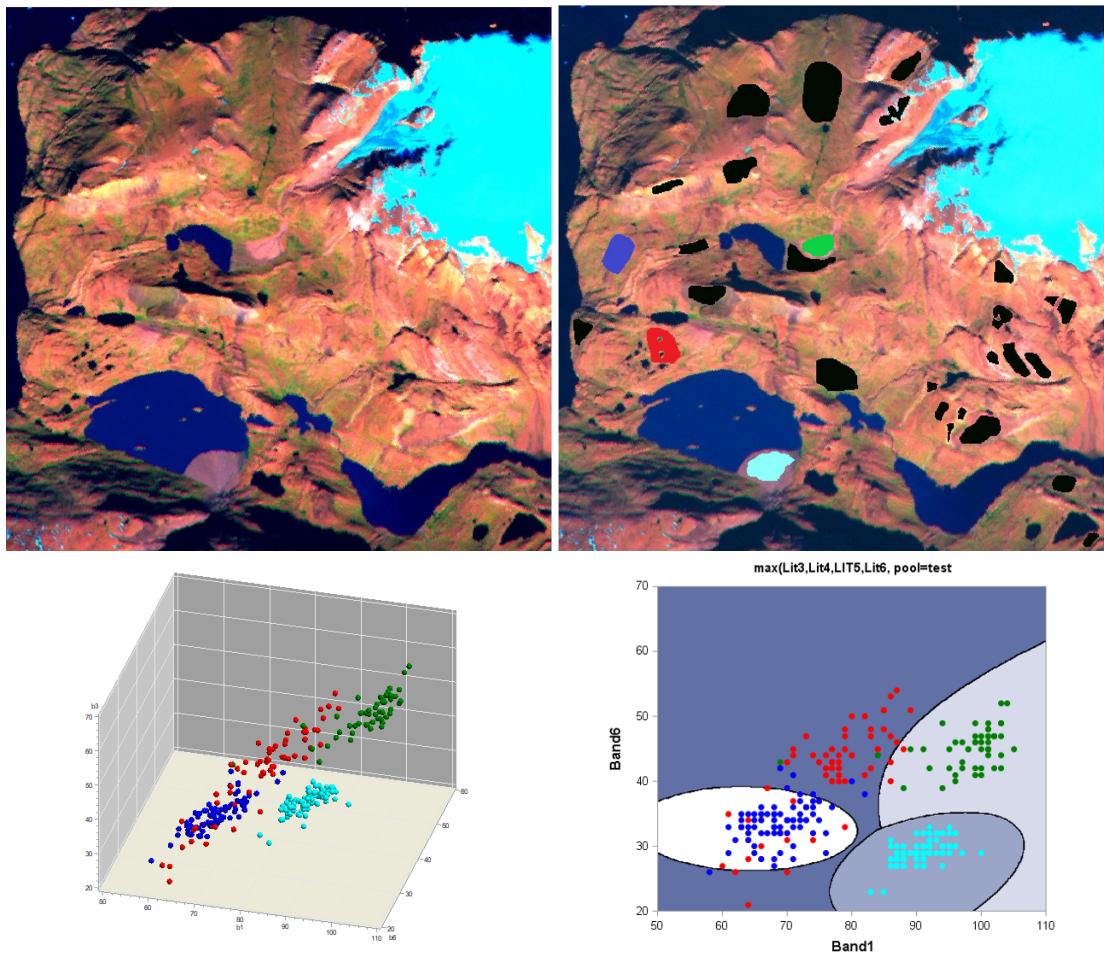


Multivariate Statistics

For the Technical Sciences

Knut Conradsen
Anders Nymark Christensen
Allan Aasbjerg Nielsen
Bjarne Kjær Ersbøll



DTU Compute, Lyngby
2019 Autumn, v. 0.94

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.	17
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.	25
1.2.6 The two-dimensional normal distribution.	27
1.3 Correlation and regression	33
1.3.1 The partial correlation coefficient.	33
1.3.2 The multiple correlation coefficient	41
1.3.3 Regression	45
1.4 The partition theorem	49
1.5 The Wishart distribution and the generalized variance	54
1.6 Complex distributions	59
1.6.1 Moments of complex distributions	59
1.6.2 The complex multivariate normal distribution	66
1.6.3 The complex multivariate Wishart distribution	72
1.7 On estimation of multidimensional parameters	83
1.7.1 Maximum likelihood estimation	83
1.7.2 Restricted Maximum Likelihood (REML)	95
1.7.3 Profile, partial, marginal, conditional, and quasi likelihood	98
2 The general linear model	100
2.1 Estimation in the general linear model	100
2.1.1 Formulation of the Model.	100
2.1.2 Estimation in the regular case	103
2.1.3 The case of $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ singular	109

2.1.4	Constrained estimation	117
2.1.5	Confidence-intervals for estimated values. Prediction-intervals	123
2.2	Tests in the general linear model	128
2.2.1	Test for a lower dimension of model space	128
2.2.2	Specification of effects and sums of squares in PROC GLM	146
2.2.3	Successive testing in the general linear model.	154
2.3	Repeated Measurements Models - RMM	158
2.3.1	Introduction	159
2.3.2	A model for repeated measurements	160
2.3.3	Some covariance structures	162
2.3.4	Subject dependent time points	163
3	Regression analysis	179
3.1	Linear regression analysis	179
3.1.1	Notation and model.	180
3.1.2	Basic properties of regression residuals	181
3.1.3	The Anatomy of Regression	182
3.1.4	Estimation of partial and multiple correlations	188
3.1.5	The partial F-test and partial correlations	194
3.1.6	The semi-partial correlation coefficient	196
3.2	Analysis of assumptions.	199
3.2.1	Analysis of residuals	200
3.2.2	On "Influence Statistics"	204
3.3	Regression using orthogonal polynomials	212
3.3.1	Definition and formulation of the model.	212
3.3.2	Determination of orthogonal polynomials.	216
3.4	Selection of the "best" regression equation	225
3.4.1	The Problem.	225
3.4.2	Examination of all regressions.	228
3.4.3	Backwards elimination.	229
3.4.4	Forward selection	230
3.4.5	Stepwise regression.	232
3.4.6	Numerical appendix.	234
3.5	Other regression models and solutions	240
3.5.1	Regularization and Ridge Regression	240
3.5.2	The Lasso, the Elastic Net, and the LARS selection algorihm	253
3.5.3	Logistic regression	262
3.5.4	Non-linear regression	269
3.5.5	Orthogonal regression (linear functional relationship).	273
4	Tests in the multidimensional normal distribution	277
4.1	Test for mean value.	277
4.1.1	Hotelling's T^2 in the One-Sample Situation	277
4.1.2	Hotelling's T^2 in the two-sample situation.	283
4.2	The multidimensional general linear model.	289

