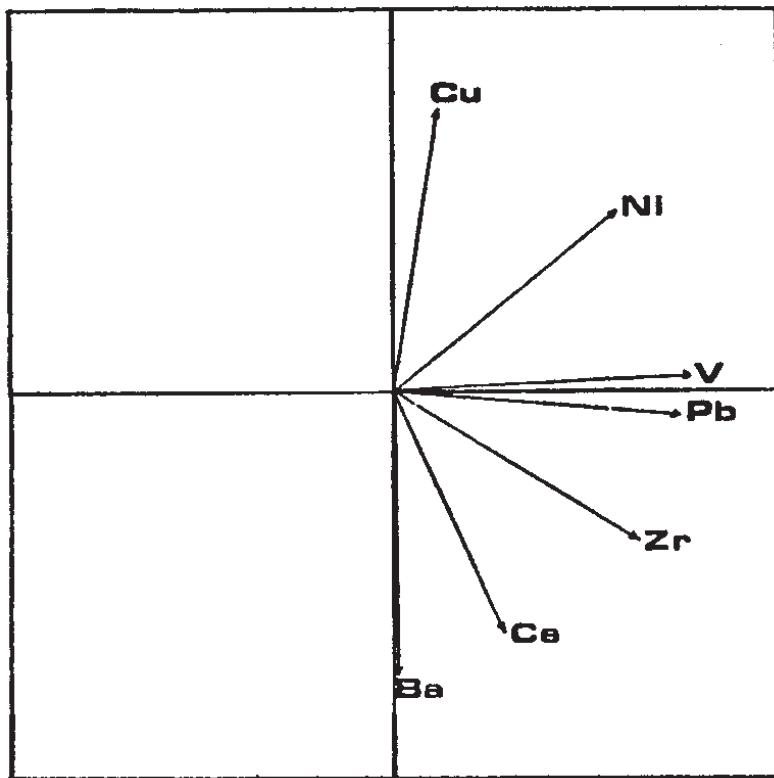


Figure 6.23: Factor weights in R-modus analysis.

Then a Q-modus analysis was performed as mentioned above. This gave a first factor which described 38% of the total variation and a second factor which described 26% of the total variation.

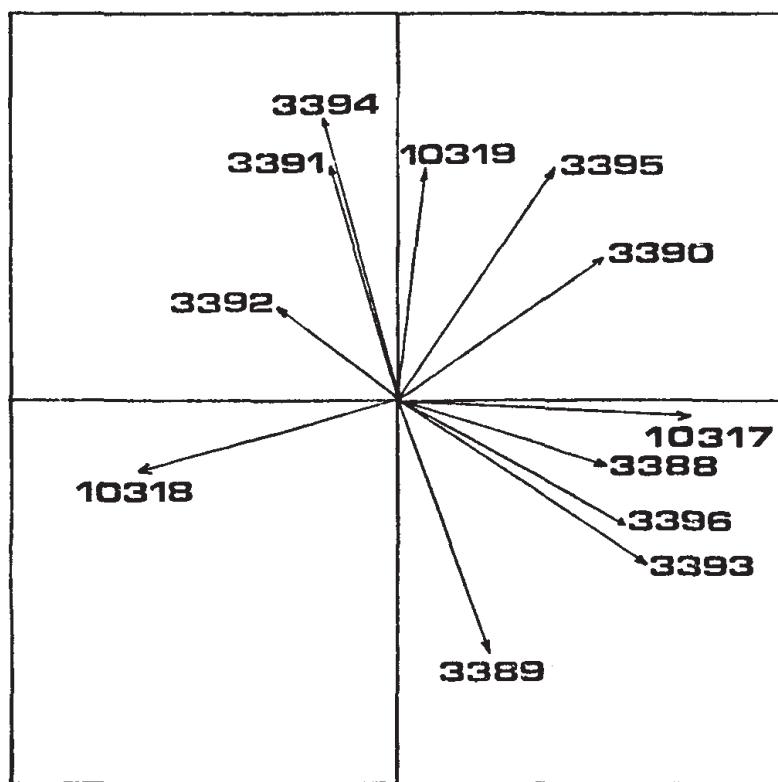


Figure 6.23: Factor

weights in Q-modus analysis.

From the figure with the factor weights we now get a direct comparison of the different samples. This could also be obtained through R-modus analysis but we would then have to go via the factor scores.

Analysis of this kind is used in mineral prospecting in the attempt to determine which samples are to be declared non-normal and thereby interesting.

When performing a Q-modus analysis one will often end up with a large amount of computations since the Q-matrix is of the order $n \times n$, where n is the number of individuals. One can then draw advantage of the theorems in section A.4.2. From these we see that the eigenvalues which are different from 0 in R and Q are equal and there is a simple relationship between the eigenvectors. Since R is only of the order $k \times k$ and the number of variables usually is considerably less than the number of individuals it is possible to save a lot of numerical work.

Finally we remark that Q-modus analysis often is not performed on $\mathbf{X}\mathbf{X}'$ but on another matrix containing some more or less arbitrarily chosen *similarity measures*. The technique is, however, unchanged and one can still obtain computational savings by using the above mentioned relation between R-modus and Q-modus analysis. For special choices of similarity measures one often calls this a *principal coordinate analysis*.

An attempt to do both analyses at one time is found in the so-called *correspondence analysis* which is due to Benzécri (1973).

6.4 PLS – Regression and Projection on Latent Structure

6.4.1 Introduction

We consider n independent random variables

$$\mathbf{Z}_i \sim N_{m+k}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $m \leq k$ and \mathbf{Z}_i and the parameters have been partitioned as follows:

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{Y}_i \\ \mathbf{X}_i \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix}.$$

We shall assume that the variables are centered and scaled appropriately (often standardized to have variance 1), i.e.

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Furthermore, the observations are organized in a data matrix

$$[\mathbf{Z}] = [\mathbf{Y} \ \mathbf{X}] = \begin{bmatrix} \mathbf{Y}_1^T & \mathbf{X}_1^T \\ \vdots & \vdots \\ \mathbf{Y}_n^T & \mathbf{X}_n^T \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1m} & X_{11} & \cdots & X_{1k} \\ \vdots & & \vdots & \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{nm} & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

Then we have the unbiased estimator $\hat{\boldsymbol{\Sigma}}$ given by

$$(n-1)\hat{\boldsymbol{\Sigma}} = [\mathbf{Y} \ \mathbf{X}]^T [\mathbf{Y} \ \mathbf{X}] = \begin{bmatrix} \mathbf{Y}^T \mathbf{Y} & \mathbf{Y}^T \mathbf{X} \\ \mathbf{X}^T \mathbf{Y} & \mathbf{X}^T \mathbf{X} \end{bmatrix}$$

6.4.2 Ordinary least squares regression.

If we want to predict \mathbf{Y} linearly based on \mathbf{X} , i.e. assume a model

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

or

$$\begin{bmatrix} \mathbf{Y}_1^T \\ \vdots \\ \mathbf{Y}_n^T \end{bmatrix} = [\mathbf{X} \ \mathbf{b}_1 \ \cdots \ \mathbf{X}\mathbf{b}_m] + \mathbf{E} = \left[\mathbf{X} \begin{bmatrix} b_{11} \\ \vdots \\ b_{k1} \end{bmatrix} \ \cdots \ \mathbf{X} \begin{bmatrix} b_{1m} \\ \vdots \\ b_{km} \end{bmatrix} \right] + \mathbf{E}$$

we get the *ordinary least squares* estimate by variable-wise estimation

$$\widehat{\mathbf{Y}}_{OLS} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{X}\widehat{\mathbf{B}}_{OLS}$$

i.e. the regression coefficients coefficients for *ordinary least squares regression* are

$$\widehat{\mathbf{B}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

6.4.3 Principal components regression

Next, we consider *principal components regression*. The first a eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$ are collected in two matrices

$$\begin{aligned} \Lambda_a &= \text{diag} [\lambda_1 \ \cdots \ \lambda_a] \\ \mathbf{P}_a &= [\mathbf{p}_1 \ \cdots \ \mathbf{p}_a] \end{aligned}$$

and we replace \mathbf{X} by the first f principal components

$\mathbf{X}\mathbf{P}_f$. Since

$$(\mathbf{X}\mathbf{P}_f)^T(\mathbf{X}\mathbf{P}_f) = \mathbf{P}_f^T\mathbf{X}^T\mathbf{X}\mathbf{P}_f = \mathbf{P}_f^T\mathbf{P}_k\Lambda_k\mathbf{P}_k^T\mathbf{P}_f = [\mathbf{I}_f \ \mathbf{0}] \Lambda_k \begin{bmatrix} \mathbf{I}_f \\ \mathbf{0} \end{bmatrix} = \Lambda_f$$

it follows that

$$\widehat{\mathbf{Y}}_{PCA} = \mathbf{X}\mathbf{P}_f \left((\mathbf{X}\mathbf{P}_f)^T(\mathbf{X}\mathbf{P}_f) \right)^{-1} (\mathbf{X}\mathbf{P}_f)^T\mathbf{Y} = \mathbf{X}\mathbf{P}_f \Lambda_f^{-1} \mathbf{P}_f^T \mathbf{X}^T \mathbf{Y}$$

i.e. the regression coefficients coefficients for *principal component regression* are

$$\widehat{\mathbf{B}}_{PCA} = \mathbf{P}_f \Lambda_f^{-1} \mathbf{P}_f^T \mathbf{X}^T \mathbf{Y}$$

6.4.4 Canonical correlation regression

.

For *canonical correlation regression* we consider the matrices A_f and B_f containing the coefficients for computing the canonical variables, i.e. we have the canonical variables for the i 'th observation

$$V_i = \begin{bmatrix} V_{i1} \\ \vdots \\ V_{if} \end{bmatrix} = [\ a_1 \ \cdots \ a_f]^T Y_i = A_f^T Y_i$$

$$W_i = \begin{bmatrix} W_{i1} \\ \vdots \\ W_{if} \end{bmatrix} = [\ b_1 \ \cdots \ b_f]^T X_i = B_f^T X_i$$

We collect those in matrices as

$$[\ V \ \ W \] = \begin{bmatrix} V_1^T & W_1^T \\ \vdots & \vdots \\ V_n^T & W_n^T \end{bmatrix} = \begin{bmatrix} Y_1^T A_f & X_1^T B_f \\ \vdots & \vdots \\ Y_n^T A_f & X_n^T B_f \end{bmatrix} = [\ Y \ \ X \] \begin{bmatrix} A_f \\ B_f \end{bmatrix}$$

From the properties of the canonical variates we immediately obtain

$$A_f^T Y^T Y A_f = B_f^T X^T X B_f = (n - 1) I_f$$

$$A_f^T Y^T X B_f = B_f^T X^T Y A_f = (n - 1) \Gamma_f = (n - 1) \cdot \text{diag}(\varrho_1, \dots, \varrho_f)$$

We replace X by the first f canonical variates $X B_f$. Since

$$(X B_f)^T (X B_f) = (n - 1) I_f$$

it follows that

$$\widehat{Y}_{CCA} = X B_f \left((X B_f)^T (X B_f) \right)^{-1} (X B_f)^T Y = \frac{\mathbf{1}}{n-1} X B_f B_f^T X^T Y$$

i.e. the regression coefficients for *canonical correlation regression* are

$$\widehat{B}_{CCA} = \frac{\mathbf{1}}{n-1} B_f B_f^T X^T Y$$

6.4.5 Reduced rank regression

We first state a useful result on matrix approximation. For given matrices A and M_f , the squared normed difference is

$$\|A - M_f\|_2^2 = \sum_{i,j} (a_{ij} - m_{ij})^2 = \text{tr}((A - M_f)^T (A - M_f))$$

We now want to minimize this squared norm with respect to M_f among matrices that has rank f . The best approximation in a least squares sense of A with a matrix \widehat{M}_f of rank f is given by the first f terms in the singular value decomposition

$$\widehat{M}_f = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \cdots + \sigma_f u_f v_f^T$$

If the singular values are different, \widehat{M}_f is unique.

We now seek a solution to the least squares problem with limited rank (f). We have

$$\|\mathbf{Y} - \mathbf{X}\mathbf{B}_f\|_2^2 = \|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{B}}_{OLS} + \mathbf{X}\widehat{\mathbf{B}}_{OLS} - \mathbf{X}\mathbf{B}_f\|_2^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}_{OLS} + \widehat{\mathbf{Y}}_{OLS} - \mathbf{X}\mathbf{B}_f\|_2^2$$

Since

$$(\mathbf{Y} - \widehat{\mathbf{Y}}_{OLS})^T (\widehat{\mathbf{Y}}_{OLS} - \mathbf{X}\mathbf{B}_f) = (\mathbf{Y} - \widehat{\mathbf{Y}}_{OLS})^T \mathbf{X}(\widehat{\mathbf{B}}_{OLS} - \mathbf{B}_f) = \mathbf{0}$$

we get

$$\|\mathbf{Y} - \mathbf{X}\mathbf{B}_f\|_2^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}_{OLS}\|_2^2 + \|\widehat{\mathbf{Y}}_{OLS} - \mathbf{X}\mathbf{B}_f\|_2^2$$

The best rank f approximation to $\mathbf{X}\widehat{\mathbf{B}}_f$ is thus given by the first f singular values of $\widehat{\mathbf{Y}}_{OLS}$. If those are given by – assuming that the rank of $\widehat{\mathbf{Y}}_{OLS}$ is m –

$$\boldsymbol{\Gamma}_m = \text{diag} [\gamma_1 \ \cdots \ \gamma_m]$$

$$\mathbf{U}_m = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_m]$$

$$\mathbf{V}_m = [\mathbf{v}_1 \ \cdots \ \mathbf{v}_m]$$

$\boldsymbol{\Gamma}_r$ is $(m \times m)$, \mathbf{U}_r $(n \times m)$, and \mathbf{V}_r is $(m \times m)$. Then

$$\widehat{\mathbf{Y}}_{OLS} = \mathbf{U}_m \boldsymbol{\Gamma}_m \mathbf{V}_m^T$$

and the rank f approximation to this is

$$\mathbf{X}\widehat{\mathbf{B}}_f = \mathbf{U}_f \boldsymbol{\Gamma}_f \mathbf{V}_f^T$$

now

$$\mathbf{V}^T \mathbf{V}_f = \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_m^T \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_f \end{bmatrix} = \begin{bmatrix} \mathbf{I}_f \\ \mathbf{0} \end{bmatrix}$$

and we have

$$\widehat{\mathbf{Y}}_{OLS} \mathbf{V}_f \mathbf{V}_f^T = \mathbf{X}\widehat{\mathbf{B}}_{OLS} \mathbf{V}_f \mathbf{V}_f^T = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^T \mathbf{V}_f \mathbf{V}_f^T = \mathbf{U}_f \boldsymbol{\Gamma}_f \mathbf{V}_f^T = \mathbf{X}\widehat{\mathbf{B}}_f$$

i.e. *the reduced rank regression coefficients* are

$$\widehat{\mathbf{B}}_{RRR} = \widehat{\mathbf{B}}_f = \widehat{\mathbf{B}}_{OLS} \mathbf{V}_f \mathbf{V}_f^T$$

6.4.6 Covariance maximization

We now consider a $k \times 1$ weight vector \mathbf{r} with length 1, i.e. $\mathbf{r}^T \mathbf{r} = 1$ and a $m \times 1$ weight vector \mathbf{q} with length 1, i.e. $\mathbf{q}^T \mathbf{q} = 1$ and have

$$Cov \left(\mathbf{r}^T \mathbf{X}_i, \mathbf{q}^T \mathbf{Y}_i \right) = \mathbf{r}^T \boldsymbol{\Sigma}_{xy} \mathbf{q}$$

and

$$\widehat{Cov} \left(\mathbf{r}^T \mathbf{X}_i, \mathbf{q}^T \mathbf{Y}_i \right) = \frac{1}{n-1} \mathbf{r}^T \mathbf{X}^T \mathbf{Y} \mathbf{q}$$

Maximizing this empirical covariance wrt. \mathbf{r} and \mathbf{q} under the constraints $\mathbf{r}^T \mathbf{r} = 1$ and $\mathbf{q}^T \mathbf{q} = 1$ is equivalent to maximizing $\mathbf{r}^T \mathbf{X}^T \mathbf{Y} \mathbf{q}$ under the same constraints. The solution to the latter problem is the largest singular value for $\mathbf{X}^T \mathbf{Y}$ obtained for \mathbf{r} and \mathbf{q} equal to the first left and right singular vectors for $\mathbf{X}^T \mathbf{Y}$. If we want to consider further directions $(\mathbf{r}_i, \mathbf{q}_i)$, $i = 1, \dots, f$, $f \leq \min(k, m)$, orthogonal to the previous ones in maximizing the covariances, i.e.

$$\max_{\mathbf{r}_i, \mathbf{q}_i} \mathbf{r}_i^T \mathbf{X}^T \mathbf{Y} \mathbf{q}_i \quad \text{subject to} \quad \mathbf{r}_i^T \mathbf{r}_i = 1, \quad \mathbf{q}_i^T \mathbf{q}_i = 1 \quad \& \quad \mathbf{r}_j^T \mathbf{r}_i = 0, \quad \mathbf{q}_j^T \mathbf{q}_i = 0, \quad j < i,$$

the solutions are the subsequent singular values and corresponding singular vectors. The relationship between the singular vectors are

$$\begin{aligned}\mathbf{r}_i &= \frac{1}{\gamma_i} \mathbf{X}^T \mathbf{Y} \mathbf{q}_i \\ \mathbf{q}_i &= \frac{1}{\gamma_i} \mathbf{Y}^T \mathbf{X} \mathbf{r}_i\end{aligned}$$

where γ_i is the i 'th singular value.

6.4.7 Partial least squares regression

We now seek relations

$$\begin{aligned}\mathbf{X} &= \mathbf{T} \mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{T} \mathbf{C}^T + \mathbf{F}\end{aligned}$$

where \mathbf{E} and \mathbf{F} are residuals, \mathbf{T} ($n \times f$) is the X factor score matrix and \mathbf{P} ($k \times f$) and \mathbf{C} ($n \times f$) are the *model effect (X) loadings* and the *dependent variables (Y) loadings* respectively. The X factor scores are assumed orthogonal, i.e. $\mathbf{T}^T \mathbf{T}$ is diagonal.

We predict \mathbf{X} and \mathbf{Y} by regression on \mathbf{T} , i.e.

$$\begin{aligned}\widehat{\mathbf{X}}_{PLS} &= \mathbf{T} (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{X} = \mathbf{T} \widehat{\mathbf{P}}^T \\ \widehat{\mathbf{Y}}_{PLS} &= \mathbf{T} (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{Y} = \mathbf{T} \widehat{\mathbf{C}}^T\end{aligned}$$

where $\widehat{\mathbf{P}} = \mathbf{X}^T \mathbf{T} (\mathbf{T}^T \mathbf{T})^{-1}$ are the *model effect (X) loadings* and $\widehat{\mathbf{C}} = \mathbf{Y}^T \mathbf{T} (\mathbf{T}^T \mathbf{T})^{-1}$ the *dependent variables (Y) loadings*.

We now consider prediction of \mathbf{Y}_i from a X -latent vector, i.e. a linear combination of the components of \mathbf{X}_i , or $\mathbf{r}^T \mathbf{X}_i$, where \mathbf{r} still is a $k \times 1$ weight vector

with length 1, i.e. $\mathbf{r}^T \mathbf{r} = 1$. In order to determine \mathbf{r} we may take a look at the covariances between $\mathbf{r}^T \mathbf{X}_i$ and the components of \mathbf{Y}_i , i.e.

$$[\text{Cov}(\mathbf{r}^T \mathbf{X}_i, Y_{i1}) \ \dots \ \text{Cov}(\mathbf{r}^T \mathbf{X}_i, Y_{im})] = \text{Cov}(\mathbf{r}^T \mathbf{X}_i, \mathbf{Y}_i) = \mathbf{r}^T \boldsymbol{\Sigma}_{xy}$$

The **sum of those m covariances squared** becomes

$$\mathbf{r}^T \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yx} \mathbf{r}$$

The estimate of this quantity is proportional to

$$\mathbf{r}^T \mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X} \mathbf{r}$$

A sensible choice for \mathbf{r} would be to maximize this sum of squared covariances under the constraint $\mathbf{r}^T \mathbf{r} = 1$. The solution is the largest eigenvalue of $\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$ obtained for $\mathbf{r} = \mathbf{r}_1$, the corresponding eigenvector. The k eigenvalues of the $k \times k$ matrix $\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$ are the squares of the singular values of the $k \times m$ matrix $\mathbf{X}^T \mathbf{Y}$, and \mathbf{r}_1 is the left singular vector corresponding to the largest singular value. In other words, we obtain the same solution for \mathbf{r} as we had earlier where we solved the covariance maximization problem

$$\max_{\mathbf{r}, \mathbf{q}} (n - 1) \widehat{\text{Cov}}(\mathbf{r}^T \mathbf{X}_i, \mathbf{q}^T \mathbf{Y}_i) = \max_{\mathbf{r}, \mathbf{q}} \mathbf{r}^T \mathbf{X}^T \mathbf{Y} \mathbf{q} \text{ subject to } \mathbf{r}^T \mathbf{r} = 1 \text{ and } \mathbf{q}^T \mathbf{q} = 1$$

i.e. the largest singular value of $\mathbf{X}^T \mathbf{Y}$ and the maximum is attained at the corresponding left and right singular vectors of $\mathbf{X}^T \mathbf{Y}$, i.e. the eigenvectors of $\mathbf{X}^T \mathbf{Y} \mathbf{Y}^T \mathbf{X}$ and $\mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Y}$ corresponding to the (common) largest eigenvalue of those matrices.

More directions are obtained by adding another constraint to the maximization problem, i.e. for $\mathbf{t}_i = \mathbf{X} \mathbf{r}_i$

$$\max_{\mathbf{r}_i, \mathbf{q}_i} \mathbf{r}_i^T \mathbf{X}^T \mathbf{Y} \mathbf{q}_i \text{ subject to } \mathbf{r}_i^T \mathbf{r}_i = 1, \mathbf{q}_i^T \mathbf{q}_i = 1 \text{ and } \mathbf{r}_j^T \mathbf{X}^T \mathbf{X} \mathbf{r}_i = \mathbf{t}_j^T \mathbf{t}_i = 0, j < i.$$

We put

$$\mathbf{S}_1 = \mathbf{X}^T \mathbf{Y}$$

giving

$$\mathbf{S}_1^T \mathbf{S}_1 = \mathbf{Y}^T \mathbf{X} \mathbf{X}^T \mathbf{Y}.$$

We define iteratively (assuming that vectors and matrices corresponding to $i = 0$ are zero)

$S_1 = X^T Y, \quad S_1^T S_1 = Y^T X X^T Y$		
$S_i = S_{i-1} - v_{i-1}(v_{i-1}^T S_{i-1})$		Deflate S_{i-1} wrt current X block factor loadings
q_i principal eigenvector of $S_i^T S_i$	$Q_i = [q_1 \cdots q_i]$	Y block factor weights
$r_i = S_i q_i / \ X S_i q_i\ $	$R_i = [r_1 \cdots r_i]$	X block factor weights
$t_i = X r_i$	$T_i = [t_1 \cdots t_i] = X R_i$	$\propto X$ block factor scores
$p_i = X^T t_i$	$P_i = [p_1 \cdots p_i] = X^T T_i$	$\propto X$ block factor loadings
$h_i = Y^T t_i$	$H_i = [h_1 \cdots h_i] = Y^T T_i$	$\propto Y$ block factor loadings
$v_i = v'_i / \ v'_i\ $, where $v'_{i-1} = p_i - v_{i-1}(v_{i-1}^T p_i)$	$V_i = [v_1 \cdots v_i]$	Orthogonalization of X block factor loadings
$u_i = Y q_i - T_{i-1}(T_{i-1}^T Y q_i)$	$U_i = [u_1 \cdots u_i]$	Y block factor scores

It follows that

$$T_f^T T_f = I_f$$

Furthermore, we introduce the “calibration” matrix

$$D = D_f = \text{diag}(\|X^T t_1\|, \dots, \|X^T t_f\|) = \text{diag}(\|p_1\|, \dots, \|p_f\|)$$

Then the X factor scores becomes

$$T = T_f D$$

giving

$$T^T T = D^T D = \text{diag}(\|X^T t_1\|^2, \dots, \|X^T t_f\|^2)$$

The *percent variation accounted for* by the SIMPLS factors are for the model effects and for the dependent variables respectively the diagonal elements in

$$P_f^T P_f = T_f^T X X^T T_f \text{ respectively } H_f^T H_f = T_f^T Y Y^T T_f$$

The *model effect (X) loadings* and the *dependent variables (Y) loadings* are

$$P = X^T T (T^T T)^{-1} = X^T T_f D^{-1},$$

respectively

$$\mathbf{C} = \mathbf{Y}^T \mathbf{T} (\mathbf{T}^T \mathbf{T})^{-1} = \mathbf{H}_f \mathbf{D}^{-1} = \mathbf{Y}^T \mathbf{T}_f \mathbf{D}^{-1}$$

and the \mathbf{Y} factor scores

$$\mathbf{U}_f.$$

By design $\mathbf{U}_f^T \mathbf{T}$ is lower triangular implying that the $\mathbf{Y} - score_i$ and the $\mathbf{X} - score_j$ values are (empirically) uncorrelated for $i > j$, i.e. the $\mathbf{Y} - score$ values are uncorrelated with the previous $\mathbf{X} - score$ values.

The *model effect weights* (\mathbf{X} block factor weights) are

$$\mathbf{R} = \mathbf{R}_f \mathbf{D}$$

and the *dependent variable weights* (\mathbf{Y} block factor weights) are

$$\mathbf{Q}_f$$

If we want to express the predicted value $\hat{\mathbf{Y}}$ as a function of the \mathbf{X} -variables, we see that

$$\mathbf{X} \mathbf{R}_f \mathbf{H}_f^T = \mathbf{X} \mathbf{R}_f \mathbf{T}_f^T \mathbf{Y} = \mathbf{T}_f \mathbf{T}_f^T \mathbf{Y} = \mathbf{T} (\mathbf{D}^{-1} \mathbf{D}^{-1}) \mathbf{T}^T \mathbf{Y} = \hat{\mathbf{Y}}$$

and thus the *partial least squares (SIMPLS) regression coefficients* are

$$\hat{\mathbf{B}}_{PLS} = \mathbf{B}_f = \mathbf{R}_f \mathbf{H}_f^T$$

|||| Appendix A

Summary of linear algebra

This chapter contains a summary of linear algebra with special emphasis on its use in statistics. The chapter is not intended to be an introduction to the subject. Rather it is a summary of an already known subject. Therefor we will not give very many examples within the areas typically covered in algebra and geometry courses. However, we will give more examples and sometimes proofs within areas which usually do not receive much attention in all-round courses, but which do enjoy significant use within algebra in statistics.

In the course of analysis of multidimensional statistical problems one often needs to invert non-regular matrices. For instance this is the case if one considers a problem given on a true sub-space of the considered n -dimensional vector-space. Instead of just considering the relevant sub-space, many authors prefer giving partly algebraic solutions by introducing the pseudoinverse of a non-regular matrix. In order to ease the reading of other literature we will introduce this concept and try to visualize it geometrically.

We note that use of pseudoinverse matrices gives a very convenient way to solve many matrix equations in an algorithmic form.

A.1 Vector space

We start by giving an overview of the definition and elementary properties in the fundamental concept of a linear vector space.

A.1.1 Definition of a vector space

||| Definition A.1

A *vector space (on the real numbers)* is a set V with a composition rule $+$ in the set $V \times V \rightarrow V$ which is called *vector addition* and a composition rule \cdot in $R \times V \rightarrow V$ called *scalar multiplication*, which obey

- i) $\forall \mathbf{u}, \mathbf{v} \in V : \mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ (*commutative law for vector addition*)
- ii) $\forall \mathbf{u}, \mathbf{v}, \mathbf{x} \in V : \mathbf{u} + (\mathbf{v} + \mathbf{x}) = (\mathbf{u} + \mathbf{v}) + \mathbf{x}$ (*associative law for vector addition*)
- iii) $\exists \mathbf{0} \in V \forall \mathbf{u} \in V : \mathbf{u} + \mathbf{0} = \mathbf{u}$ (*existence of a neutral element*)
- iv) $\forall \mathbf{u} \in V \exists -\mathbf{u} \in V : \mathbf{u} + (-\mathbf{u}) = \mathbf{0}$ (*existence of an inverse element*)
- v) $\forall \lambda \in R \forall \mathbf{u}, \mathbf{v} \in V : \lambda(\mathbf{u} + \mathbf{v}) = \lambda\mathbf{u} + \lambda\mathbf{v}$ (*distributive law for scalar multiplication*)
- vi) $\forall \lambda_1, \lambda_2 \in R \forall \mathbf{u} \in V : (\lambda_1 + \lambda_2)\mathbf{u} = \lambda_1\mathbf{u} + \lambda_2\mathbf{u}$ (*distributive law for scalar multiplication*)
- vii) $\forall \lambda_1, \lambda_2 \in R \forall \mathbf{u} \in V : (\lambda_1\lambda_2)\mathbf{u} = \lambda_1(\lambda_2\mathbf{u})$ (*associative law for scalar multiplication*)
- viii) $\forall \mathbf{u} \in V : 1\mathbf{u} = \mathbf{u}$

||| Example A.2

It is readily shown that all ordered n -tuples

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

of real numbers constitute a vector space if the compositions are defined element by element, i.e.

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

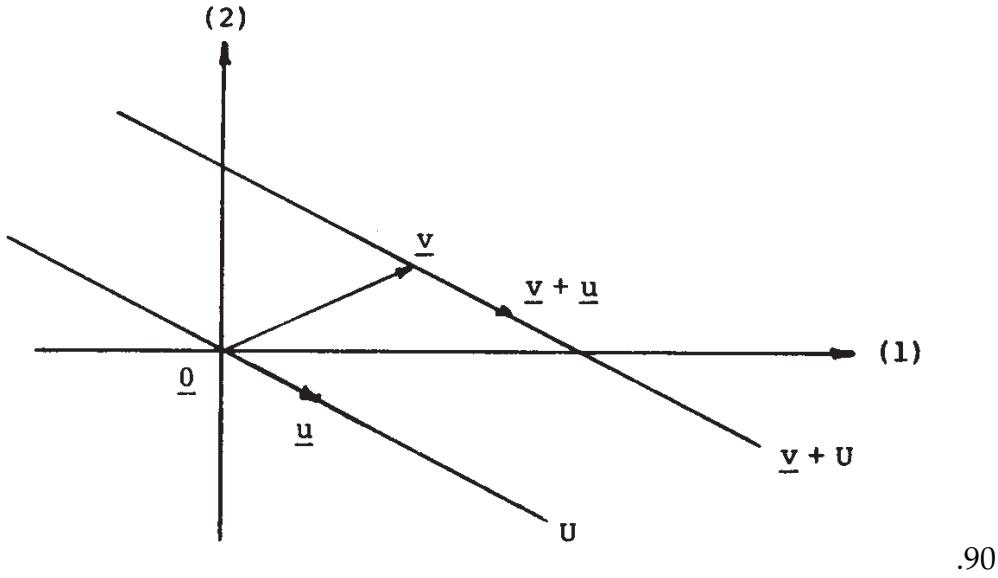
and

$$\lambda \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{bmatrix}$$

This vector space is denoted R^n

A vector space U which is subset of a vector space V is called a *subspace* in V . On the other hand, if we consider vectors $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$, we can define the *linear span* of those vectors

$$\text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$$

Figure A.1: Sub-space and corresponding side-subspace in R^2 .

as the smallest subspace of V , which contains $\{v_1, \dots, v_k\}$. It is easily shown that

$$\text{span}\{v_1, \dots, v_k\} = \left\{ \sum_{i=1}^k \alpha_i v_i \mid \alpha_i \in R, \quad i = 1, \dots, k \right\}.$$

A vector of the form $\sum \alpha_i v_i$ is called a *linear combination* of the vectors $v_i, i = 1, \dots, k$. The above result can then be expressed such that $\text{span}\{v_1, \dots, v_k\}$ precisely consists of all linear combinations of the vectors v_1, \dots, v_k . Generally we define

$$\text{span}(U_1, \dots, U_p)$$

where $U_i \subseteq V$, as the smallest subspace of V , which contains all $U_i, i = 1, \dots, p$.

A side-subspace is a set of the form

$$v + U = \{v + u \mid u \in U\},$$

where U is a sub-space in V .

The situation is sketched in fig. A.1.

Vectors v_1, \dots, v_n are said to be *linearly independent* if the relation

$$\alpha_1 v_1 + \cdots + \alpha_n v_n = 0$$

implies that

$$\alpha_1 = \cdots = \alpha_n = 0$$

In the opposite case they are said to be *linearly dependent* and at least one of them can be expressed as a linear combination of the other two.

A *basis* for the vector space V is a set of linearly independent vectors which span all of V . Any vector can be expressed unambiguously as a linear combination of vectors in a basis. The number of elements in different bases of a vector space is always the same. If this number is finite it is called the *dimension* of the vector space and it is written $\dim(V)$.

||| Example A.3

\mathbb{R}^n has the basis

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

and is therefore n -dimensional

In an expression like

$$\mathbf{v} = \sum_{i=1}^n \alpha_i \mathbf{v}_i$$

where $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ is a basis for V , we call the set $\alpha_1, \dots, \alpha_n$ \mathbf{v} 's *coordinates* with respect to the basis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$. <U+FFF><U+FFF>

A.1.2 Direct sum of vector spaces

Let V be a vector space (of finite dimension) and let U_1, \dots, U_k be sub-spaces of V . We then say that V is the *direct sum* of the sub-spaces U_1, \dots, U_k , and we write

$$V = U_1 \oplus \dots \oplus U_k = \bigoplus_{i=1}^k U_i,$$

if an arbitrary vector $\mathbf{v} \in V$ in exactly one way can be expressed like

$$\mathbf{v} = \mathbf{u}_1 + \dots + \mathbf{u}_k, \quad \mathbf{u}_1 \in U_1, \dots, \mathbf{u}_k \in U_k \tag{A-1}$$

This condition is equivalent to that for vectors $\mathbf{u}_i \in U_i$ the following holds true

$$\mathbf{u}_1 + \dots + \mathbf{u}_k = \mathbf{0} \Rightarrow \mathbf{u}_1 = \dots = \mathbf{u}_k = \mathbf{0}.$$

This is again equivalent to

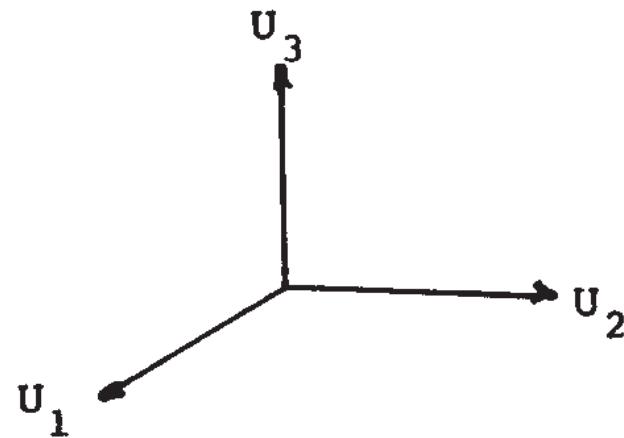
$$\dim(\text{span}(U_1, \dots, U_k)) = \sum_{i=1}^k \dim U_i = \dim V$$

Finally, this is equivalent to that all unions of some of the U_i 's are $\mathbf{0}$. Of course, it is a general condition that $\text{span}(U_1, \dots, U_k) = V$, i.e. that it is at all possible to find an expression like A-1. It is the unambiguity of A-1 which implies that we may call the "sum" direct.