4.3	Multivariate Analyses of Variance (MANOVA)	302
4.3.1	One-way multi-dimensional analysis of variance	302
4.3.2	Two-way multidimensional analysis of variance	304
4.4	Tests regarding variance-covariance matrices	312
4.4.1	Tests regarding a single variance-covariance matrix	312
4.4.2	Test for equality of several variance-covariance matrices .	315
5	Discriminant analysis and classification	317
5.1	Discrimination between two populations	319
5.1.1	Logistic regression	319
5.1.2	Bayes and minimax solutions	319
5.1.3	Discrimination between two normal populations	323
5.1.4	Discrimination with unknown parameters	332
5.1.5	Test for best discrimination function	334
5.2	Discrimination between several populations	335
5.2.1	The Bayes solution	335
5.2.2	The Bayes' solution in the case with several normal distributions	337
5.2.3	The case with several normal distributions and unknown parameters	340
5.2.4	Short about kernel estimates and nearest neighbor estimates	343
5.3	Evaluation	347
5.3.1	Some performance measures for a classifier	347
5.3.2	Terminology from non-statistical communities.	349
5.3.3	Comparing classifiers: The ROC curve and McNemar's test.	350
5.4	Feature selection and extraction.	354
5.4.1	Test for further information	354
5.4.2	Principal Component Analysis	359
5.4.3	Canonical Discriminant Analysis	359
6	Principal components, canonical variables and correlations, factor analysis, and regression and latent structures	367
6.1	Principal components	368
6.1.1	Definition and simple characteristics	368
6.1.2	Estimation and Testing	374
6.2	Canonical variables and correlations	381
6.2.1	Definition and properties	382
6.2.2	Estimation and testing	388
6.3	Factor analysis	402
6.3.1	Model and assumptions	402
6.3.2	Estimation of factor loadings	405
6.3.3	Factor rotation	408
6.3.4	Computation of the factor scores	412
6.3.5	Briefly on maximum likelihood factor analysis	416
6.3.6	Q-mode analysis	419
6.4	PLS – Regression and Projection on Latent Structure	420

6.4.1	Introduction	420
6.4.2	Ordinary least squares regression.	420
6.4.3	Principal components regression	421
6.4.4	Canonical correlation regression	422
6.4.5	Reduced rank regression	423
6.4.6	Covariance maximization	424
6.4.7	Partial least squares regression	425
A	Summary of linear algebra	429
A.1	Vector space	429
A.1.1	Definition of a vector space	429
A.1.2	Direct sum of vector spaces	432
A.2	Linear transformations and matrices	433
A.2.1	Linear transformations	433
A.2.2	Matrices	436
A.2.3	Linear transformations using matrix-formulation	437
A.2.4	Coordinate transformation	438
A.2.5	Rank of a matrix	440
A.2.6	Determinant of a matrix	441
A.2.7	Block matrices	443
A.3	Pseudoinverse or generalised inverse matrix	446
A.4	Eigenvalue problems. Quadratic forms	455
A.4.1	Eigenvalues and eigenvectors for symmetric matrices . . .	455
A.4.2	Singular value decomposition of an arbitrary matrix. <i>Q</i> - and <i>R</i> -mode analysis	461
A.4.3	Quadratic forms and positive semi-definite matrices . . .	464
A.4.4	The general eigenvalue problem for symmetrical matrices	469
A.4.5	The trace of a matrix	473
A.4.6	Differentiation of linear form and quadratic form	474
A.5	Tensor- or Kronecker product of matrices	476
A.6	Inner products and norms	477

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 image in the upper right corner shows in black, areas with a known, homogeneous geology (lithological unit), and the coloured areas 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, 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.

We end this section with a remark, collecting the various results derived and

comparing them to the univariate case.

|||| Remark 1.10 Rules for computing moments of simple functions

$$\begin{aligned} E(a + bX) &= a + bE(X) \\ E(X + Y) &= E(X) + E(Y) \end{aligned}$$

$$\begin{aligned} V(a + bX) &= b^2 V(X) \\ V(X + Y) &= V(X) + V(Y) + 2\text{Cov}(X, Y) \\ &= V(X) + V(Y) \\ &\quad \text{iff } X, Y \text{ independent} \end{aligned}$$

$$\begin{aligned} \text{Cov}(X, X) &= V(X) \\ \text{Cov}(aX, bY) &= ab\text{Cov}(X, Y) \\ \text{Cov}(X + U, Y) &= \text{Cov}(X, Y) + \text{Cov}(U, Y) \\ \text{Cov}(X, Y + V) &= \text{Cov}(X, Y) + \text{Cov}(X, V) \end{aligned}$$

$$\begin{aligned} E(\mathbf{A} + \mathbf{X}) &= \mathbf{A} + E(\mathbf{X}) \\ E(\mathbf{AX}) &= \mathbf{A}E(\mathbf{X}) \\ E(\mathbf{XB}) &= E(\mathbf{X})\mathbf{B} \\ E(\mathbf{X} + \mathbf{Y}) &= E(\mathbf{X}) + E(\mathbf{Y}) \\ D(\mathbf{b} + \mathbf{X}) &= D(\mathbf{X}) \\ D(\mathbf{AX}) &= \mathbf{A}D(\mathbf{X})\mathbf{A}^T \\ D(\mathbf{X} + \mathbf{Y}) &= D(\mathbf{X}) + D(\mathbf{Y}) + C(\mathbf{X}, \mathbf{Y}) + C(\mathbf{Y}, \mathbf{X}) \\ &= D(\mathbf{X}) + D(\mathbf{Y}) \\ &\quad \text{iff } X, Y \text{ independent} \end{aligned}$$

$$\begin{aligned} C(\mathbf{X}, \mathbf{X}) &= D(\mathbf{X}) \\ C(\mathbf{X}, \mathbf{Y}) &= C(\mathbf{Y}, \mathbf{X})^T \\ C(\mathbf{AX}, \mathbf{BY}) &= \mathbf{A}C(\mathbf{X}, \mathbf{Y})\mathbf{B}^T \\ C(\mathbf{X} + \mathbf{U}, \mathbf{Y}) &= C(\mathbf{X}, \mathbf{Y}) + C(\mathbf{U}, \mathbf{Y}) \\ C(\mathbf{X}, \mathbf{Y} + \mathbf{V}) &= C(\mathbf{X}, \mathbf{Y}) + C(\mathbf{X}, \mathbf{V}) \end{aligned}$$

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 1.11**

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

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

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

|||| **Remark 1.12**

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.24 follows that for any positive semidefinite matrix $\boldsymbol{\Sigma}$ there exists a matrix \mathbf{A} with $\mathbf{A} \mathbf{A}^T = \boldsymbol{\Sigma}$, so the expression $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ makes sense for any positive semidefinite $p \times p$ matrix $\boldsymbol{\Sigma}$ and any p -dimensional vector $\boldsymbol{\mu}$.