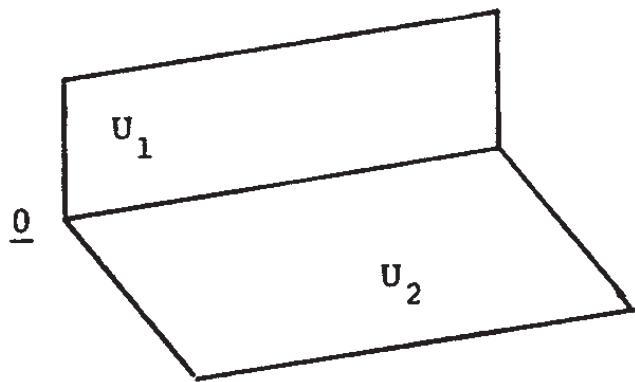
We sketch some examples below in fig. A.2.

If V is partitioned into a direct sum

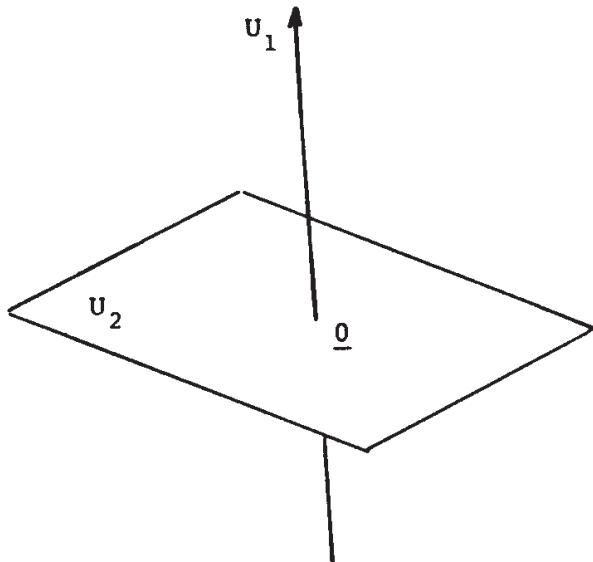
$$V = U_1 \oplus \dots \oplus U_k$$



$U_1 \oplus U_2 \oplus U_3 = R^3$ The sum is direct because for instance $\dim U_1 + \dim U_2 + \dim U_3 = 3$



R^3 is not a direct sum of U_1 and U_2 ; because $\dim U_1 + \dim U_2 = 4$



Here $U_1 \oplus U_2 = R^3$ because for instance U_1 and U_2 besides spanning R^3 also satisfy $U_1 \cap U_2 = \mathbf{0}$

Figure A.2:

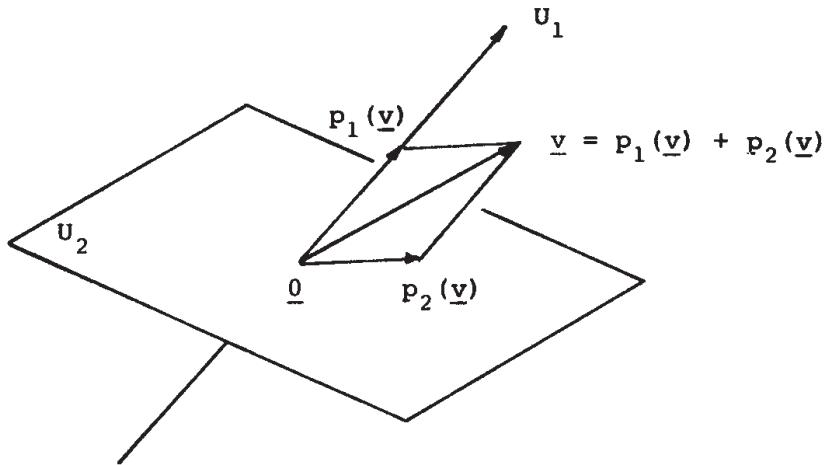


Figure A.3: Projection of a vector.

then we call any arbitrary vector v 's component in U_i for v 's *projection* onto U_i (by the direction determined by $U_1, \dots, U_{i-1}, U_{i+1}, \dots, U_k$) and we denote it $p_i(v)$. The situation is sketched in fig. A.3.

The projection p_i is *idempotent*, i.e. $p_i \circ p_i(v) = p_i(v), \forall v$ where $f \circ g$ denotes the combination of f and g .

A.2 Linear transformations and matrices

We start with a section on *linear transformations* (or *linear mappings*).

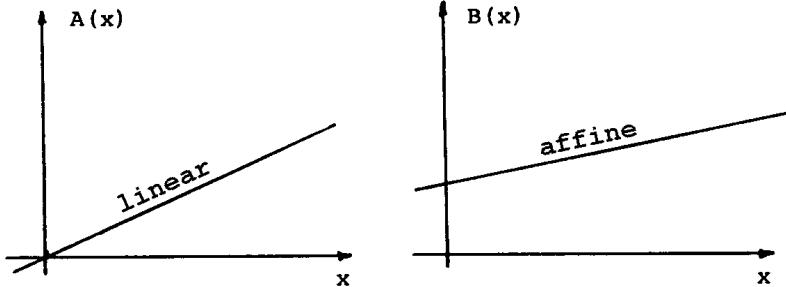
A.2.1 Linear transformations

A transformation (or mapping) $A : U \rightarrow V$, where U and V are vector spaces is said to be *linear* if

$$\begin{aligned} \forall \lambda_1, \lambda_2 \in R \quad & \forall \mathbf{u}_1, \mathbf{u}_2 \in U : A(\lambda_1 \mathbf{u}_1 + \lambda_2 \mathbf{u}_2) = \\ & \lambda_1 A(\mathbf{u}_1) + \lambda_2 A(\mathbf{u}_2) \end{aligned}$$

||| Example A.4

A transformation $A : R \rightarrow R$ is linear if its graph is a straight line through $(0,0)$. If the graph is a straight line which does not pass through $(0,0)$ we say the transformation is *affine*.



10

Figure A.4: Graphs for a linear and an affine transformation $R \rightarrow R$.

By the *null-space* $N(A)$ of a linear transformation $A : U \rightarrow V$ we mean the sub-space

$$A^{-1}(\mathbf{0}) = \{\mathbf{u} | A(\mathbf{u}) = \mathbf{0}\}$$

The following formula holds connecting the dimension of image space and null-space

$$\dim N(A) + \dim A(U) = \dim U$$

In particular we have

$$\dim A(U) \leq \dim U$$

with equality if A is injective (i.e. unambiguous). If A is bijective we readily see that $\dim U = \dim V$. We say that such a transformation is an *isomorphism* and that U and V are isomorphic. It can be shown that any n -dimensional (real) vector space is isomorphic with R^n . In the following we will therefore often identify an n -dimensional vector space with R^n .

It can be shown that the transformations mentioned in the previous section are linear transformations.

A.2.2 Matrices

By a *matrix* A we understand a rectangular table of numbers like

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}.$$

We will often use the abbreviated notation

$$A = (a_{ij}).$$

More specifically we call A an $m \times n$ matrix because there are m rows and n columns. If $m = 1$ then the matrix can be called a row-vector and if $n = 1$ it can be called column-vector.

The matrix one gets by interchanging rows and columns is called the *transposed* matrix of A and we denote it by A' or by A^T , i.e.

$$A' = A^T = \begin{bmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{mn} \end{bmatrix}$$

An $m \times n$ matrix is square if $n = m$. A square matrix for which $A = A'$ is called a *symmetric matrix*. The elements a_{ii} , $i = 1, \dots, n$ are called the *diagonal elements*.

An especially important matrix is the identity matrix of order n

$$I_n = I = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

A matrix which has zeroes off the diagonal is called a *diagonal matrix*. We use the notation

$$\Delta = \text{diag}(\delta_1, \dots, \delta_n) = \begin{bmatrix} \delta_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \delta_n \end{bmatrix}.$$

For given $n \times m$ matrices A and B one defines the *matrix sum*

$$A + B = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1m} + b_{1m} \\ \vdots & & \vdots \\ a_{n1} + b_{n1} & \cdots & a_{nm} + b_{nm} \end{bmatrix}.$$

Scalar multiplication is defined by

$$cA = \begin{bmatrix} ca_{11} & \cdots & ca_{1m} \\ \vdots & & \vdots \\ ca_{n1} & \cdots & ca_{nm} \end{bmatrix},$$

i.e. element-wise multiplication.

For an $m \times n$ matrix C and an $n \times p$ matrix D we define the *matrix product* $P = CD$ by having that P is a $m \times p$ matrix with the (i, j) 'th element

$$p_{ij} = \sum_{k=1}^n c_{ik}d_{kj}$$

We note that the matrix product is not commutative, i.e. that CD generally does not equal DC .

For transposition we have the following rules

$$\begin{aligned} (\mathbf{A} + \mathbf{B})' &= \mathbf{A}' + \mathbf{B}' \\ (c\mathbf{A})' &= c\mathbf{A}' \\ (\mathbf{C}\mathbf{D})' &= \mathbf{D}'\mathbf{C}' \end{aligned}$$

A.2.3 Linear transformations using matrix-formulation

It can be shown that for any linear transformation $A : R^n \rightarrow R^m$ there is a corresponding $m \times n$ matrix \mathbf{A} , such that

$$\forall \mathbf{x} \in R^n : A(\mathbf{x}) = \mathbf{A}\mathbf{x}$$

Conversely an A defined by this relation is a linear transformation. \mathbf{A} is easily found as the matrix which as columns has the coordinates of the transformation of the unit vectors in R^n . E.g. we have

$$\mathbf{A}\mathbf{e}_2 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} = \mathbf{a}_2$$

If we also have a linear transformation $B : R^m \rightarrow R^k$ with corresponding matrix \mathbf{B} ($k \times m$), then we have that $B \circ A \leftrightarrow \mathbf{B}\mathbf{A}$ i.e.

$$\forall \mathbf{x} \in R^n : B \circ A(\mathbf{x}) = B(A(\mathbf{x})) = \mathbf{B}\mathbf{A}\mathbf{x}$$

Here we note, that an $n \times n$ matrix \mathbf{A} is said to be regular if the corresponding linear transformation is bijective. This is equivalent with the existence of an *inverse matrix*, i.e. a matrix \mathbf{A}^{-1} , which satisfies

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

where \mathbf{I} is the identity matrix of order n . Furthermore we have

$$\begin{aligned} (\mathbf{A}^{-1})^{-1} &= \mathbf{A} \\ (k\mathbf{A})^{-1} &= \frac{1}{k}\mathbf{A}^{-1} \\ (\mathbf{A}')^{-1} &= (\mathbf{A}^{-1})' \end{aligned}$$

and for invertible matrices \mathbf{A}, \mathbf{B} we have

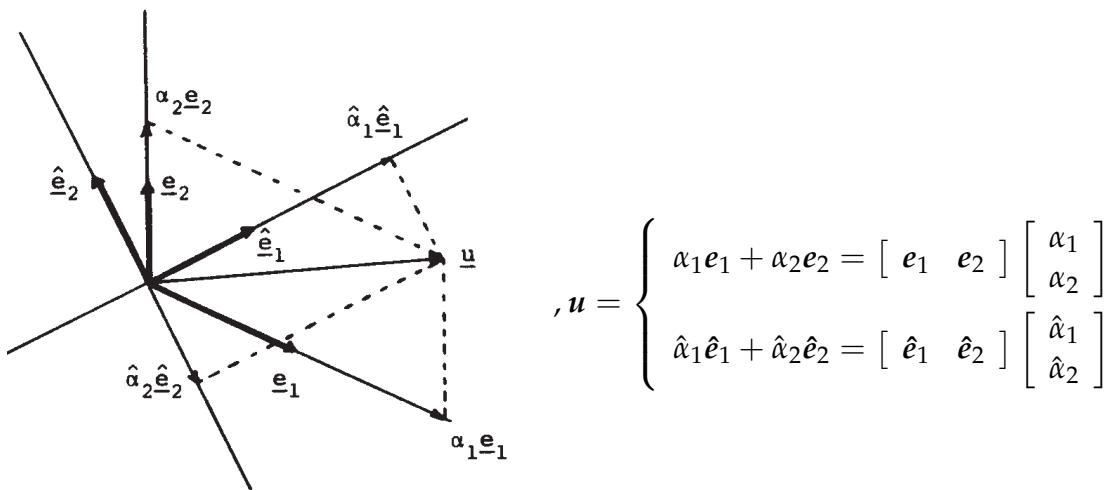


Figure A.4: Sketch of the coordinate transformation problem.

$$(AB)^{-1} = B^{-1}A^{-1}$$

A square matrix which corresponds to an idempotent transformation is itself called *idempotent*. It is readily seen that a matrix A is *idempotent* if and only if

$$AA = A$$

We note that if an idempotent matrix is regular, then it equals the identity matrix, i.e. the corresponding transformation is the identity.

A.2.4 Coordinate transformation

In this section we give formulas for the matrix formulation of a linear mapping (transformation) by going from one basis to another.

We first consider the change of coordinates going from one coordinate system to another. Normally, we choose not to distinguish between a vector u and its set of coordinates. This gives a simple notation and does not lead to confusion. However, when several coordinate systems are involved we do need to be able to make this distinction. In R^n we consider two coordinate systems (e_1, \dots, e_n) and $(\hat{e}_1, \dots, \hat{e}_n)$. The coordinates of a vector u in each of the two coordinate systems is denoted respectively $(\alpha_1, \dots, \alpha_n)'$ and $(\hat{\alpha}_1, \dots, \hat{\alpha}_n)'$, cf. figure A.4.

Let the "new" system $(\hat{e}_1, \dots, \hat{e}_n)$ be given by

$$(\hat{e}_1, \dots, \hat{e}_n) = (e_1, \dots, e_n)S$$

i.e.

$$\hat{e}_i = s_{1i}e_1 + \dots + s_{ni}e_n, \quad i = 1, \dots, n.$$

The columns in the S -matrix are thus equal to the "new" systems "old" coordinates. S is called the *coordinate transformation matrix*.

||| Remark A.5

However, many references use the expression coordinate transformation matrix about the matrix S^{-1} . It is therefore important to be sure which matrix one is talking about. Since

$$(e_1 \cdots e_n) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = (\hat{e}_1 \cdots \hat{e}_n) \begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{pmatrix},$$

(cf. fig. A.4), the connection between a vectors "old" and "new" coordinates becomes

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = S \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{bmatrix} \iff \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{bmatrix} = S^{-1} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}.$$

We now consider a linear mapping $A : R^n \rightarrow R^m$, and let A 's matrix formulation w.r.t. the bases (e_1, \dots, e_n) and (f_1, \dots, f_m) be

$$\beta = A \alpha$$

and the formulation w.r.t. the bases $(\hat{e}_1, \dots, \hat{e}_n) = (e_1, \dots, e_n)S$ and $(\hat{f}_1, \dots, \hat{f}_m) = (f_1, \dots, f_m)T$ be

$$\hat{\beta} = \hat{A} \hat{\alpha}$$

Then we have

$$\hat{A} = T^{-1} A S,$$

which is readily found by use of the rules of coordinate transformation on the coordinates.

If we are concerned with mappings $R^n \rightarrow R^n$ and we use the same coordinate transformation, then we get the relation

$$\hat{A} = S^{-1} A S.$$

The matrices A and $\hat{A} = S^{-1} A S$ are then called *similar matrices*.

A.2.5 Rank of a matrix

By rank of a linear projection $A : R^n \rightarrow R^m$ we mean the dimension of the image space, i.e.

$$\text{rk}(A) = \text{rank}(A) = \dim A(R^n).$$

By *rank* of a matrix A we mean the rank of the corresponding linear projection.

We see that $\text{rk}(A)$ exactly equals the number of linearly independent column vectors in A . Trivially we therefore have

$$\operatorname{rk}(A) \leq n.$$

If we introduce the transposed matrix A' it is easily shown that $\operatorname{rk}(A) = \operatorname{rk}(A')$ i.e. we have

$$\operatorname{rk}(A) \leq \min(m, n).$$

If A and B are two $m \times n$ matrices, then

$$\operatorname{rk}(A + B) \leq \operatorname{rk}(A) + \operatorname{rk}(B).$$

This relation is obvious when one remembers that for the corresponding projections A and B we have $(A + B)(\mathbb{R}^n) \subseteq A(\mathbb{R}^n) \cup B(\mathbb{R}^n)$.

If A is an $(m \times n)$ -matrix and B is an $(k \times m)$ -matrix we have

$$\operatorname{rk}(BA) \leq \operatorname{rk}(A).$$

If B is regular $(m \times m)$ we have

$$\operatorname{rk}(BA) = \operatorname{rk}(A).$$

These relations are immediate consequences of the relation $\dim B(A(\mathbb{R}^n)) \leq \dim(A(\mathbb{R}^n))$, where we have equality if B is injective. There are of course analogue relations for an $(n \times p)$ -matrix C :

$$\operatorname{rk}(AC) \leq \operatorname{rk}(A)$$

with equality if C is a regular $(n \times n)$ -matrix. From these we can deduce for regular B and C that

$$\operatorname{rk}(BAC) = \operatorname{rk}(A).$$

Finally we mention that an $(n \times n)$ -matrix A is regular if $\operatorname{rk}(A) = n$.