Trivially, we note that

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

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

If $\boldsymbol{\Sigma}$ has full rank, then the distribution has the density given in

|||| **Theorem 1.13**

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

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

where the norm used is the one defined by $\boldsymbol{\Sigma}^{-1}$,

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

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.31 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 Σ^{-1} is often called the *precision* of the normal distribution.

If Σ 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 1.14

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

|||| Remark 1.15

If we restrict the considerations to the affine support, then X is regularly distributed and has a density as shown in theorem 1.13.

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

|||| Theorem 1.16

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

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

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 1.17**

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 1.18**

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 1.19**

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

$$\begin{aligned}\lambda_1 &= 9 \quad \wedge \quad \mathbf{p}_1 = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix}, \\ \lambda_2 &= 2 \quad \wedge \quad \mathbf{p}_2 = \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}, \\ \lambda_3 &= 0 \quad \wedge \quad \mathbf{p}_3 = \begin{bmatrix} \frac{2\sqrt{2}}{3} \\ -\frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \end{bmatrix}.\end{aligned}$$

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 \mathbf{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

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

This again gives the same description of V .

The affine support for X 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 1.20

From the example the proofs of theorems 1.16-1.18 can nearly be deduced completely.

We now formulate a trivial but useful theorem.

|||| Theorem 1.21

Let $X \sim N(\mu, \Sigma)$. Then

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

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

|||| 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

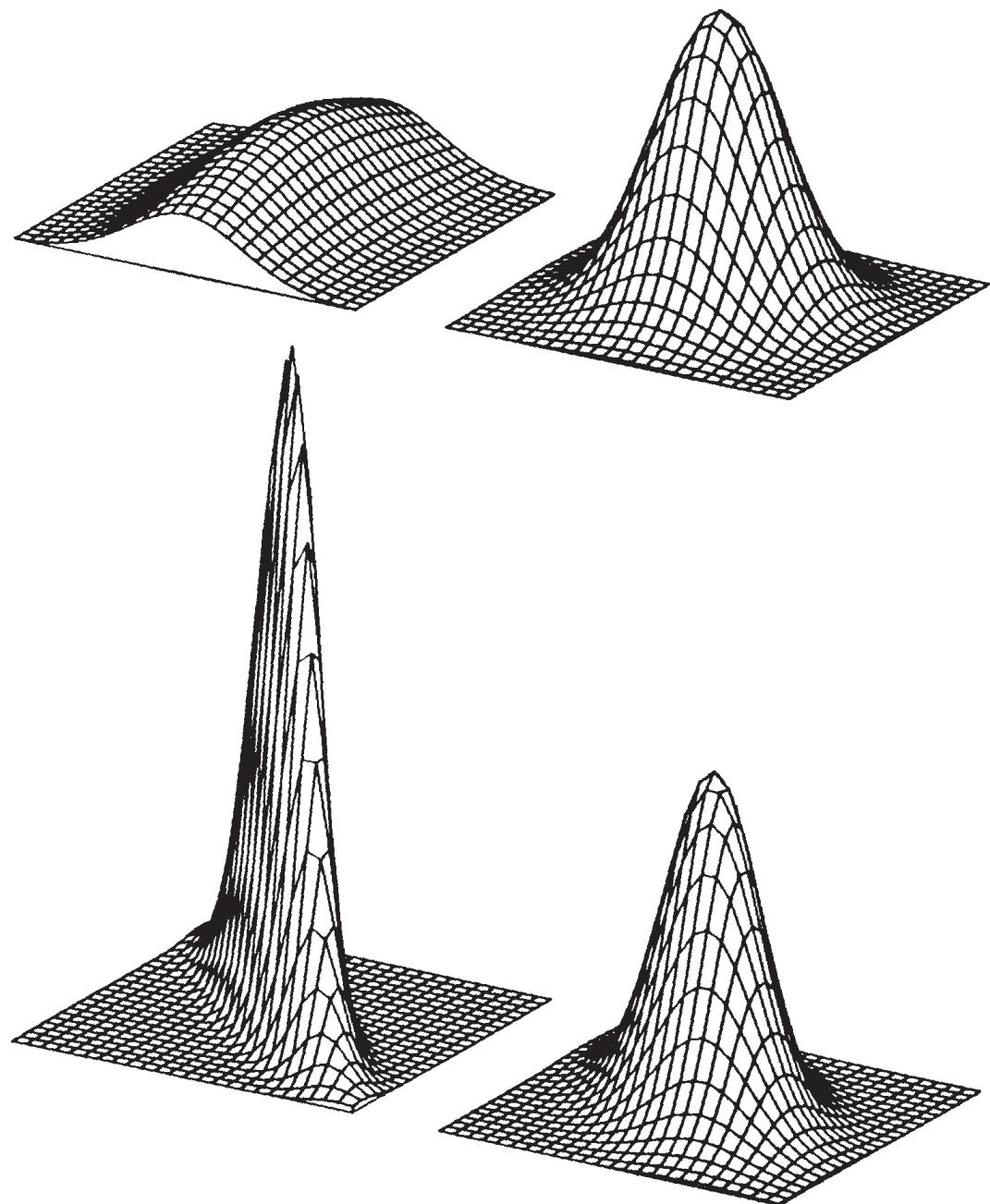


Figure 1.1 – Density functions for two-dimensional normal distributions with the variance-covariance matrices:

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

|||| **Theorem 1.22**

Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N \left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right).$$

Then

$$X_i \sim N(\mu_i, \Sigma_{ii}),$$

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 $\mathbf{X} \sim N(\mu, \Sigma)$ 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 $\mathbf{X} \sim N(\mu, \Sigma)$ and let Σ have the ortho-normed eigenvectors p_1, \dots, p_n . We now consider a coordinate system, with origo in μ and the vectors p_1, \dots, p_n as base-vectors. The coordinates in this system are called y .

If we let

$$\mathbf{P} = (p_1, \dots, p_n),$$

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

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

cf. p. 439.

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 1.23**

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

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

where Λ is a diagonal matrix with Σ 's eigenvalues on the diagonal.

|||| **Proof**

Follows from theorem 1.21 and theorem A.24.

■

|||| **Remark 1.24**

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.25, 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

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

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 $\mathbf{X} \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 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.40. From theorem A.40 is also seen that the major axes in these ellipsoids are the eigenvectors for $\boldsymbol{\Sigma}^{-1}$, but from theorem A.27 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 $\boldsymbol{\Sigma}$, and

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

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 1.25

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

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

If $\boldsymbol{\Sigma}$ has full rank p then

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

|||| **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

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

■

|||| **Remark 1.26**

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 1.35, where we consider the two-dimensional normal distribution.

1.2.3 Conditional distributions

In this section we consider the partitioning of a random variable $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, 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 1.27**

If X_2 is regularly distributed, i.e. if Σ_{22} has full rank, then the distribution of X_1 conditioned on $X_2 = x_2$ is again a normal distribution, and the following holds

$$\begin{aligned} E(X_1 | X_2 = x_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ D(X_1 | X_2 = x_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

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

|||| **Proof omitted**

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

|||| **Remark 1.28**

It is seen that the conditional dispersion of X_1 is independent of 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 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 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 reproductivity and the central limit theorem.

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

|||| Theorem 1.29 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 1.30 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. [Rao \(1955\)](#), 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 1.31

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. [Anderson \(1958\)](#), chapter 3.

|||| Theorem 1.32

Let the situation be as stated above. Then the $100(1 - \alpha)\%$ confidence ellipsoid for the unknown mean μ is

$$\{\boldsymbol{\mu} | (\boldsymbol{\mu} - \bar{\mathbf{x}})^T \mathbf{s}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}) \leq \frac{p(n-1)}{(n-p)n} F(p, n-p)_{1-\alpha}\}$$

and the $100(1 - \alpha)\%$ prediction ellipsoid for a coming observation \mathbf{x} is

$$\{\mathbf{x} | (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{s}^{-1} (\mathbf{x} - \bar{\mathbf{x}}) \leq \frac{p(n-1)(n+1)}{(n-p)n} F(p, n-p)_{1-\alpha}\}$$

|||| Proof

Follows immediately from results on Hotellings T^2 (see Chapter 4)

■

|||| **Remark 1.33**

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

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

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 $(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{bmatrix}.$$

Since

$$\det(\Sigma) = \sigma_y^2 \sigma_x^2 - \sigma_{yx}^2$$

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

$$\Sigma^{-1} = \frac{1}{\sigma_y^2 \sigma_x^2 - \sigma_{yx}^2} \begin{bmatrix} \sigma_x^2 & -\sigma_{yx} \\ -\sigma_{yx} & \sigma_y^2 \end{bmatrix}.$$

Introducing the correlation coefficient ρ

$$\rho = \frac{\sigma_{yx}}{\sigma_y \sigma_x},$$

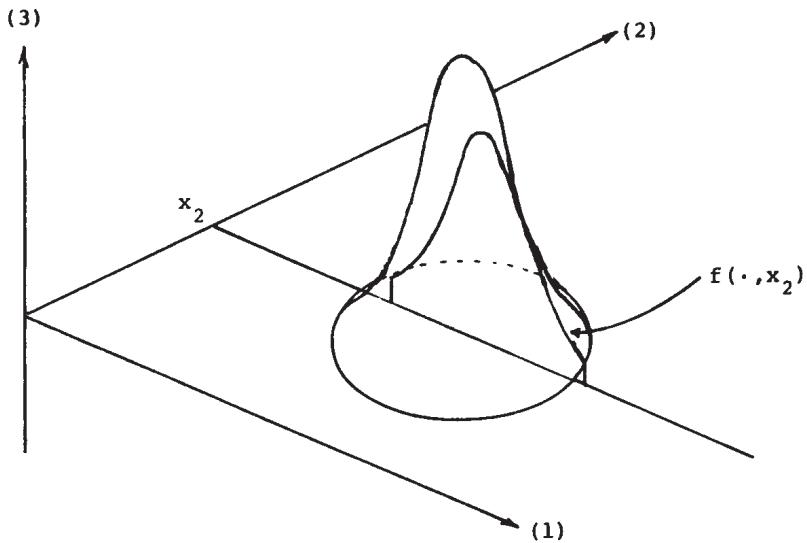


Figure 1.2 – The density of a two-dimensional normal distribution.

we get

$$\Sigma^{-1} = \frac{1}{1 - \rho^2} \begin{bmatrix} \frac{1}{\sigma_y^2} & \frac{-\rho}{\sigma_y \sigma_x} \\ \frac{-\rho}{\sigma_y \sigma_x} & \frac{1}{\sigma_x^2} \end{bmatrix},$$

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 Σ is a diagonal matrix.

The conditional distribution of Y conditioned on $X = x$ 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

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

where c is a normalisation constant. We have

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

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 1.27 for the case $n = 2$. Note, that the conditional mean depends linearly (or more correctly: affinely) upon x , and that the conditional variance is independent of x . 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 1.34

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.

Method	I	2	5	15	16	16	19	26	24	16	36
	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

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

In figure A 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 1.25) 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:

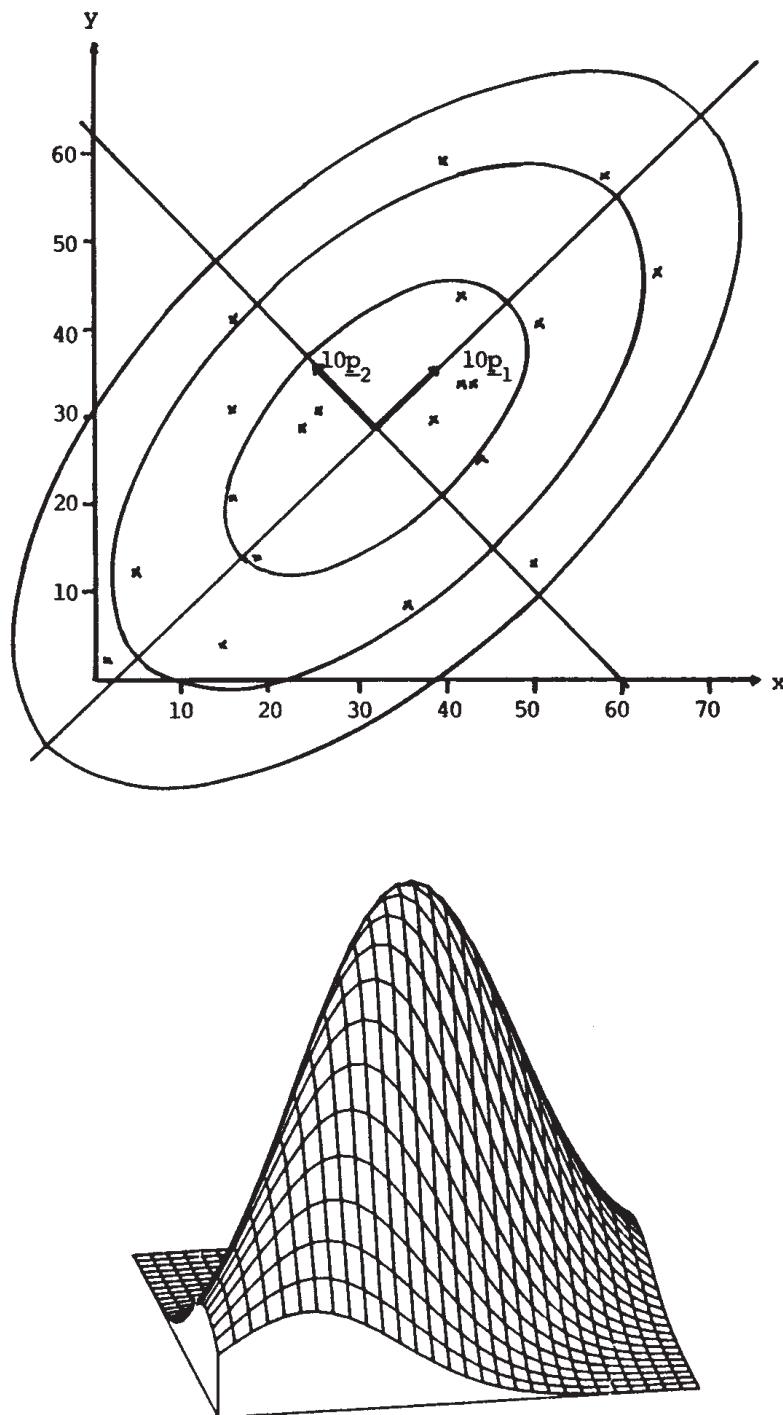


Figure A: Estimated contour ellipses and estimated density function corresponding to the data.

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

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 1.27, 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

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

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 1.35

(Data are from [Pedersen and Skjøth \(1976\)](#)).

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 $\hat{\rho}_{C_3A, \text{Strength}3} = 0.120$. 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).

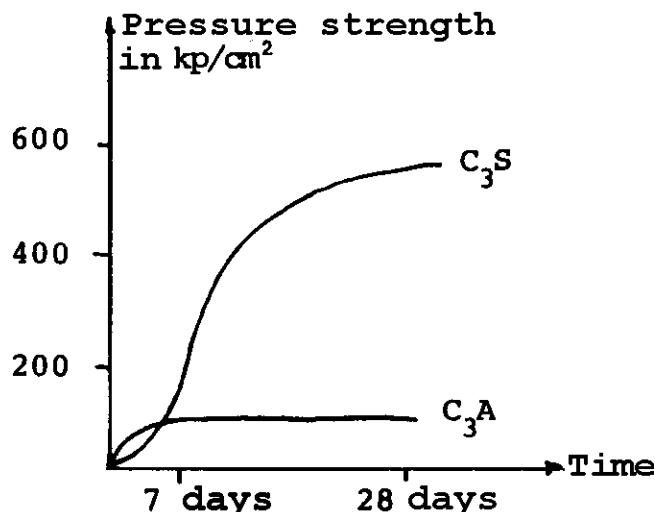


Figure A: Strength by pressure test at ordinary temperature of paste of C_3S and C_3A seasoned for different amounts of time. (from Knudsen (1975)).

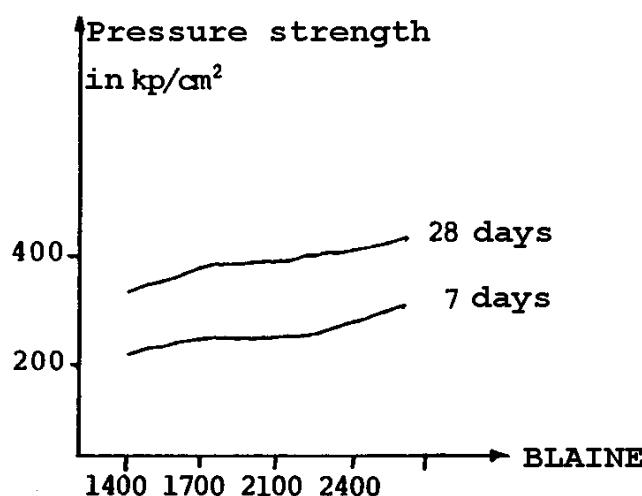


Figure B: Pressure strengths for different fine-grainedness of the cement. (from Knudsen (1975)).

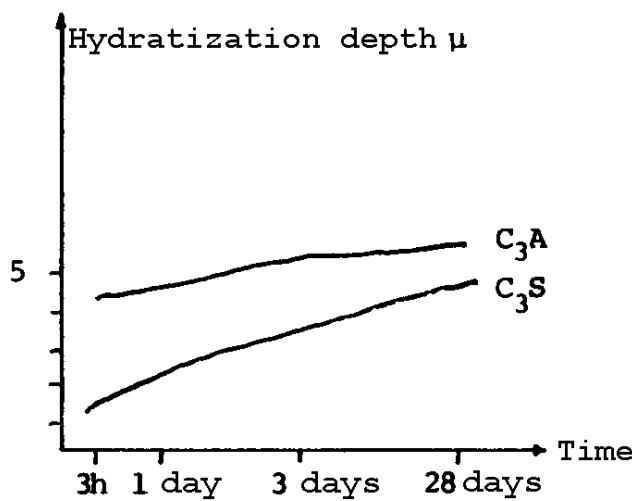


Figure C: Degree of hydratization for cement minerals and their dependence on time (from Knudsen (1975)).