A.2.6 Determinant of a matrix

The abstract definition of the determinant of a square $p \times p$ matrix A is

$$\det(A) = \sum_{\text{alle } \sigma} \pm a_{1\sigma(1)} \cdots a_{p\sigma(p)},$$

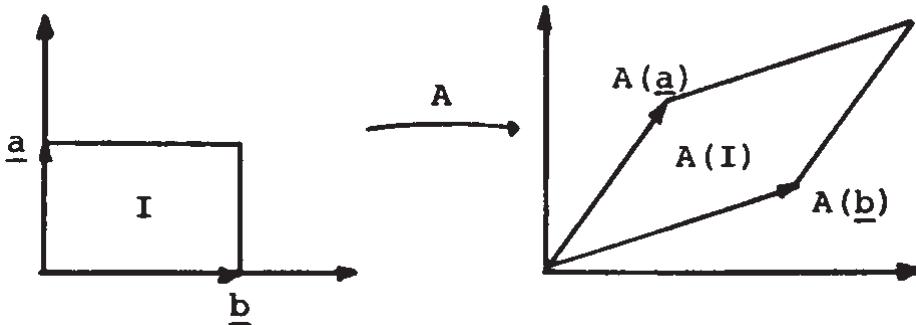


Figure A.5: A rectangle and its image after a linear projection.

where σ is a permutation of the numbers $1, \dots, p$ and where we use the + sign if the permutation is even (i.e. it can be composed of an even number of neighbour swaps) and - if it is odd. If confusion with the absolute value of a real number is unlikely we sometimes use the notation $|A| = \det(A)$

We will not go into the background of this definition. We note that the determinant represents the volume ratio of the corresponding linear projection i.e. for an $(n \times n)$ -matrix A

$$|\det(A)| = \frac{\text{vol}(A(I))}{\text{vol}(I)},$$

where I is an n -dimensional box and $A(I)$ is the image of I (being an n -dimensional parallelepiped) found by the corresponding projection.

The situation is sketched in 2 dimensions in fig. A.5. For 2×2 and 3×3 matrices the definition of the determinant becomes

$$\begin{aligned} \det \begin{pmatrix} a & b \\ c & d \end{pmatrix} &= ad - bc \\ \det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} &= aei + bfg + cdh - gec - hfa - idb. \end{aligned}$$

For determinants of higher order (here n 'th order) we can develop the determinant by the i 'th row i.e.

$$\det(A) = \sum_{j=1}^n a_{ij} (-1)^{i+j} \det(A_{ij}),$$

where A_{ij} is the matrix we get after deleting the i 'th row and the j 'th column of A . The number

$$A_{ij} = (-1)^{i+j} \det(A_{ij})$$

is also called the *cofactor* of element a_{ij} . Of course an analogue procedure exists for development by columns.

When one explicitly must evaluate a determinant the following three rules are handy:

- i) interchanging 2 rows (columns) in A multiplies $\det(A)$ by -1 .
- ii) multiplying a row (column) by a scalar multiplies $\det(A)$ by the scalar.
- iii) multiplying the matrix with a scalar multiplies $\det(A)$ by the scalar raised to the power of p .
- iv) adding a multiplum of a row (column) to another row (column) leaves $\det(A)$ unchanged.

When determining the rank of a matrix it can be useful to remember that the rank is the largest number r for which the matrix has a determinant of the minor which different from 0 and of r 'th order. We find as a special case that A is regular if and only if $\det A \neq 0$. This also seems intuitively obvious when one considers the determinant being the volume. If it is 0 then the projection must in some sense "reduce the dimension".

For a square matrix A we have

$$\det(A') = \det(A)$$

For square matrices A and B we have

$$\det(AB) = \det(A)\det(B)$$

For a diagonal matrix $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ we have

$$\det(\Lambda) = \lambda_1 \dots \lambda_n$$

For a triangular matrix C with diagonal elements c_1, \dots, c_n we have

$$\det(C) = c_1 \dots c_n$$

By means of determinants one can directly state the inverse of a regular matrix A . We have

$$A^{-1} = \frac{1}{\det(A)} (A_{ij})',$$

i.e. the inverse of a regular matrix A is the transposed of the matrix we get by substituting each element in A by its cofactor divided by $\det A$. However, note that this formular is

not directly applicable for the inversion of large matrices because of the large number of computations involved in the calculation of determinants.

Something similar is true for Cramér's theorem on solving a linear system of equations: Consider the regular matrix $A = (A_1, \dots, A_n)$. Then the solution to the equation

$$A x = b$$

is given by

$$x_i = \frac{\det(a_1, \dots, a_{i-1}, b, a_{i+1}, \dots, a_n)}{\det A}$$

A.2.7 Block matrices

By a *block matrix* we mean a matrix of the form

$$B = \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & & \vdots \\ B_{m1} & \cdots & B_{mn} \end{bmatrix}$$

where the *blocks* B_{ij} are matrices of order $m_i \times n_j$. A block matrix is also called a *partitioned matrix*.

When adding and multiplying one can use the usual rules of calculation for matrices and just consider the blocks as elements. For instance we find

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} R \\ S \end{bmatrix} = \begin{bmatrix} AR + BS \\ CR + DS \end{bmatrix},$$

under the obvious condition that the involved products exist etc.

First we give a result on determinants of the "triangular" matrix.

||| Theorem A.6

Let the square matrix A be partitioned into block matrices

$$A = \begin{bmatrix} B & C \\ \mathbf{0} & D \end{bmatrix}$$

where B and D are square and $\mathbf{0}$ is a matrix only containing 0's. Then we have

$$\det(A) = \det(B) \det(D)$$

||| Proof

We have that

$$\begin{bmatrix} B & C \\ \mathbf{0} & D \end{bmatrix} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix} \begin{bmatrix} B & C \\ \mathbf{0} & I \end{bmatrix}$$

where the I 's are identity-matrices, not necessarily of same order. If one develops the first matrix by its 1st row we see that it has the same determinant as the matrix one gets by deleting the first row and column. By repeating this until the remaining minor is D , we see that

$$\det \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & D \end{bmatrix} = \det(D)$$

Analogously we find that the last matrix has the determinant $\det B$ and the result follows. ■

The following theorem expands this result.

||| Theorem A.7

Let the matrix Σ be partitioned into block matrices

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Then we have

$$\det(\Sigma) = \det(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \det(\Sigma_{22}),$$

under the condition that Σ_{22} is regular.

||| Proof

Since

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -\Sigma_{22}^{-1}\Sigma_{21} & I \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & \Sigma_{12} \\ \mathbf{0} & \Sigma_{22} \end{bmatrix},$$

the result follows immediately from the previous theorem. ■

Remark. The matrix

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

is called the *Schur complement* of the block Σ_{22} .

The last theorem gives a useful result on inversion of matrices which are partitioned into block matrices.

|||| Theorem A.8

For the symmetrical matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

we have

$$\Sigma^{-1} = \begin{bmatrix} B^{-1} & -B^{-1}A' \\ -AB^{-1} & \Sigma_{22}^{-1} + AB^{-1}A' \end{bmatrix},$$

where

$$\begin{aligned} A &= \Sigma_{22}^{-1}\Sigma_{21} \\ B &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \end{aligned}$$

conditioned on the existence of the inverses involved.

|||| Proof

The result follows immediately by multiplication of Σ and Σ^{-1} .

■

A.3 Pseudoinverse or generalised inverse matrix of a non-regular matrix

We consider a linear transformation

$$A : E \rightarrow F$$

where E is an n -dimensional and F an m -dimensional (euclidian) vector space. The matrix corresponding to A is usually called A and it has the dimensions $m \times n$. We equal the null space of A to U , i.e.

$$U = A^{-1}(\mathbf{0}),$$

and call its dimension r . The image space

$$V = A(E)$$

has dimension $s = n - r$, cf. section A.2.1.

We now consider an arbitrary s -dimensional space $U^* \subseteq E$, which is complementary to U , and an arbitrary $m - s$ -dimensional subspace $V^* \subseteq F$, which is complementary to V .

An arbitrary vector $x \in E$ can now be written as

$$x = u + u^*, \quad u \in U \quad \text{og} \quad u^* \in U^*,$$

since u and u^* are given by

$$\begin{aligned} u &= x - p_{U^*}(x) \\ u^* &= p_{U^*}(x) \end{aligned}$$

Here p_{U^*} denotes the projection of E onto U^* along the sub-space U . Similarly any $y \in F$ can be written

$$y = (y - p_V(y)) + p_V(y) = v^* + v$$

where

$$p_V : F \rightarrow V$$

is the projection of F onto V along V^* .

Since

$$A(x) = A(u + u^*) = A(u^*),$$

we see that A is constant on the side-spaces

$$u^* + U = \{u^* + u \mid u \in U\}$$

and it follows that A 's restriction on U^* is a bijective projection of U^* onto V . This projection therefore has an inverse

$$B_1 : V \rightarrow U^*$$

given by

$$B_1(v) = u^* \Leftrightarrow A(u^*) = v$$

We are now able to formulate the definition of the pseudoinverse transformation.

||| Definition A.9

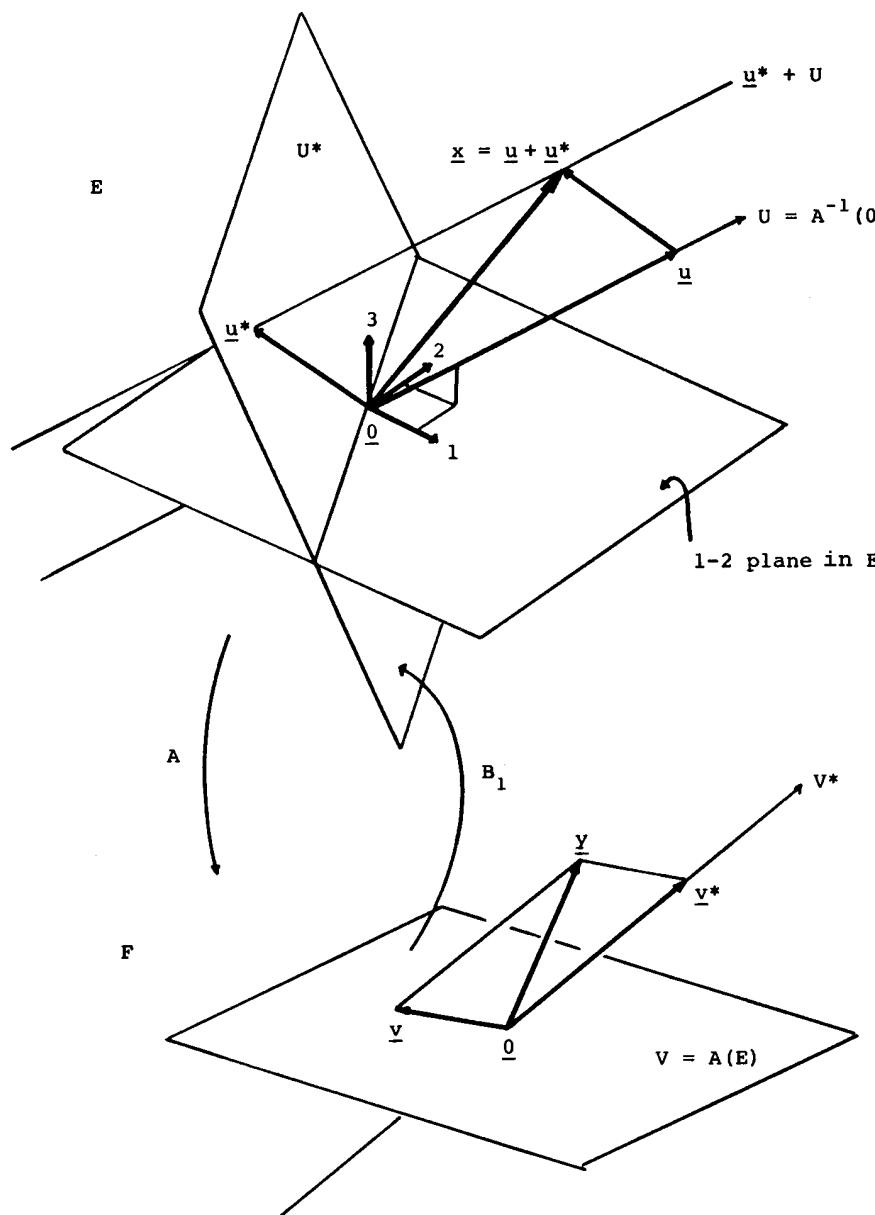
By a *pseudoinverse* or *generalised inverse* transformation of the transformation A we mean a transformation

$$B = B_1 \circ p_V : F \rightarrow E,$$

where p_V and B_1 are as mentioned previously.

||| Remark A.10

The pseudoinverse is thus the combined transformation onto V along V^* and the inverse of A 's restriction to U^* .



10

Figure A.6: Sketch showing pseudoinverse transformation.

||| Remark A.11

The pseudoinverse is of course by no means unambiguous, because we get one for each choice of the sub-spaces U^* and V^* .

We can now state some obvious properties of the pseudoinverse in the following

||| Theorem A.12

The pseudoinverse B of A has the following properties

- i) $\text{rk}(B) = \text{rk}(A) = s$
- ii) $A \circ B = p_V : F \rightarrow V$
- iii) $B \circ A = p_{U^*} : E \rightarrow U^*$

It can be shown that these properties also characterise pseudoinverse transformations,

because we have

||| Theorem A.13

Let $A : E \rightarrow F$ be linear with rank s . Assume that B also has rank s , and that $A \circ B$ and $B \circ A$ both are projections of rank s . Then B is a pseudoinverse of A as defined above.

||| Proof

Omitted (relatively simple exercise in linear algebra).

■

We now give a matrix formulation of the above mentioned definitions.

||| Definition A.14

Let A be an $(m \times n)$ -matrix of rank s . An $(n \times m)$ -matrix B , which satisfies

- i) $A B$ idempotent with rank s
- ii) $B A$ idempotent with rank s ,

is called a *pseudoinverse* or a *generalised inverse* of A .

By means of the pseudoinverse we can characterise the set of possible solutions of a system of linear equations. This is due to the following

||| Theorem A.15

Let A and B be as in definition A.14. The general solution of the equation

$$A x = \mathbf{0}$$

is

$$(I - BA)z, \quad z \in R^n,$$

and the general solution of the equation (which is assumed to be consistent)

$$A x = y,$$

is

$$B y + (I - BA)z, \quad z \in R^n.$$

||| Proof

We first consider the homogeneous equation. A solution x is obviously a point in the null-space $N(A) = A^{-1}(\mathbf{0})$ of the linear projection corresponding to A . The matrix BA according to theorem A.6 - corresponds precisely to the projection onto U^* . Therefore $I - BA$ corresponds to the projection onto the null-space $U = N(A)$. Therefore, an arbitrary $x \in N(A)$ can be written

$$x = (I - BA)z, \quad z \in R^n.$$

The statement regarding the homogeneous equation has now been proved.

The equation $A x = y$ only has a solution (i.e. is only consistent) if y lies in the image space of A . For such a y we have

$$ABy = y,$$

according to theorem A.12.

The result for the complete solution follows readily.

■

In order to illustrate the concept we now give

||| Example A.16

We consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

A obviously has the rank 2.

We will consider the linear projection corresponding to A which is

$$A : E \rightarrow F$$

where E and F are 3-dimensional vector spaces with bases $\{e_1, e_2, e_3\}$ og $\{f_1, f_2, f_3\}$. The coordinates of these bases are denoted by small x 's and y 's respectively, such that A can be formulated in the coordinates

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

First we will determine the null-space

$$U = N(A) = A^{-1}(\mathbf{0})$$

for A . We have

$$\begin{aligned} x \in U &\Leftrightarrow Ax = \mathbf{0} \\ &\Leftrightarrow x_1 + x_2 + 2x_3 = 0 \quad \wedge \quad 2x_1 + x_2 + x_3 = 0 \\ &\Leftrightarrow x_1 = x_3 \quad \wedge \quad -3x_1 = x_2 \\ &\Leftrightarrow x' = x_1(1, -3, 1). \end{aligned}$$

The null-space is then

$$U = \left\{ t \cdot \begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix} \mid t \in R \right\} = \{t \cdot u_3 \mid t \in R\}$$

As complementary sub-space we choose to consider the orthogonal complement U^* . This has the equation

$$(1, -3, 1)x = 0,$$

or

$$U^* = \{x \mid x_1 - 3x_2 + x_3 = 0\}$$

We now consider a new basis for E , namely $\{u_1, u_2, u_3\}$. Coordinates in this are denoted using small z 's. The conversion from z -coordinates to x -coordinates is given by

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

or

$$x = Sz.$$

The columns of the S matrix are known to be the u 's coordinates in the e -system.

A 's image space V is 2-dimensional and is spanned by A 's columns. We can for instance choose the first two, i.e.