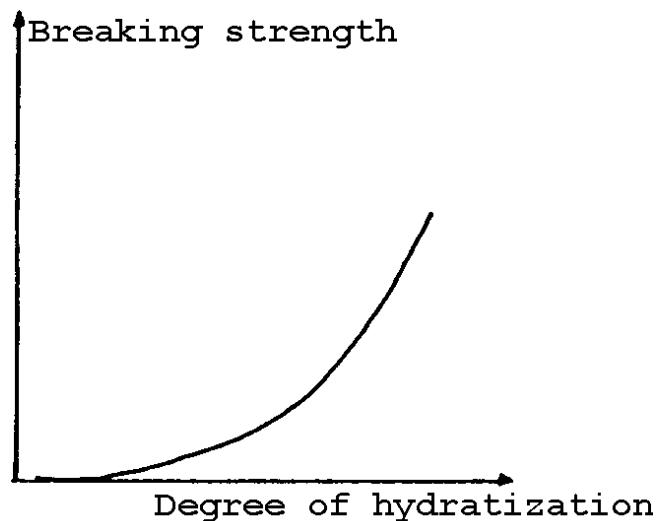


Figure D: Relationship between degree of hydratization and strength (from Knudsen (1975)).

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 C_3A from the effects which arise from a C_3A -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_3S	C_3A	Strength 3	Strength 28
C_3S	1	-0.333	0.137	0.333
C_3A	-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 C_3A 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 C_3A will tend to have lower strengths. This might indicate that the large 3-day strength for cements with high content of C_3A rather depends on these cements having a large BLAINE (that they are crushed somewhat easier) than that C_3A hydrates quickly!

We see a corresponding effect on the correlation between C_3A and 28-day strength. Here the unconditional correlation is -0.168 and the partial correlation for fixed BLAINE has become -0.246.

||| Remark 1.36

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 C_3A 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 C_3A 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 1.37

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 1.38

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 1.35 in

|||| Example 1.39

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 1.40

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 1.41

Let us determine a 95% confidence interval for $\rho_{24|3}$ in example 1.39. 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. 29

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

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 1.42

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. 23) on Σ_i by use of the matrix formulas in section A.2.7.

■

|||| Remark 1.43

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. 35 regarding the estimation of partial coefficients holds.

In the next example we continue example 1.39.

|||| Example 1.44

To get an impression of to which degree the content of C₃A and C₃S in example 1.39 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 1.35. 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 1.45

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 1.46

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

Theorem 1.45 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 1.47

Consider the situation in example 1.44. 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 1.48**

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 1.27 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 1.49.

||| Example 1.49

In this example we will determine the linear relations from a measurement by one of the two methods stated in example 1.34 to the other measurement.

We find the regressions

$$\begin{aligned}\hat{E}(X_1|X_2 = x_2) &= \bar{x}_1 + \hat{\rho} \frac{s_1}{s_2} (x_2 - \bar{x}_2) \\ &= 32.35 + 0.62 \frac{\sqrt{311}}{\sqrt{279}} (x_2 - 29.05) \\ &= 0.65x_2 + 13.43\end{aligned}$$

and

$$\begin{aligned}\hat{E}(X_2|X_1 = x_1) &= \bar{x}_2 + \hat{\rho} \frac{s_2}{s_1} (x_1 - \bar{x}_1) \\ &= 29.05 + 0.62 \frac{\sqrt{279}}{\sqrt{311}} (x_1 - 32.35) \\ &= 0.58x_1 + 10.14.\end{aligned}$$

These lines are shown in figure 1.49.

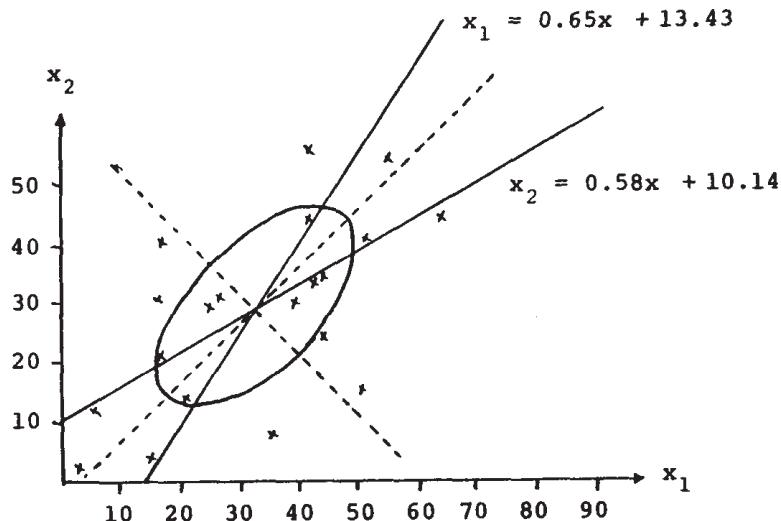


Figure 1.49

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 1.50 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

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

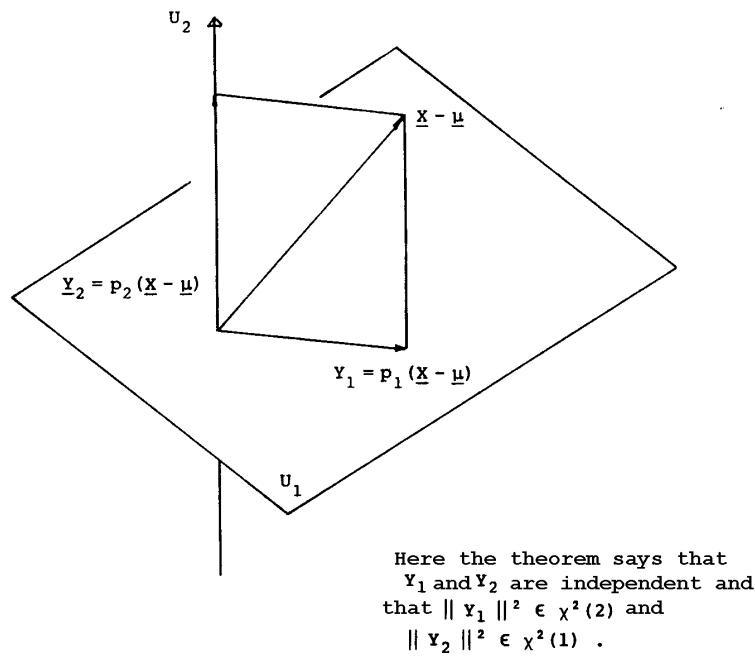


Figure 1.4: Here the theorem says that Y_1 and Y_2 are independent and that $\|Y_2\|^2 \in \chi^2(1)$.