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

As complementary sub-space V^* we choose V 's orthogonal complement. This is produced by making the cross-product of \mathbf{v}_1 and \mathbf{v}_2 :

$$\mathbf{v}_1 \times \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \mathbf{v}_3$$

We now consider the new basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ for F . The coordinates in this are denoted using small w 's. The conversion from w -coordinates to y -coordinates is given by

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix},$$

or in compact notation

$$\mathbf{y} = \mathbf{T} \mathbf{w}.$$

We will now find coordinate expressions for A in z - and w -coordinates. Since

$$\mathbf{y} = \mathbf{A} \mathbf{x}$$

we have

$$\mathbf{T} \mathbf{w} = \mathbf{A} \mathbf{S} \mathbf{z}$$

or

$$\mathbf{w} = \mathbf{T}^{-1} \mathbf{A} \mathbf{S} \mathbf{z}.$$

Now we have

$$\mathbf{T}^{-1} = \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix},$$

wherefore

$$\begin{aligned} \mathbf{T}^{-1} \mathbf{A} \mathbf{S} &= \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 0 \\ -3 & 11 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Since $\{\mathbf{u}_1, \mathbf{u}_2\}$ spans U^* and $\{\mathbf{v}_1, \mathbf{v}_2\}$ spans V , we note that the condition

$$A : U^* \rightarrow V$$

has the coordinate expression

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ -3 & 11 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

It has the inverse projection

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{3}{22} & \frac{1}{11} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

If we consider the points as points in E and F - and not just as points in U^* and V then we get

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{3}{22} & \frac{1}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \quad (\text{A-2})$$

The projection of F onto V along V^* has the formulation in coordinates

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \rightarrow \begin{bmatrix} w_1 \\ w_2 \\ 0 \end{bmatrix} \quad (\text{A-3})$$

This is the $z - w$ coordinate formulation for the pseudoinverse B of the projection A . However, we want a description in $x - y$ coordinates. Since

$$z = S^{-1}x = Cw = CT^{-1}y$$

we get

$$x = SCT^{-1}y,$$

where C is the matrix in formula A-1.

We therefore have

$$\begin{aligned} B &= SCT^{-1} \\ &= \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{3}{22} & \frac{1}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= \frac{1}{22} \begin{bmatrix} -8 & 7 & 7 \\ 2 & 1 & 1 \\ 14 & -4 & -4 \end{bmatrix} \end{aligned}$$

This matrix is a pseudoinverse of A .

As it is seen from the previous example it is rather tedious just to use the definition in order to calculate a pseudoinverse. Often one may utilise the following

||| Theorem A.17

Let the $m \times n$ matrix A have rank s and let

$$A = \begin{bmatrix} C & D \\ E & F \end{bmatrix},$$

where C is regular with dimension $s \times s$. A (possible) pseudoinverse of A is then

$$A^- = \begin{bmatrix} C^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where the 0-matrices have dimensions such that A^- has the dimension $n \times m$.

||| Proof

We have

$$AA^{-}A = \begin{bmatrix} C & D \\ E & F \end{bmatrix} \begin{bmatrix} C^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} C & D \\ E & F \end{bmatrix} = \begin{bmatrix} C & D \\ E & EC^{-1}D \end{bmatrix}.$$

Since $\text{rk}(A) = s$, then the last $n - s$ columns can be written as linear combinations of the first s columns, i.e. there exists a matrix H , so

$$\begin{bmatrix} D \\ F \end{bmatrix} = \begin{bmatrix} C \\ E \end{bmatrix} H$$

or

$$\begin{aligned} D &= CH \\ F &= EH \end{aligned}$$

From this we find

$$F = EC^{-1}D.$$

If we insert this in the top formula we have

$$AA^{-}A = A$$

By pre-multiplication with A^{-} and post-multiplication with A^{-} respectively, we see that $A^{-}A$ and AA^{-} are idempotent. The theorem is now derived from the definition page 353. ■

We illustrate the use of the theorem in the following

||| Example A.18

We consider the matrix given in example A.16

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

Since

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} -1 & 1 \\ 2 & -1 \end{bmatrix},$$

we can use as pseudoinverse:

$$A^{-} = \begin{bmatrix} -1 & 1 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

The advantage of using the procedure given in example A.16 instead of the far more simple one given in example A.18, is that one obtains a precise geometrical description of the situation.

||| Remark A.19

Finally, we note that the literature has a number of definitions of pseudoinverses and generalised inverses, so it is necessary to specify exactly what the definition is. A case of special interest is the so-called *Moore-Penrose inverse* A^+ of a matrix A . It satisfies the following

- i) $A A^+ A = A$
- ii) $A^+ A A^+ = A^+$
- iii) $(A A^+)' = A A^+$
- iv) $(A^+ A)' = A^+ A$

Many authors reserve the name pseudoinverse to the Moore-Penrose inverse. It is obvious that condition i) is equivalent to the general conditions for being a generalised inverse. A matrix that satisfies i) and ii) is called a *g2 inverse*. This is often used in estimation in the so-called General Linear Model. All 4 conditions guarantee that a least squares solution of an inconsistent equation find a solution with minimal norm. We will not pursue this further here, only refer the interested reader to the literature e.g. [23].

A.4 Eigenvalue problems. Quadratic forms

We begin with the fundamental definitions and theorems in

A.4.1 Eigenvalues and eigenvectors for symmetric matrices

The definition of an eigenvector and an eigenvalue given below are valid for arbitrary square matrices. However, in the sequel we will always assume the involved matrices are symmetrical unless explicitly stated otherwise.

An *eigenvalue* λ of the symmetric $n \times n$ matrix A is a solution to the equation

$$\det(A - \lambda I) = 0.$$

There are n (real-valued) eigenvalues (some may have equal values). If λ is an eigenvalue, then vectors $x \neq 0$, exist such that

$$A x = \lambda x,$$

i.e. vector exist such that the linear projection corresponding to A leads to a multiplum of its self. Such vectors are called *eigenvectors* corresponding to the eigenvalue λ . The number

of eigenvalues different from 0 equals $\text{rk}(A)$. An eigenvalue is to be counted as many times as its multiplicity indicates. A more interesting theorem is

||| Theorem A.20

If λ_i and λ_j are different eigenvalues, and if x_i and x_j are the corresponding eigenvectors, then x_i and x_j are orthogonal, i.e. $x_i'x_j = 0$.

||| Proof

We have

$$\begin{aligned} A x_i &= \lambda_i x_i \\ A x_j &= \lambda_j x_j \end{aligned}$$

Here we readily find

$$\begin{aligned} x_j' A x_i &= \lambda_i x_j' x_i \\ x_i' A x_j &= \lambda_j x_i' x_j. \end{aligned}$$

We transpose the first relationship and get

$$x_i' A' x_j = \lambda_i x_i' x_j.$$

Since A is symmetric this implies that

$$\lambda_i x_i' x_j = \lambda_j x_i' x_j,$$

and since $\lambda_i \neq \lambda_j$ then $x_i' x_j = 0$ i.e. $x_i \perp x_j$.

■

The result in theorem A.20 can be supplemented with the following theorem given without proof.

||| Theorem A.21

If λ is an eigenvalue with multiplicity m , then the set of eigenvectors corresponding to λ forms an m -dimensional sub-space. This has the special implication that there exists m orthogonal eigenvectors corresponding to λ .

By combining these two theorems one readily sees the following

||| Corollary A.22

For an arbitrary symmetric matrix A a basis exists for R^n consisting of mutually orthogonal eigenvectors of A .

If such a basis consisting of orthogonal eigenvectors is normed then one gets an *orthonormal basis* (p_1, \dots, p_n) . If we let P equal the $n \times n$ matrix whos columns are the coordinates of these vectors, i.e.

$$P = (p_1, \dots, p_n)$$

we get

$$P'P = I$$

P is therefore by definition an *orthogonal matrix*, and

$$AP = P\Lambda$$

where Λ is a diagonal matrix with the eigenvalues for A (repeated corresponding to multiplicity) on the diagonal. By means of this we get the following

||| Theorem A.23

Let A be a symmetric matrix. Then an orthogonal matrix P exists, such that

$$P'AP = \Lambda$$

where Λ is a diagonal matrix with A 's eigenvalues on the diagonal (repeated corresponding to the multiplicity). As P one can choose a matrix, whos columns are orthonormed eigenvectors of A .

||| Proof

Obvious from the above relation.

■

||| Theorem A.24

Let A be a symmetric matrix with non-negative eigenvalues. Then a regular matrix B exists such that

$$B'B = E,$$

where E is a diagonal matrix having 0's or 1's on the diagonal. The number of 1's equals $\text{rk}(A)$. If A is of full rank then E becomes an identity matrix.

||| Proof

By (post-) multiplication of \mathbf{P} with a diagonal matrix \mathbf{C} which has the following diagonal elements

$$c_i = \begin{cases} \frac{1}{\sqrt{\lambda_i}} & \lambda_i > 0 \\ 1 & \lambda_i = 0 \end{cases},$$

we readily find the theorem with $\mathbf{B} = \mathbf{P} \mathbf{C}$. ■

The relation in theorem A.23 is equivalent to

$$\mathbf{A} = \mathbf{P} \Lambda \mathbf{P}'$$

or

$$\mathbf{A} = (\mathbf{p}_1 \dots \mathbf{p}_n) \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{p}'_1 \\ \vdots \\ \mathbf{p}'_n \end{bmatrix},$$

i.e. we have the following partitioning of the matrix

$$\mathbf{A} = \lambda_1 \mathbf{p}_1 \mathbf{p}'_1 + \cdots + \lambda_n \mathbf{p}_n \mathbf{p}'_n.$$

This partitioning of the symmetrical matrix \mathbf{A} is often called its *spectral decomposition*, since the eigenvalues $\{\lambda_1, \dots, \lambda_n\}$ are called the *spectrum* of the matrix.

With the obvious definition of $\Lambda^{\frac{1}{2}}$ being $\text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})$, we note that we can write

$$\mathbf{A} = (\mathbf{P} \Lambda^{\frac{1}{2}})(\mathbf{P} \Lambda^{\frac{1}{2}})' = \mathbf{G} \mathbf{G}'.$$

Remark. Here we mention that if \mathbf{A} is positive definite, then there is a relation

$$\mathbf{A} = \mathbf{L} \mathbf{L}',$$

where \mathbf{L} is a lower triangular matrix. This relation is called the Cholesky decomposition of \mathbf{A} (see e.g. [24]). It is unique.

Finally we have

||| Theorem A.25

Let \mathbf{A} be a regular symmetrical matrix. Then \mathbf{A} and \mathbf{A}^{-1} have the same eigenvectors corresponding to reciprocal eigenvalues.

||| Proof

Let λ be an eigenvalue of A and \underline{x} be a corresponding eigenvector, i.e.

$$A\underline{x} = \lambda\underline{x}.$$

Since A is regular then this is equivalent to

$$A^{-1}\underline{x} = \frac{1}{\lambda}\underline{x},$$

which concludes the proof. ■

Finally, we note that

$$\det A = \prod_i \lambda_i.$$

||| Example A.26

Orthogonal transformations of the plane. In order to give a geometrical understanding of the transformations which reduce a symmetrical matrix into diagonal form, we state the orthogonal transformations of the plane.

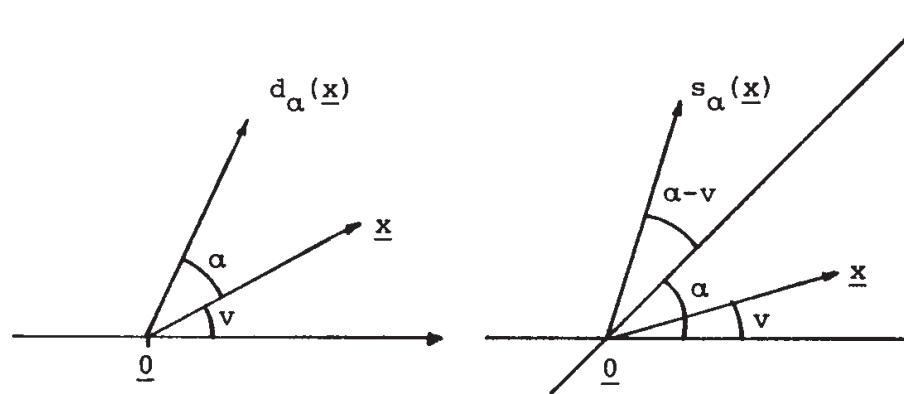
By utilising the orthogonality conditions $P'P = I$ we readily see, that the only orthogonal 2×2 -matrices are matrices of the form

$$\begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \quad \text{og} \quad \begin{bmatrix} \cos \alpha & \sin \alpha \\ \sin \alpha & -\cos \alpha \end{bmatrix}.$$

We will now show that these correspond to *rotations* around the origin and *reflections* in straight lines.

We do this by determining coordinate expressions for the linear transformations d_α and s_α , which respectively represent a rotation of the plane of the angle α and a reflection in the line having the angle α with the 1.st axis.

The transformations are illustrated in figure A.26.



Rotation and reflection as determined by the angle α .

Since $\mathbf{x} = r(\cos v, \sin v)'$, where r is equal to 1, we have

$$\begin{aligned} d_\alpha(\mathbf{x}) &= \begin{bmatrix} \cos(\alpha + v) \\ \sin(\alpha + v) \end{bmatrix} = \begin{bmatrix} \cos \alpha \cos v - \sin \alpha \sin v \\ \sin \alpha \cos v + \cos \alpha \sin v \end{bmatrix} \\ &= \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}. \end{aligned}$$

From this we find d_α has the matrix representation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Analogously we find

$$\begin{aligned} s_\alpha(\mathbf{x}) &= \begin{bmatrix} \cos(2\alpha - v) \\ \sin(2\alpha - v) \end{bmatrix} = \begin{bmatrix} \cos 2\alpha \cos v + \sin 2\alpha \sin v \\ \sin 2\alpha \cos v - \cos 2\alpha \sin v \end{bmatrix} \\ &= \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}. \end{aligned}$$

so that s_α has the matrix representation

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \rightarrow \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This concludes the proof of the introductory statement.

It is often useful to have the following relations between rotations and reflections of the plane in mind

$$\begin{aligned} s_{\frac{\pi}{4}} \circ d_\alpha &= s_{\frac{\pi}{4} - \frac{\alpha}{2}} \\ s_\alpha &= s_{\frac{\pi}{4}} \circ d_{\frac{\pi}{2} - 2\alpha}. \end{aligned}$$

The first relation follows from

$$\begin{aligned} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} &= \\ \begin{bmatrix} \sin \alpha & \cos \alpha \\ \cos \alpha & -\sin \alpha \end{bmatrix} &= \begin{bmatrix} \cos(\frac{\pi}{4} - \alpha) & \sin(\frac{\pi}{4} - \alpha) \\ \sin(\frac{\pi}{4} - \alpha) & -\cos(\frac{\pi}{4} - \alpha) \end{bmatrix}. \end{aligned}$$

The last two relations are found from the first by substituting α with $\frac{\pi}{2} - 2\alpha$.

Part of the following section will be devoted to consider the problem of generalising the spectral decomposition of an arbitrary matrix.

A.4.2 Singular value decomposition of an arbitrary matrix. Q - and R -mode analysis

We first state the main result, also known as Eckart-Young's theorem.

||| Theorem A.27

Let x be an arbitrary $n \times p$ matrix of rank r . Then orthogonal matrices U ($p \times r$) and V ($n \times r$) exist, as do positive numbers $\gamma_1, \dots, \gamma_r$, such that

$$x = V \Gamma U' = [v_1 \cdots v_r] \begin{bmatrix} \gamma_1 & & 0 \\ & \ddots & \\ 0 & & \gamma_r \end{bmatrix} \begin{bmatrix} u'_1 \\ \vdots \\ u'_r \end{bmatrix} = \gamma_1 v_1 u'_1 + \cdots + \gamma_r v_r u'_r,$$

where $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_r)$ and v_1, \dots, v_r are the columns of V and u_1, \dots, u_r are the columns of U .

||| Proof

Omitted. See e.g. [25].

■

Remark. The numbers $\gamma_1, \dots, \gamma_r$ are called x 's *singular values*. The vectors v_1, \dots, v_r are called the *left singular vectors* of x and the vectors u_1, \dots, u_r the *right singular vectors*. The factorization of x in the theorem is called the *Singular Value Decomposition* (SVD) of x .

In the sequel we will investigate the relationship between x 's singular values and the eigenvalue problems for the symmetrical matrices $x x'$ ($n \times n$) and $x' x$ ($p \times p$).

However, first we will state

||| Theorem A.28

For an arbitrary (real valued) matrix x it holds that $x' x$ and $x x'$ have non-negative eigenvalues and

$$\text{rk}(x' x) = \text{rk}(x x') = \text{rk}(x)$$

||| Proof

It suffices to prove the results for $\mathbf{x}'\mathbf{x}$. It is obvious that $\mathbf{x}'\mathbf{x}$ is symmetric, so an orthogonal matrix \mathbf{P} , exists such that

$$\mathbf{P}'\mathbf{x}'\mathbf{x}\mathbf{P} = \Lambda$$

i.e.

$$(\mathbf{x}\mathbf{P})'(\mathbf{x}\mathbf{P}) = \Lambda.$$

By letting $\mathbf{x}\mathbf{P} = \mathbf{B} = (b_{ij})$, we find $\mathbf{B}'\mathbf{B} = \Lambda$, i.e.

$$\lambda_i = \sum_j b_{ij}^2 > 0,$$

i.e. $\mathbf{x}'\mathbf{x}$ has non-negative eigenvectors. Furthermore we see that

$$\begin{aligned} \text{rk}(\mathbf{x}'\mathbf{x}) &= \text{card}(\lambda_i \neq 0) \\ &= \text{card}\{\text{columns } \mathbf{b}_j \text{ in } \mathbf{B} \text{ , which are } \neq 0\} \end{aligned}$$

Since $\mathbf{b}_i'\mathbf{b}_j = 0$ for $i \neq j$ (due to equation A-1) we have

$$\text{rk}(\mathbf{x}'\mathbf{x}) = \text{rk}(\mathbf{B})$$

Since \mathbf{P} is regular, and using a result in section A.2.5, we find

$$\text{rk}(\mathbf{B}) = \text{rk}(\mathbf{x}\mathbf{P}) = \text{rk}(\mathbf{x}).$$

■

We state a small corollary to the theorem.

||| Corollary A.29

Let Σ be symmetrical and positive definite. Then for an arbitrary matrix \mathbf{x} it holds that

$$\text{rk}(\mathbf{x}'\Sigma^{-1}\mathbf{x}) = \text{rk}(\mathbf{x}),$$

under the condition that the involved products exist.

||| Proof

Since Σ^{-1} is also regular and positive definite, an orthogonal matrix \mathbf{P} exists, such that

$$\mathbf{P}'\Sigma^{-1}\mathbf{P} = \Lambda,$$

where Λ is a diagonal matrix. This implies

$$\Sigma^{-1} = \mathbf{P}\Lambda\mathbf{P}' = \mathbf{P}\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\mathbf{P}' = \mathbf{P}\Lambda^{\frac{1}{2}}(\mathbf{P}\Lambda^{\frac{1}{2}})' = \mathbf{B}\mathbf{B}'.$$

Here $\Lambda^{\frac{1}{2}}$ denotes the diagonal matrix, whose diagonal elements are the square roots of the corresponding elements of Λ . It is obvious that B is regular. This relation is inserted and we find

$$\mathbf{x}' \Sigma^{-1} \mathbf{x} = \mathbf{x}' B B' \mathbf{x} = (B' \mathbf{x})' B' \mathbf{x},$$

i.e.

$$\text{rk}(\mathbf{x}' \Sigma^{-1} \mathbf{x}) = \text{rk}(B' \mathbf{x}) = \text{rk}(\mathbf{x}),$$

which concludes the proof. ■

Using the notation from theorem A.28 we have.

||| Theorem A.30

The matrix $\mathbf{x}' \mathbf{x}$ ($n \times n$) has r positive eigenvalues and $n - r$ eigenvalues equal to 0. The positive eigenvalues are $\gamma_1^2, \dots, \gamma_r^2$, where $\gamma_1, \dots, \gamma_r$ are the singular values of \mathbf{x} . The corresponding eigenvectors are v_1, \dots, v_r .

Similarly $\mathbf{x}' \mathbf{x}$ ($p \times p$) has r positive and $(p - r)$ 0-eigenvalues. The positive eigenvalues are $\gamma_1^2, \dots, \gamma_r^2$ and the corresponding eigenvectors are u_1, \dots, u_r .

The positive eigenvalues of $\mathbf{x}' \mathbf{x}$ and $\mathbf{x}' \mathbf{x}$ are therefore equal and the relationship between the corresponding eigenvectors is ($m = 1, \dots, r$)

$$v_m = \frac{1}{\gamma_m} \mathbf{x} u_m \quad \text{and} \quad u_m = \frac{1}{\gamma_m} \mathbf{x}' v_m,$$

or in a more compact notation

$$V = \mathbf{x} U \Gamma^{-1} \quad \text{og} \quad U = \mathbf{x}' V \Gamma^{-1}$$

||| Proof

Follows by use of Eckart-Young's theorem. ■

||| Remark A.31

Analysis of the matrix $\mathbf{x}' \mathbf{x}$ is called R -mode analysis and the analysis of $\mathbf{x}' \mathbf{x}$ is called Q -mode analysis. These names originate from factor analysis, cf. chapter 6.

||| Remark A.32

The theorem implies that one can find the results for an R-mode analysis from a Q-mode analysis ad vice versa. For practical use one should therefore consider which of the matrices $\mathbf{x}'\mathbf{x}$ and $\mathbf{x}\mathbf{x}'$ has lowest order.

A.4.3 Quadratic forms and positive semi-definite matrices

In this section we still consider symmetrical matrices only.

By the *quadratic form* corresponding to the symmetrical matrix A we mean the mapping

$$\mathbf{x} \rightarrow \mathbf{x}'A\mathbf{x} = \sum a_{ii}x_i^2 + 2 \sum_{1 < j} a_{ij}x_i x_j.$$

We say that a symmetrical matrix A is *positive definite* respectively *positive semi-definite* if the corresponding quadratic form is positive respectively non-negative for vectors different from the 0-vector, i.e. if

$$\forall \mathbf{x} \neq \mathbf{0} : \mathbf{x}'A\mathbf{x} > 0,$$

respectively

$$\forall \mathbf{x} \neq \mathbf{0} : \mathbf{x}'A\mathbf{x} \geq 0.$$

We then also say the quadratic form is *positive definite* respectively *positive semi-definite*.

We have the following

||| Theorem A.33

The symmetrical matrix A is positive definite respectively semi-definite, if all A 's eigenvalues are positive respectively non-negative.

||| Proof

With P as in theorem A.23 we have

$$\begin{aligned} \mathbf{x}'A\mathbf{x} &= \mathbf{x}'P'PAPPP'\mathbf{x} = (\mathbf{P}'\mathbf{x})'\Lambda(\mathbf{P}'\mathbf{x}) \\ &= \mathbf{y}'\Lambda\mathbf{y} = \lambda_1y_1^2 + \cdots + \lambda_ny_n^2. \end{aligned}$$

Another useful result is

|||| Theorem A.34

A symmetrical $n \times n$ matrix A is positive definite if the determinants of all *principal minors*

$$d_i = \det \begin{bmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} \end{bmatrix}, \quad i = 1, \dots, n,$$

are positive.

|||| Proof

Omitted

■

We now state a very important theorem on extrema of quadratic forms

|||| Theorem A.35

If we let the eigenvalues for the symmetrical matrix A equal $\lambda_1 \geq \dots \geq \lambda_n$ with corresponding eigenvectors p_1, \dots, p_n , and we define

$$R(x) = \frac{x'Ax}{x'x},$$

and

$$M_k = \{x | x'p_i = 0, \quad i = 1, \dots, k-1\}.$$

Then it holds that

$$\begin{aligned} \sup_x R(x) &= R(p_1) = \lambda_1, \\ \inf_x R(x) &= R(p_n) = \lambda_n, \\ \sup_{x \in M_k} R(x) &= R(p_k) = \lambda_k. \end{aligned}$$

|||| Proof

An arbitrary vector x can be written

$$x = \alpha_1 p_1 + \dots + \alpha_n p_n.$$

If $p_i'x = 0, i = 1, \dots, k-1$, we find $\alpha_1 = \dots = \alpha_{k-1} = 0$, i.e.

$$x = \alpha_k p_k + \dots + \alpha_n p_n.$$

Therefore we have

$$x'Ax = \alpha_k^2 \lambda_k + \dots + \alpha_n^2 \lambda_n,$$

and

$$R(x) = \frac{x'Ax}{x'x} = \frac{\alpha_k^2 \lambda_k + \dots + \alpha_n^2 \lambda_n}{\alpha_k^2 + \dots + \alpha_n^2}$$

It is obvious that this expression is maximal for

$$(\alpha_k, \dots, \alpha_n) = (\alpha_k, 0, \dots, 0),$$

where it takes the value λ_k . The result with inf is proved analogously. ■

||| Remark A.36

The theorem say for $k = 1$, that the unit vector, i.e. the "direction", for which the quadratic form takes its maximal value, is the eigenvector corresponding to the largest eigenvalue. If we only consider the quadratic form in unit vectors which are orthogonal to eigenvectors corresponding to the $k-1$ largest eigenvalues, then the theorem says that maximum is in the direction corresponding to the eigenvector which corresponds to the k 'th largest eigenvalue.

||| Remark A.37

$R(x)$ is also called *Rayleigh's coefficient* or *quotient*.

We will now describe the level sets for positive definite forms.

||| Theorem A.38

Let A be positive definite. Then the set of solutions for the equation

$$x'Ax = c, \quad c > 0,$$

is an ellipsoid with principle axes in the directions of the eigenvectors. The first principle axis corresponds with the smallest eigenvalue, the second to the second smallest eigenvalue etc.

||| Proof

We consider the matrix $P = (p_1, \dots, p_n)$, whose columns are the coordinates of orthonormal eigenvectors of A . Assuming $y = P'x$ the following holds

$$\begin{aligned} x'Ax &= y'\Lambda y \\ &= \lambda_1 y_1^2 + \dots + \lambda_n y_n^2 \\ &= \frac{y_1^2}{(1/\sqrt{\lambda_1})^2} + \dots + \frac{y_n^2}{(1/\sqrt{\lambda_n})^2} \end{aligned} \quad (\text{A-4})$$

The matrix equation

$$y = P'x \Leftrightarrow x = P y$$

corresponds to a change of basis from the original orthonormal basis $\{e_1, \dots, e_n\}$ to the orthonormal basis $\{p_1, \dots, p_n\}$.

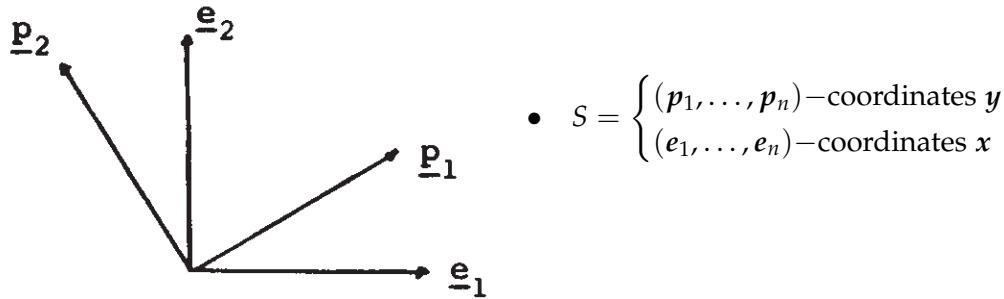


Illustration showing change of basis

This is seen by letting S be a point whose $\{e_1, \dots, e_n\}$ -coordinates are called x and whose $\{p_1, \dots, p_n\}$ -coordinates are called y . Then it holds that

$$x_1 e_1 + \dots + x_n e_n = y_1 p_1 + \dots + y_n p_n,$$

or

$$(e_1 \cdots e_n)x = (p_1 \cdots p_n)y,$$

i.e.

$$I x = P y,$$

where I is a unit matrix.

The expression in A-4 therefore shows the equation of the set of solutions in y -coordinates corresponding to the coordinate system consisting of orthonormal eigenvectors. This shows that we are dealing with an ellipsoid. The rest of the theorem now follows by noting that the 1st principle axis corresponds to the y_i , for which $1/\sqrt{\lambda_i}$ is maximal, i.e. for which λ_i is minimal.

■

||| Remark A.39

If the matrix is only positive semi-definite then the set of solutions to the equation correspond to an elliptical cylinder. This can be seen by change of base to the base $\{p_1, \dots, p_n\}$ consisting of orthonormal eigenvectors, where we for simplicity assume that p_1, \dots, p_r corresponds to the eigenvalues which are different from 0. We then have

$$\begin{aligned} \mathbf{x}' \mathbf{A} \mathbf{x} = c &\Leftrightarrow \lambda_1 y_1^2 + \dots + \lambda_r y_r^2 + 0y_{r+1}^2 + \dots + 0y_n^2 = c \\ &\Leftrightarrow \lambda_1 y_1^2 + \dots + \lambda_r y_r^2 = c. \end{aligned}$$

This leads to the statement. If we consider the restriction of the quadratic form to the subspace spanned by the eigenvectors corresponding to eigenvectors > 0 , then the set of solutions becomes an ellipsoid.

||| Example A.40

We consider the symmetrical positive definite matrix

$$\mathbf{A} = \begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix}.$$

The quadratic form corresponding to \mathbf{A} is

$$\mathbf{x}' \mathbf{A} \mathbf{x} = 3x_1^2 + 2x_2^2 + 2\sqrt{2}x_1x_2,$$

so the unit ellipse corresponding to \mathbf{A} is the set of solutions to the equation

$$3x_1^2 + 2x_2^2 + 2\sqrt{2}x_1x_2 = 1.$$

In order to determine the principle axes we determine \mathbf{A} 's eigenvalues. We find

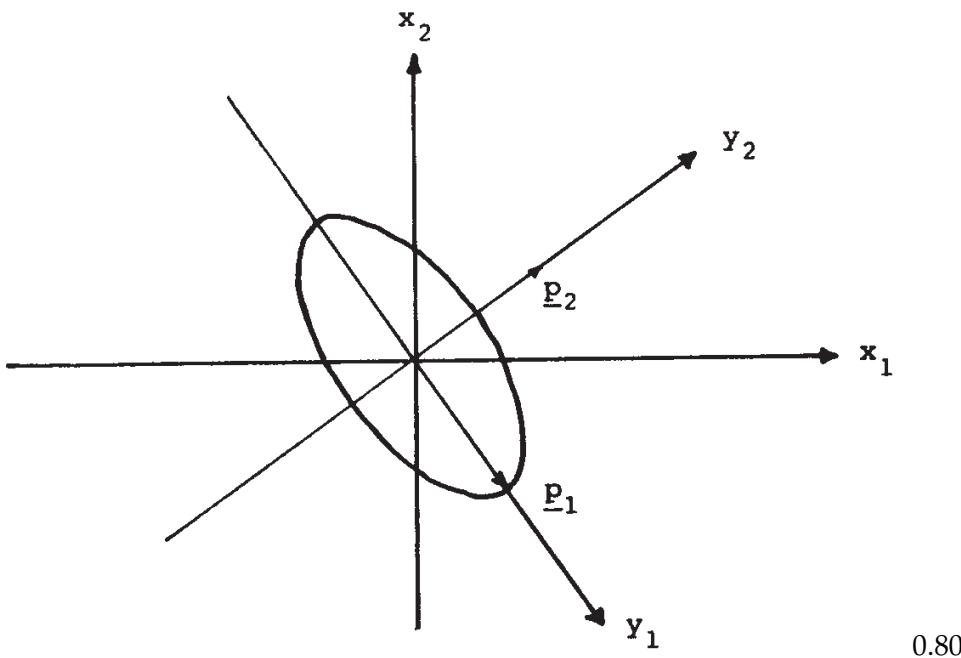
$$\begin{aligned} (\mathbf{A} - \lambda \mathbf{I}) = 0 &\Leftrightarrow \lambda^2 - 5\lambda + 4 = 0 \\ &\Leftrightarrow \lambda = 1 \quad \vee \quad \lambda = 4. \end{aligned}$$

Eigen vectors corresponding to $\lambda = 1$ respectively $\lambda = 4$ are seen to be of the form $t(1, -\sqrt{2})$ respectively $t(1, \sqrt{2}/2)$. We norm these and get

$$\mathbf{p}_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{6}}{3} \end{bmatrix}, \quad \mathbf{p}_2 = \begin{bmatrix} \frac{\sqrt{6}}{3} \\ -\frac{\sqrt{3}}{3} \end{bmatrix}.$$

If we choose the base $\{\mathbf{p}_1, \mathbf{p}_2\}$, then the coordinate representation of the quadratic form becomes

$$\mathbf{y} \rightarrow y_1^2 + 4y_2^2,$$



Ellipse determined by the quadratic form given in example A.40. The ellipse has the equation

$$\frac{y_1^2}{1^2} + \frac{y_2^2}{\frac{1}{2}^2} = 1.$$

It is illustrated in figure A.40

Since

$$\begin{aligned} p_1 &= \begin{bmatrix} \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{6}}{3} \end{bmatrix} = \begin{bmatrix} 0.577 \\ -0.820 \end{bmatrix} \\ &\simeq \begin{bmatrix} \cos(-54.7^\circ) \\ \sin(-54.7^\circ) \end{bmatrix}, \end{aligned}$$

the new coordinate system corresponds to a rotation of the old one with the angle -54.7° .

A.4.4 The general eigenvalue problem for symmetrical matrices

For use with the theory of canonical correlations and in discriminant analysis we will need a slightly more general concept of eigenvalues than seen in the previous sections. We introduce the concept in

||| Definition A.41

Let A and B be real-valued $m \times m$ symmetrical matrices and let B be of full rank. A number λ , for which

$$\det(A - \lambda B) = 0,$$

is termed an *eigenvalue of A w.r.t. B* . For such a λ it is possible to find an $x \neq 0$ such that

$$Ax = \lambda Bx.$$

Such a vector x is called an *eigenvector for A w.r.t. B* .

||| Remark A.42

The concepts given above can be traced back to eigenvalues and eigenvectors for the **non-symmetrical** matrix $B^{-1}A$.

||| Theorem A.43

We consider again the situation in the definition A.41 and further let B be positive definite. There are then m real eigenvalues of A w.r.t. B . If A is positive semi-definite, then these will be non-negative and if A is positive definite then they will be positive.

||| Proof

According to theorem A.24 there is a regular matrix T where

$$T'B T = I.$$

Let

$$D = T'A T$$

D is obviously symmetrical, and since

$$x'Dx = (Tx)'A(Tx),$$

we see that D and A are at the same time respectively positive semi-definite and positive definite.