Now for $i \neq j$ it follows from the lemma on page 480 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.55 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.36 p. 464) therefore there exists an orthogonal (and even orthonormal) matrix \mathbf{P}' (theorem A.24) 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 1.51

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

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

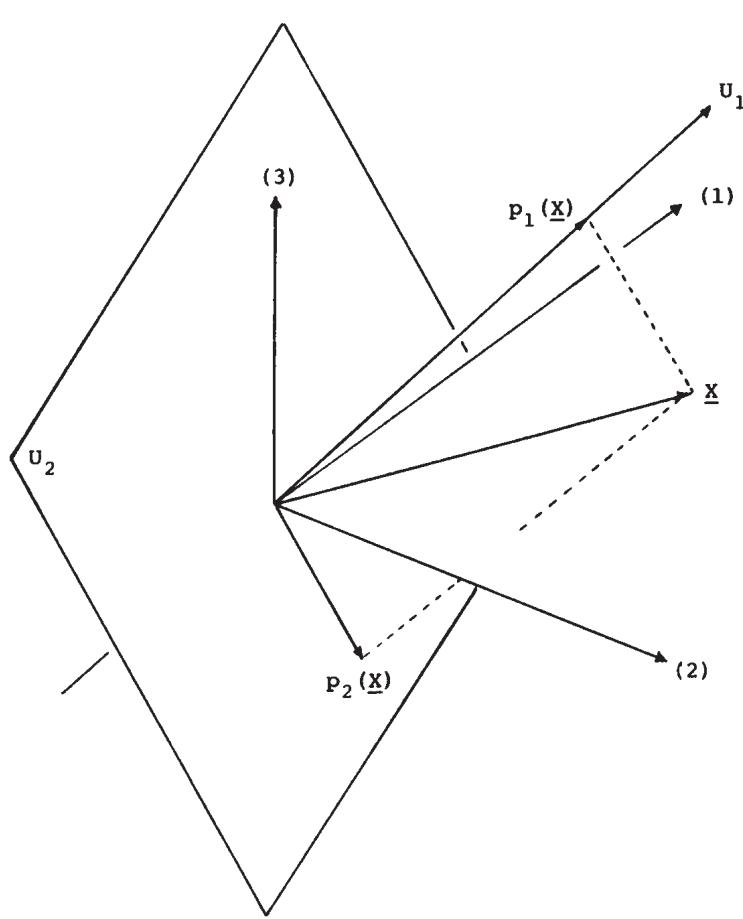


Figure 1.51:

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 1.52

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 1.53

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 1.54

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 1.55

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 1.56**

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

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

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 1.57**

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 1.58

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 1.59

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 1.60

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 1.61

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 1.62

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 1.63**

Let $\mathbf{X}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i = 1, \dots, n$, be stochastically independent. Then the empirical generalised variance follows the same distribution as

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

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 1.64**

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

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

1.6 Complex distributions

1.6.1 Moments of complex distributions

||| Definition 1.65

Let X and Y be real random variables. Then we say that $Z = U + iV$ is a *complex random variable*. Assuming that the involved real moments exist, we define the *expectation* or *mean value* as

$$E(Z) = E(U) + iE(V) = \mu_u + i\mu_v = \mu_z$$

and the *variance* as

$$V(Z) = E\left\{(Z - \mu_z)(\overline{Z - \mu_z})\right\} = V(U) + V(V)$$

For two complex random variables X and Y we define the *covariance* as

$$\text{Cov}(X, Y) = E\left\{(X - \mu_x)(\overline{Y - \mu_y})\right\}$$

||| Theorem 1.66

For complex numbers a, b, c, d and Z, X, Y as above, we have rules very similar to the real case, i.e. we have

1. $E(\overline{Z}) = \overline{E(Z)}$
2. $E(aZ+b) = aE(Z) + b$
3. $E(X+Y) = E(X) + E(Y)$
4. $V(Z) = E(Z\overline{Z}) - E(Z)\overline{E(Z)}$
5. $V(aZ+b) = a\overline{a}V(Z)$
6. $V(X+Y) = V(X) + V(Y)$ if X and Y are independent
7. $V(X+Y) = V(X) + V(Y) + 2\text{Re}(\text{Cov}(X, Y))$
8. $\text{Cov}(X, Y) = E(X\overline{Y}) - E(X)\overline{E(Y)}$
9. $\text{Cov}(X, Y) = \overline{\text{Cov}(X, Y)}$
10. $\text{Cov}(Z, X+Y) = \text{Cov}(Z, X) + \text{Cov}(Z, Y)$
 $\text{Cov}(X+Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$
11. $\text{Cov}(aX+b, cY+d) = a\overline{c}\text{Cov}(X, Y)$

|||| Proof omitted

Straightforward calculations.

A vector of complex random variables is called a *multivariate complex random variable*, and a matrix of complex random variables is called a *complex random matrix*. Similarly to the real case we generalize the concepts of mean, variance and covariance to the multivariate case.

|||| Definition 1.67 Complex Mean

The complex random matrix

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

has the *expected value* or *mean value*

$$E(\mathbf{Z}) = \begin{bmatrix} E(Z_{11}) & \cdots & E(Z_{1k}) \\ \vdots & & \vdots \\ E(Z_{n1}) & \cdots & E(Z_{nk}) \end{bmatrix}$$

||| Definition 1.68 Complex Dispersion

The *dispersion (variance-covariance matrix)* of a complex random vector

$$\begin{aligned} \mathbf{X} &= \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \\ D(\mathbf{X}) &= E \left\{ (\mathbf{X} - \boldsymbol{\mu}_x) \overline{(\mathbf{X} - \boldsymbol{\mu}_x)}^T \right\} \\ &= E \left\{ (\mathbf{X} - \boldsymbol{\mu}_x) (\mathbf{X} - \boldsymbol{\mu}_x)^H \right\} \\ &= \begin{bmatrix} V(X_1) & \cdots & Cov(X_1, X_p) \\ \vdots & & \vdots \\ Cov(X_p, X_1) & \cdots & V(X_p) \end{bmatrix} \\ &= \begin{bmatrix} E \left\{ (X_1 - \mu_{x1}) \overline{(X_1 - \mu_{x1})} \right\} & \cdots & E \left\{ (X_1 - \mu_{x1}) \overline{(X_p - \mu_{xp})} \right\} \\ \vdots & & \vdots \\ E \left\{ (X_p - \mu_{xp}) \overline{(X_1 - \mu_{x1})} \right\} & \cdots & E \left\{ (X_p - \mu_{xp}) \overline{(X_p - \mu_{xp})} \right\} \end{bmatrix} \end{aligned}$$