Now we have

$$\begin{aligned} (D - \lambda I)v = 0 &\Leftrightarrow (T'A T - \lambda T'B T)v = 0 \\ &\Leftrightarrow (A - \lambda B)(Tv) = 0 \end{aligned}$$

From this we deduce that D 's eigenvalues equal A 's eigenvalues w.r.t. B , and that the eigenvectors of A w.r.t. B are found by using the transformation T on D 's eigenvectors. The result regarding the sign of the eigenvalues follows trivially. ■

||| Theorem A.44

Let the situation be as above. Then a basis exists for R^m consisting of eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_m$ of \mathbf{A} w.r.t. \mathbf{B} . These vectors can be chosen as conjugated vectors both w.r.t. \mathbf{A} as well as w.r.t. \mathbf{B} , i.e.

$$\mathbf{u}'_i \mathbf{A} \mathbf{u}_j = \mathbf{u}'_i \mathbf{B} \mathbf{u}_j = 0.$$

||| Proof

Follows from the proof of the above theorem and of the corollary to theorem A.21, remembering that

$$0 = \mathbf{v}'_i \mathbf{v}_j = (\mathbf{v}'_i \mathbf{T}') \mathbf{T}'^{-1} \mathbf{T}^{-1}(\mathbf{T} \mathbf{v}_j) = \mathbf{u}'_i \mathbf{B} \mathbf{u}_j,$$

where $\mathbf{v}_1, \dots, \mathbf{v}_m$ is an orthonormal basis for R^m consisting of eigenvectors of \mathbf{D} .

Finally we have

$$\mathbf{u}'_i \mathbf{A} \mathbf{u}_j = \lambda_j \mathbf{u}'_i \mathbf{B} \mathbf{u}_j = 0$$

■

||| Theorem A.45

Let \mathbf{A} be symmetrical and let \mathbf{B} be positive definite. Then a regular matrix \mathbf{R} exists with

$$\mathbf{R}' \mathbf{A} \mathbf{R} = \mathbf{\Lambda} = \mathbf{diag}(\lambda_1, \dots, \lambda_n),$$

and

$$\mathbf{R}' \mathbf{B} \mathbf{R} = \mathbf{I},$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of \mathbf{A} w.r.t. \mathbf{B} . If the i 'th column in \mathbf{R}'^{-1} is termed \mathbf{s}_i then these relations can be written

$$\mathbf{A} = \lambda_1 \mathbf{s}_1 \mathbf{s}'_1 + \dots + \lambda_m \mathbf{s}_m \mathbf{s}'_m,$$

and

$$\mathbf{B} = \mathbf{s}_1 \mathbf{s}'_1 + \dots + \mathbf{s}_m \mathbf{s}'_m.$$

||| Proof

From the proof of theorem A.43 we consider the $\mathbf{D} = \mathbf{T}' \mathbf{A} \mathbf{T}$. Since \mathbf{D} is symmetrical, according to theorem A.23 there exists an orthogonal matrix \mathbf{C} with

$$\mathbf{C}' \mathbf{D} \mathbf{C} = \mathbf{\Lambda},$$

because we have that D 's eigenvalues are A 's eigenvalues w.r.t. B .

If we choose $R = T C$, then we have that

$$R' B R = C' T' B T C = C' C = I,$$

and

$$R' A R = C' T' A T C = C' D C = \Lambda.$$

■

Finally we state an analogue of theorem A.35 in the following

||| Theorem A.46

Let A be positive semi-definite and let B be positive definite. Let A 's eigenvalues w.r.t. B be $\lambda_1 \geq \dots \geq \lambda_m$ and let v_1, \dots, v_m denote a basis for \mathbb{R}^m consisting of the corresponding eigenvectors with $v_i' B v_j = 0 \quad i \neq j$. We define the general Rayleigh quotient

$$R(x) = \frac{x' A x}{x' B x}$$

and put

$$M_k = \{x | x' B v_1 = \dots = x' B v_{k-1} = 0\}.$$

Then we obtain

$$\begin{aligned} \sup_x R(x) &= R(v_1) = \lambda_1 \\ \inf_x R(x) &= R(v_m) = \lambda_m \\ \sup_{x \in M_k} R(x) &= R(v_k) = \lambda_k. \end{aligned}$$

||| Proof

Without loss of generality the v_i 's can be chosen so that $v_i' B v_i = 1$, and since an arbitrary vector x can be written

$$x = \alpha_1 v_1 + \dots + \alpha_m v_m,$$

we find

$$R(x) = \frac{\sum \alpha_i^2 v_i' A v_i}{\sum \alpha_i^2 v_i' B v_i} = \frac{\sum \lambda_i \alpha_i^2}{\sum \alpha_i^2}.$$

From this the two first statements are easily seen. If $x \in M_k$, then x can be written

$$x = \alpha_k v_k + \dots + \alpha_m v_m,$$

and

$$R(\mathbf{x}) = \frac{\lambda_k \alpha_k^2 + \cdots + \lambda_m \alpha_m^2}{\alpha_1^2 + \cdots + \alpha_m^2},$$

which leads to the desired result. ■

A.4.5 The trace of a matrix

By the term *trace* of the (square) matrix \mathbf{A} we mean the sum of the diagonal elements. i.e.

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii}.$$

Obviously

$$\text{tr}(\mathbf{A}') = \text{tr}(\mathbf{A}).$$

For (square) matrices \mathbf{A} and \mathbf{B} the following holds

$$\text{tr}(\mathbf{A} \mathbf{B}) = \text{tr}(\mathbf{B} \mathbf{A}). \quad (\text{A-5})$$

Furthermore we have that the trace equals the sum of eigenvalues, i.e.

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i.$$

This follows trivially from A-5 and theorem A.23

For positive semi-definite matrices the trace is therefore another measure of "size" of a matrix. If the trace is large then at least some of the eigenvalues are large. On the other hand this measure is not sensitive to if some eigenvalues might be 0, i.e. if the matrix is degenerate. The determinant is sensitive to that, since we recall

$$\det(\mathbf{A}) = \prod_{i=1}^n \lambda_i.$$

We note further that for an idempotent matrix \mathbf{A} we have that

$$\text{tr}(\mathbf{A}) = \text{rk}(\mathbf{A}).$$

Further we have

$$\text{tr}(\mathbf{B} \mathbf{B}^-) = \text{rk}(\mathbf{B}),$$

where \mathbf{B}^- is an arbitrary pseudoinverse of \mathbf{B} .

Finally we note that for a regular matrix \mathbf{S} we have that

$$\text{tr}(\mathbf{S}^{-1}\mathbf{B}\mathbf{S}) = \text{tr}(\mathbf{B}).$$

A.4.6 Differentiation of linear form and quadratic form

Let $f : R^n \rightarrow R$. We will use the following notation for the vector of partial derivatives

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

The following theorem holds for differentiation of certain forms

|||| Theorem A.47

For a symmetrical $(n \times n)$ -matrix \mathbf{A} and an arbitrary n -dimensional vector \mathbf{b} it holds that

- i) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}'\mathbf{x}) = \mathbf{b}$
- ii) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{x}) = 2\mathbf{x}$
- iii) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2\mathbf{A}\mathbf{x}$.

|||| Proof

The proof of i) and ii) are trivial. iii) is (strangely) proved most easily by means of the definition. For an arbitrary vector \mathbf{h} we have that

$$(\mathbf{x} + \mathbf{h})'\mathbf{A}(\mathbf{x} + \mathbf{h}) = \mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{h}'\mathbf{A}\mathbf{h} + 2\mathbf{h}'\mathbf{A}\mathbf{x}$$

By choosing $\mathbf{h} = (0, \dots, h, \dots, 0)'$ we see that

$$\frac{\partial}{\partial x_i}(\mathbf{x}'\mathbf{A}\mathbf{x}) = 2 \sum_{j=1}^h a_{ij}x_j,$$

and the result follows readily.

■

We will illustrate the use of the theorem in the following

||| Example A.48

We want to find the minimum of the function

$$g(\theta) = (\mathbf{y} - \mathbf{A}\theta)' \mathbf{B} (\mathbf{y} - \mathbf{A}\theta),$$

where \mathbf{y} , \mathbf{A} and \mathbf{B} are given and \mathbf{B} is further positive semidefinite (and symmetrical). Since $g(\theta)$ is convex (a paraboloid, possibly degenerate), then the point corresponding to the minimum is found by solving the equation

$$\frac{\partial}{\partial \theta} g(\theta) = \mathbf{0}.$$

First we rewrite g . We have that

$$\begin{aligned} g(\theta) &= \mathbf{y}' \mathbf{B} \mathbf{y} - \theta' \mathbf{A}' \mathbf{B} \mathbf{y} + \theta' \mathbf{A}' \mathbf{B} \mathbf{A} \theta - \mathbf{y}' \mathbf{B} \mathbf{A} \theta \\ &= \mathbf{y}' \mathbf{B} \mathbf{y} - 2\mathbf{y}' \mathbf{B} \mathbf{A} \theta + \theta' \mathbf{A}' \mathbf{B} \mathbf{A} \theta. \end{aligned}$$

Here we have used that

$$\theta' \mathbf{A}' \mathbf{B} \mathbf{y} = \mathbf{y}' \mathbf{B} \mathbf{A} \theta$$

(both 1×1 matrices, i.e. a scalar, and each others transposed). From this follows that

$$\frac{\partial g}{\partial \theta} = -2\mathbf{A}' \mathbf{B} \mathbf{y} + 2\mathbf{A}' \mathbf{B} \mathbf{A} \theta,$$

and it is seen that

$$\frac{\partial g}{\partial \theta} = \mathbf{0} \Leftrightarrow \mathbf{A}' \mathbf{B} \mathbf{A} \theta = \mathbf{A}' \mathbf{B} \mathbf{y}.$$

This equation has as mentioned always at least one root. If $\mathbf{A}' \mathbf{B} \mathbf{A}$ is regular then we have

$$\theta_{\min} = (\mathbf{A}' \mathbf{B} \mathbf{A})^{-1} \mathbf{A}' \mathbf{B} \mathbf{y}.$$

If the matrix is singular, then we can write

$$\theta_{\min} = (\mathbf{A}' \mathbf{B} \mathbf{A})^- \mathbf{A}' \mathbf{B} \mathbf{y},$$

where $(\mathbf{A}' \mathbf{B} \mathbf{A})^-$ denotes a pseudoinverse of $\mathbf{A}' \mathbf{B} \mathbf{A}$.

We are now able to find an alternative description of the principle axes in an ellipsoid, due to

||| Theorem A.49

Let \mathbf{A} be a positive definite symmetrical matrix. The principle directions of the ellipsoid E_c with the equation

$$\mathbf{x}' \mathbf{A} \mathbf{x} = c, \quad c > 0$$

are those directions where $\mathbf{x}' \mathbf{x}$, $\mathbf{x} \in E_c$, has stationary points.

Proof

We may assume that $x = 1$. We then need to find the stationary points for

$$f(x) = x'x$$

with the condition that

$$x'Ax = 1$$

We apply a Lagrange multiplier technique and define

$$\varphi(x, \lambda) = x'x - \lambda(x'Ax - 1).$$

By differentiation we obtain

$$\frac{\partial \varphi}{\partial x} = 2x - 2\lambda Ax.$$

If this quantity is to equal 0, then

$$x = \lambda Ax$$

or

$$Ax = \frac{1}{\lambda}x,$$

i.e. x must be an eigenvector.

■

A.5 Tensor- or Kronecker product of matrices

It is an advantage to use this product when treating the multidimensional general linear model.

Definition A.50

Let A be an $m \times n$ matrix and let B be a $k \times \ell$ matrix. By the term *tensor - or Kronecker product* of A and B we mean the matrix

$$A \otimes B = (a_{ij}B) = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \quad (\text{A-6})$$

This concept corresponds to the tensor product of linear projections, which can be stated independently of coordinate system (see e.g. [26]). If this is introduced in coordinate form then we can either use A-6 or equivalently, $A \otimes B = (Ab_{ij})$. This only corresponds to changing the order of the coordinates, i.e. to changing row and columns in the respective matrices.

We briefly give some rules of calculation for the tensor-product. These are proved trially by

means of the definition.

- i) $\mathbf{O} \otimes \mathbf{A} = \mathbf{A} \otimes \mathbf{O} = \mathbf{O}$
- ii) $(\mathbf{A}_1 + \mathbf{A}_2) \otimes \mathbf{B} = \mathbf{A}_1 \otimes \mathbf{B} + \mathbf{A}_2 \otimes \mathbf{B}$
- iii) $\mathbf{A} \otimes (\mathbf{B}_1 + \mathbf{B}_2) = \mathbf{A} \otimes \mathbf{B}_1 + \mathbf{A} \otimes \mathbf{B}_2$
- iv) $\alpha \mathbf{A} \otimes \beta \mathbf{B} = \alpha \beta \mathbf{A} \otimes \mathbf{B}$
- v) $\mathbf{A}_1 \mathbf{A}_2 \otimes \mathbf{B}_1 \mathbf{B}_2 = (\mathbf{A}_1 \otimes \mathbf{B}_1)(\mathbf{A}_2 \otimes \mathbf{B}_2)$
- vi) $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$, if the inverses exist
- vii) $(\mathbf{A} \otimes \mathbf{B})^- = \mathbf{A}^- \otimes \mathbf{B}^-$
- viii) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$
- ix) Let \mathbf{A} be symmetrical and $p \times p$, have eigenvalues $\alpha_1, \dots, \alpha_p$ and eigenvectors \mathbf{x}_i , and let \mathbf{B} , be symmetrical and $q \times q$, have eigenvalues β_1, \dots, β_q and eigenvectors $\mathbf{y}_1, \dots, \mathbf{y}_q$. Then $\mathbf{A} \otimes \mathbf{B}$ will have the eigenvalues $\alpha_i \beta_j$, $i = 1, \dots, p$, $j = 1, \dots, q$, with corresponding eigenvectors.
- x) $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^q (\det \mathbf{B})^p$

$$\mathbf{x}_i \otimes \mathbf{y}_j = \begin{bmatrix} x_{1i}y_j \\ \vdots \\ x_{pi}y_j \end{bmatrix}$$

A.6 Inner products and norms

For n -dimensional vectors we note that the *inner product* or *scalar product* or *dot product* of \mathbf{x} and \mathbf{y} is defined by

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}' \mathbf{y} = (x_1 \dots x_n) \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i,$$

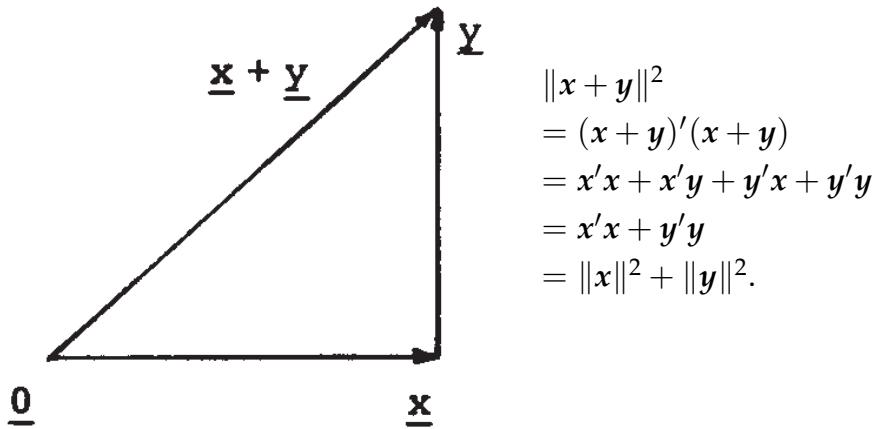
and we note that \mathbf{x} and \mathbf{y} are orthogonal if and only if

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}' \mathbf{y} = 0.$$

The corresponding norm is

$$\|\mathbf{x}\| = (\mathbf{x} \cdot \mathbf{x})^{\frac{1}{2}} = (\mathbf{x}' \mathbf{x})^{\frac{1}{2}} = \sqrt{x_1^2 + \dots + x_n^2}$$

We note that $\|\mathbf{x} - \mathbf{y}\|$ represents the euclidian distance between the points \mathbf{x} and \mathbf{y} .



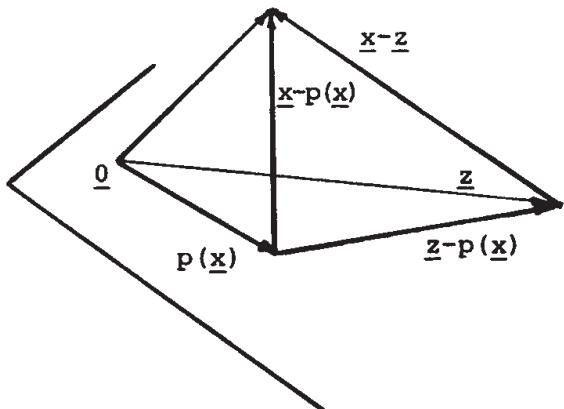
For orthogonal vectors x and y (i.e. $x \perp y$) we have the pythagorean theorem

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2;$$

see figure A.6. Further we note that the (orthogonal) projection $p(x)$ of a vector x onto the sub-space U can be determined by means of the norm, since we have that $p(x)$ is given by

$$\|x - p(x)\| = \min_{z \in U} \|x - z\|$$

||| Proof



Due to the Pythagorean theorem we have that

$$\begin{aligned}
 & \|x - p(x)\|^2 - \|z - p(x)\|^2 \\
 &= \|x - z\|^2, \\
 &\text{i.e. the minimal value of} \\
 &= \|x - z\|^2, \text{ and therefore of} \\
 &= \|x - z\| \text{ is achieved for} \\
 &z = p(x).
 \end{aligned}$$

It is now very easy to show that the validity of the above results only depend on 4 fundamental properties of the inner product. If we term the inner product of x and y by $(x|y)$

then they are

- IP1 : $(\mathbf{x}|\mathbf{y}) = (\mathbf{y}|\mathbf{x})$
- IP2 : $(\mathbf{x} + \mathbf{y}|\mathbf{z}) = (\mathbf{x}|\mathbf{z}) + (\mathbf{y}|\mathbf{z})$
- IP3 : $(k\mathbf{x}|\mathbf{y}) = k(\mathbf{x}|\mathbf{y})$
- IP4 : $\mathbf{x} \neq \mathbf{0} \Rightarrow (\mathbf{x}|\mathbf{x}) > 0.$

For an arbitrary bi-linear form $(\cdot|\cdot)$, which satisfies the above one can define a concept of orthogonality by

$$\mathbf{x} \perp \mathbf{y} \Leftrightarrow (\mathbf{x}|\mathbf{y}) = 0.$$

For an arbitrary positive definite symmetrical matrix A we can define an inner product by

$$(\mathbf{x}|\mathbf{y})_A = \mathbf{x}' A \mathbf{y}.$$

It is trivial to prove that IP 1-4 are satisfied. for this inner product and the corresponding norm given by

$$\|\mathbf{x}\|_A = \sqrt{(\mathbf{x}|\mathbf{x})_A} = \sqrt{\mathbf{x}' A \mathbf{x}},$$

we will - whenever it does not lead to confusion - use the terms $(\mathbf{x}|\mathbf{y})$ and $\|\mathbf{x}\|$.

We note that the set of points with constant A -norm equal to 1 is the set

$$\{\mathbf{x} | \|\mathbf{x}\|^2 = 1\} = \{\mathbf{x} | \mathbf{x}' A \mathbf{x} = 1\},$$

i.e. the points on an ellipsoid.

Conversely, to any non-degenerate ellipsoid there is a corresponding positive definite matrix A , so

$$E = \{\mathbf{x} | \mathbf{x}' A \mathbf{x} = 1\} = \{\mathbf{x} | \|\mathbf{x}\|_A^2 = 1\}.$$

In this way we have brought about a connection between the set of possible inner products and the set of ellipsoids.

Two vectors \mathbf{x} and \mathbf{y} are *orthogonal (with respect to A)*, if

$$\mathbf{x}' A \mathbf{y} = 0,$$

i.e. if \mathbf{x} and \mathbf{y} are conjugate directions in the ellipsoid corresponding to A .

It is also possible to introduce a *concept of angle* by means of the definition

$$\cos(\angle \mathbf{a}, \mathbf{b}) = \frac{(\mathbf{a}|\mathbf{b})}{\|\mathbf{a}\| \|\mathbf{b}\|}.$$

We now give a lemma which we will need for the theorems of independence of projections of normally distributed stochastic variables.

||| Lemma A.51

Let \mathbb{R}^n be partitioned in a direct sum

$$\mathbb{R}^n = U_1 \oplus \cdots \oplus U_k$$

of n_i dimensional sub-spaces that are orthogonal w.r.t. the positive definite matrix Σ^{-1} , i.e.

$$\mathbf{x} \perp \mathbf{y} \Leftrightarrow \mathbf{x}'\Sigma^{-1}\mathbf{y} = 0.$$

For $i = 1, \dots, k$ we let the projection p_i onto U_i be given by the matrix C_i . Then

$$C_i\Sigma C_i' = 0$$

for all $i \neq j$. Furthermore, we have

$$\Sigma^{-1}C_i = C_i'\Sigma^{-1} = C_i'\Sigma C_i.$$

||| Proof

Since $p_i \circ p_i = p_i$, we have

$$C_i C_i = C_i,$$

and since

$$p_i(x) \perp x - p_i(x),$$

(cf. the illustration) we have

$$p_i(x)' \Sigma^{-1} (x - p_i(x)) = 0,$$

i.e.

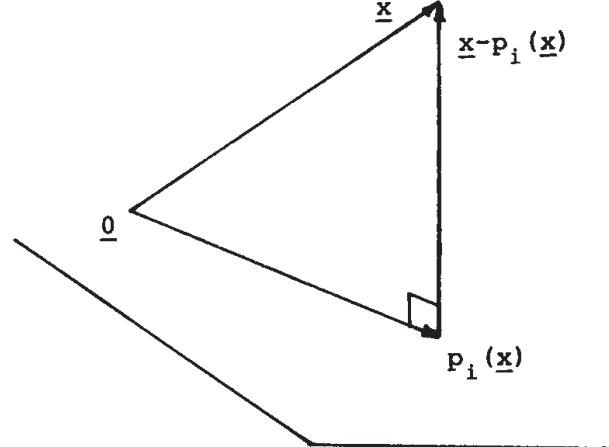
$$x C_i' \Sigma^{-1} [x - C_i x] = 0.$$

This holds for all x , and therefore

$$C_i' \Sigma^{-1} (I - C_i) = 0,$$

or

$$C_i' \Sigma^{-1} = C_i' \Sigma^{-1} C_i.$$



The right hand side of the equation is obviously symmetrical, so that

$$C_i' \Sigma^{-1} = \Sigma^{-1} C_i.$$

By pre- and post-multiplication with Σ we get

$$\Sigma C_i' = C_i \Sigma,$$

so

$$C_i \Sigma C_i' = C_i C_i \Sigma = C_i \Sigma.$$

This gives

$$\mathbf{C}_i \Sigma \mathbf{C}'_j = \mathbf{C}_i \Sigma \mathbf{C}'_i \mathbf{C}'_j = \mathbf{C}_i \Sigma \mathbf{0} = \mathbf{0}.$$

The second-last equal sign follows from the fact that the sum is direct, so for all \mathbf{x} it holds that

$$p_j(p_i(\mathbf{x})) = \mathbf{0},$$

i.e.

$$\mathbf{C}_j \mathbf{C}_i \mathbf{x} = \mathbf{0}.$$

Since \mathbf{x} - as was mentioned previously - is arbitrary, then this implies

$$\mathbf{C}_j \mathbf{C}_i = \mathbf{0},$$

or

$$\mathbf{C}'_i \mathbf{C}'_j = \mathbf{0}.$$

■

Bibliography

- [1] C. R. Rao. *Estimation and tests of significance in factor analysis*. 1955.
- [2] Ttest Wtest Anderson. *An Introduction to Multivariable Statistical Analysis*. John Wiley & sons, New York 1958.
- [3] Steen Tokkesdal Pedersen and Poul Skjøth. Statistisk analyse af data fra cementfabrikation. Eksamensprojekt, IMSOR, DTU, 1976.
- [4] Jan Gunder Knudsen. En statistisk analyse af cementstyrke. Eksamensprojekt, IMSOR, DTU, 1975.
- [5] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 2. Charles Griffin & Co., London 1967.
- [6] Otest Ltest Davies, editor. *Design and Analysis of Industrial Experiments*. Oliver and Boyd, second edition, London 1967.
- [7] H. Spliid. *User Guide for Stepwise Regression Program REGRGO*. IMSOR, Lyngby 1974.
- [8] A. E. Hoerl and R. W. Kennard. Ridge regression. biased estimation for nonorthogonal problems. In *Technometrics*, volume 12, No.1, pages 55–67. 1970.
- [9] D. W. Marquardt. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. In *Technometrics*, volume 12, No.3, pages 591–612. 1970.
- [10] D. W. Marquardt and R.D.Snee. Ridge regression in practice. *The American Statistician*, 29:3–20, 1975.
- [11] G. Wahba. Spline models for observational data. *Society for Industrial and Applied Mathematics*, 1990.
- [12] Eva Salomonsen. Fjernelse af klorider fra forhistorisk jern. Master's thesis, Konservatorskolen, København 1977.
- [13] D. N. Lawley. *The estimation of factor loadings by the method of maximum likelihood*. 1940.
- [14] John C. Davis. *Statistics and data analysis in geology*. John Wiley, New York 1973.
- [15] Ftest Ptest Agterberg. *Geomathematics. Mathematical background and geo-science applications*. Elsevier, Amsterdam 1973.

- [16] Paul S. Dwyer. *The contribution of an orthogonal multiple facto solution to multiple correlation*, volume 4. John Wiley & Sons, 1939.
- [17] Harry H. Harman. *Modern Factor Analysis*. The University of Chicago Press, second edition, Chicago 1967.
- [18] Raymond Cattell. Factor analysis: An introduction to essentials. i.the purpose and underlying models. ii.the role of factor analysis in research. *Biometrics* 21, pages 190–215, 405–435, 1965.
- [19] Donald F. Morrison. *Multivariate Statistical Methods*. McGraw-Hill, New York 1967.
- [20] K. G. Jöreskog. Some contributions to maximum likelihood factor analysis. In *Psychometrika*, volume 32. 1967.
- [21] H. F. Kaiser. *The varimax criterion for analytic rotation in factor analysis*. 1958.
- [22] P. M. Larsen. Geokemisk oversigtsprospektering. multivariable statistiske metoders anvendelighed ved interpretation af regionale geokemiske data. Eksamensprojekt, IM-SOR, DTU, 1976.
- [23] Ctest Rtest Rao and Stest Ktest Mitra. *Generalized Inverse of Matrices and Its Applications*. John Wiley, New York 1971.
- [24] J. H. Wilkinson. Error analysis of direct methods of matrix inversion. *Journal of the Association of Computing Machinery*, 8:281–330, 1961.
- [25] R. M. Johnson. On a theorem stated by Eckart and Young. In *Psychometrika*, volume 28, pages 259–263. 1963.
- [26] Nicolas Bourbaki. *Algebre*, chapter 2 Algebre Lineaire. Hermann, Paris 1967.

Index

- associative law, 335
- distributive law, 335
- inverse element, 335
- commutative law, 335
- neutral element, 335
- pseudoinverse matrix, 381
- transpose, 381
- vector-space, 334
- accuracy, 43
- adjusted R-square, 127
- affine support, 13
- affine transformation, 340
- algebra, 334
- analysis of variance, 207
 - two-sided, multidimensional, 210
 - two-way, multidimensional, 210
- Andersons U, 200
- angle, 383
- backwards elimination, 152
- basis, 336
- block matrix, 348
- BLUE - Best Linear Unbiased Estimators, 79
- canonical correlation, 283
- canonical variable, 283
- central limit therorem, 23
- Cholesky decomposition, 362
- column-vector, 340
- communalities, 307
- computational formulas, 120
- condition number, 138
- conditional distribution, 22, 44
- confidence region for
 - mean value, 187, 190
- confidenceinterval for
 - correlation-coefficient, 38
 - partial correlation-coefficient, 38
- confidenceintervals for
 - estimated value, 97
- conjugate directions, 383
- conjugated vectors, 375
- constrained estimation, 92
- contour ellipsoid, 18, 21
- Cook's D, 134
- coordinate transformation, 343
- coordinate transformation matrix, 343
- coordinates, 337
- correlation matrix, 7
- correlation-coefficient, 26, 28
- correspondence analysis, 325
- covariance, 8
- covariance-matrix, 5
- COVRATIO, 135
- Cramér's theorem, 348
- Cramér-Rao's inequality, 66
- cross validation, 174
- data-matrix, 24
- definite, 368
- deletion formula, 134
- determinant, 345, 348, 349
- DFBETAS, 136
- DFFITS, 136
- diag, 341
- diagonal element, 341
- diagonal matrix, 341
- differentiation of
 - linear form, 378
 - quadratic form, 378
- dim, 336
- dimension, 336
- direct sum, 337, 384
- Discrimination between two populations, 224
- dispersion matrix, 5
- dot product, 381
- Eckart-Young's theorem, 365
- eigenvalue, 359, 367, 377

- eigenvalue problem, general, 373
eigenvalue w.r.t. matrix, 374
eigenvector, 359, 367
eigenvector w.r.t. matrix, 374
ellipsoid, 370
elliptical cylinder, 372
empirical generalised variance, 56
empirical partial correlation, 37
empirical variance-covariance matrix, 25
estimation of
 variance-covariance matrix, 24
estimation of/in
 factor scores, 315
 eigenvalue of variance-covariance matrix,
 277
 factor loadings, 308
 variance-covariance matrix, 183, 188, 196
euclidian distance, 381
expectation, 3
expected value, 3

factor analysis, 305
 principal factor solution, 310
 estimation of loadings, 308
 maximum likelihood analysis, 319
 Q-mode analysis, 322
 rotation, 310
 test for model, 321
factor scores, 305
factors, common, 305
faktors, unique, 306
forward selection, 154
functional relation, 166

g2 inverse, 359
Gauss-Markov's theorem, 79, 196
general linear model, 102
 multidimensional, 194
generalised inverse matrix, 353
generalised inverse transformation, 351
generalised variance, 52, 273
generalized variance, 56
geodesy, 97

hat matrix, 106
hessian matrix, 63
Hotelling-Lawley's Trace, 202
Hotellings T^2
 one-sample situation, 182

two-sample situation, 188

idempotent matrix, 343, 377
idempotent projection, 339
identity matrix, 361
independence, 18
influence statistics, 132
information matrix, Fisher, 62
inner product, 381
intercept, 108
inverse matrix, 342, 347, 362, 381
isomorphism, 340

Jacobian, 64

knot, 180
Kroenecker product, 380

least squares estimate, 79
left singular vectors, 365
leverage, 136
likelihood
 conditional, 73
 marginal, 73
 partial, 73
 profile, 72
 quasi, 73
likelihood equations, 59
likelihood ratio test, 104
linear combination, 336
linear equations, solution, 354
linear functional relation, 166
linear independency, 336
linear mapping, 339
linear regression analyse, 123
linear restrictions., 83
linear transformation, 339, 342
linearity in parameters, 124
Little Jiffy, 321
logistic curve, 178
logit, 179

matrix, 340
 product, 341
 regular, 342
 sum, 341
Maximum likelihood estimate, 62
mean squared error, 170
Mean Squared Error (MSE), 43
mean value, 3

- Minimum Mean Squared Error, 44
Moore-Penrose inverse, 359
multicollinearity, 137, 169
multidimensional analysis of variance, 207, 209, 211
multidimensional general linear model, 194, 196, 207, 209, 211
multiple correlation coefficient, 126
multiple correlation-coefficient, 39
multivariate general linear model
 multiple correlation-coefficient, 41
 normal distribution, 24
 partial correlation-coefficient, 33
 $N_p(\mu, \Sigma)$, 11
Newton-Raphson algorithm, 64
norm, 381
normal distribution
 multi-dimensional, 11
 multivariate, 11
 two-dimensional, 26
normal equation, 77
null-space, 340
observed information matrix, 63
orthogonal matrix, 361
orthogonal polynomials, 138
orthogonal regression, 47, 165
orthogonal transformation, 363
orthogonal vectors, 360, 381, 383
orthonormal basis, 361
partial correlation coefficient, 31, 127
partitioned matrix, 348
partition theorem, 47
Pillai's Trace, 202
positive definite, 368
positive semi-definite, 368
precision, 13, 43
prediction, 149, 168, 176
prediction interval, 97, 99
prediction variance, 44
predictor, 43
principal component, 277
principal components, 168, 271
principal coordinate analysis, 324
principal minor, 369
projection, 339, 382
pseudoinverse matrix, 334, 350, 353, 378
pseudoinverse transformation, 351, 352
pythagorean theorem, 382
Q-mode, 322, 367
Q-modus, 365
quadratic form, 368, 378
quartimax rotation, 311
 R^2 , 126
 R^n , 335, 337, 340
R-mode, 367
R-modus, 365
random matrix, 3
range of variation, 132
rank of matrix, 344, 365, 366
rank of projection, 344
Rayleigh's coefficient, 370
Rayleigh's quotient, 370
reflection, 363
regression, 43, 44
regression analysis, 123
 by orthogonal polynomials, 138
 multidimensional, 195, 204
 non-linear, 177
regression equation
 all possible regressions, 151
 backwards elimination, 152
 choice of best, 148
 forward selection, 154, 159
 stepwise regression, 157
regular matrix, 342, 345
regularization, 171
REML, 69
reproductivity theorem for
 normal distribution, 23
residual, 81, 130
residual plot, 130
residual sum of squares, 81
Restricted Maximum Likelihood, 69
ridge estimator, 170
ridge regression, 168
ridge trace, 174
right singular vectors, 365
rotation, 363
row-vector, 340
Roy's Maximum Root, 202
RSTUDENT, 135
scalar multiplication, 335, 341

- scalar product, 381
scale-invariance, 319
Schur complement, 349
score vector, 59
scoring equations, 65
semi-definite, 368
side-subspace, 336
similar matrices, 344
similarity measures, 324
singular value, 365
Singular Value Decomposition (SVD), 365
span, 335
spectral decomposition of matrix, 362
spectrum of matrix, 362
spline function, 180
square matrix, 341
SS, 105, 209, 211
stepwise regression, 157
stochastic matrix, 3
STUDENT RESIDUAL, 135
studentised residual, 135
subspace, 335
successive testing, 113
sum of squares, SS, 105
support, 13
surveying, 97
symmetric matrix, 341, 361

tensor produkt, 380
test for/in
assumptions in regr. analysis, 128
correlation, 37
diagonal structure of variance-covariance matrix, 217
eigenvalue of variance-covariance matrix, 278
equal variance-covariance matrices, 219
factor model, 321
independence, 217
mean value, 182, 188
multidimensional analysis of variance, 207
multidimensional general linear model, 200
multiple correlation, 42
partial correlation, 37
proportional variance-covariance, 218
tolerance, 137
total variance, 274

tr, 377
trace of a matrix, 377
transpose, 342
transposed matrix, 341
U, 200
 $U(p, q, r)$, 201
uncorrelated, 10
uniqueness, 307

variance components, 70
variance inflation, 137
variance-covariance matrix, 5
variation
between groups, 208
partitioning of total, 208
partitioning total, 141, 211
within groups, 208
Varimax rotation, 311
VC, 10
vector addition, 335
vector of principal components, 272

 $W(n, \Sigma)$, 52
weighted regression, 124
Wilks' Λ , 202
Wilks' Λ , 200
Wishart distribution, 52