||| Definition 1.69 Complex Covariance

The *covariance* between two complex random vectors

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

is

$$\begin{aligned} C(\mathbf{X}, \mathbf{Y}) &= E \left\{ (\mathbf{X} - \boldsymbol{\mu}_x) \overline{(\mathbf{Y} - \boldsymbol{\mu}_y)}^T \right\} \\ &= E \left\{ (\mathbf{X} - \boldsymbol{\mu}_x) (\mathbf{Y} - \boldsymbol{\mu}_y)^H \right\} \\ &= \begin{bmatrix} Cov(X_1, Y_1) & \cdots & Cov(X_1, Y_q) \\ \vdots & & \vdots \\ Cov(X_p, Y_1) & \cdots & Cov(X_p, Y_q) \end{bmatrix} \end{aligned}$$

|||| **Theorem 1.70**

Let \mathbf{Z} , \mathbf{Z}_1 , and \mathbf{Z}_2 be complex random matrices and \mathbf{T} , \mathbf{X} , and \mathbf{Y} be complex random vectors. Furthermore, let \mathbf{A} , \mathbf{B} and \mathbf{C} be constant matrices and \mathbf{b} a constant vector with dimensions so that the matrix sums and products in the sequel exist. Matrices and vectors with identical symbols in different formulas are not necessarily identical. Then

1. $E(\mathbf{Z}^T) = (E(\mathbf{Z}))^T$, $E(\overline{\mathbf{Z}}) = \overline{E(\mathbf{Z})}$, $E(\mathbf{Z}^H) = (E(\mathbf{Z}))^H$
2. $E(\mathbf{A}\mathbf{Z}\mathbf{B} + \mathbf{C}) = \mathbf{A}E(\mathbf{Z})\mathbf{B} + \mathbf{C}$
3. $E(\mathbf{Z}_1 + \mathbf{Z}_2) = E(\mathbf{Z}_1) + E(\mathbf{Z}_2)$
4. $D(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^H) - E(\mathbf{X})\{E(\mathbf{X})\}^H$
5. $D(\mathbf{A}\mathbf{X} + \mathbf{b}) = \mathbf{A}D(\mathbf{X})\mathbf{A}^H$
6. $D(\mathbf{X} + \mathbf{Y}) = D(\mathbf{X}) + D(\mathbf{Y})$ if \mathbf{X} and \mathbf{Y} are independent
7. $D(\mathbf{X} + \mathbf{Y}) = D(\mathbf{X}) + D(\mathbf{Y}) + C(\mathbf{X}, \mathbf{Y}) + C(\mathbf{Y}, \mathbf{X})$
8. $D(\mathbf{X}) = C(\mathbf{X}, \mathbf{X})$
9. $C(\mathbf{X}, \mathbf{Y}) = E(\mathbf{X}\mathbf{Y}^H) - E(\mathbf{X})\{E(\mathbf{Y})\}^H$
10. $C(\mathbf{X}, \mathbf{Y}) = 0$ if \mathbf{X} and \mathbf{Y} are independent
11. $C(\mathbf{X}, \mathbf{Y}) = \{C(\mathbf{Y}, \mathbf{X})\}^H$
12. $C(\mathbf{A}\mathbf{X}, \mathbf{B}\mathbf{Y}) = \mathbf{A}C(\mathbf{X}, \mathbf{Y})\mathbf{B}^H$
13. $C(\mathbf{T}, \mathbf{X} + \mathbf{Y}) = C(\mathbf{T}, \mathbf{X}) + C(\mathbf{T}, \mathbf{Y})$, $C(\mathbf{X} + \mathbf{Y}, \mathbf{T}) = C(\mathbf{X}, \mathbf{T}) + C(\mathbf{Y}, \mathbf{T})$

|||| **Proof omitted**

Straightforward calculations.

An important result is

||| Theorem 1.71

For a complex multivariate random variable \mathbf{X} , the dispersion matrix $\Sigma = D(\mathbf{X})$ is Hermitian and positive semidefinite, i.e.

$$\Sigma = \Sigma^H$$

and

$$\mathbf{c}^H \Sigma \mathbf{c} \geq 0$$

for any complex vector \mathbf{c} . If Σ has full rank, it is positive definite.

||| Proof omitted

||| Definition 1.72

The *pseudo covariance matrix* or the *relation matrix* of a complex multivariate random variable \mathbf{X} is the matrix

$$C(\mathbf{X}, \bar{\mathbf{X}}) = \Gamma = \begin{bmatrix} E\{(X_1 - \mu_{x1})(X_1 - \mu_{x1})\} & \cdots & E\{(X_1 - \mu_{x1})(X_p - \mu_{xp})\} \\ \vdots & & \vdots \\ E\{(X_p - \mu_{xp})(X_1 - \mu_{x1})\} & \cdots & E\{(X_p - \mu_{xp})(X_p - \mu_{xp})\} \end{bmatrix}$$

If

$$\mathbf{X} = \mathbf{U} + i\mathbf{V} = \begin{bmatrix} U_1 + iV_1 \\ \vdots \\ U_p + iV_p \end{bmatrix}$$

then the $2p$ -dimensional, real valued vector $\begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix}$ has a dispersion matrix of the form

$$D \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} = D \begin{bmatrix} U_1 \\ \vdots \\ U_p \\ V_1 \\ \vdots \\ V_p \end{bmatrix} = \begin{bmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{bmatrix}$$

The dispersion matrix and the relation matrix for \mathbf{X} may be determined from the components in this real variance-covariance matrix. We give the results in the next

|||| **Theorem 1.73**

Let X, U , and V be as above. Then

$$\Sigma = \Sigma_{uu} + \Sigma_{vv} + i(\Sigma_{vu} - \Sigma_{uv})$$

$$\Gamma = \Sigma_{uu} - \Sigma_{vv} + i(\Sigma_{vu} + \Sigma_{uv})$$

|||| **Proof**

We only consider the first expression. Let us assume that the means are equal to zero. Then consider $j > k$

$$\begin{aligned} E\{X_j \bar{X}_k\} &= E\{(U_j + iV_j)(U_k - iV_k)\} \\ &= E\{U_j U_k + V_j V_k + i(V_j U_k - U_j V_k)\} \\ &= (\Sigma_{uu})_{jk} + (\Sigma_{vv})_{jk} + i((\Sigma_{vu})_{jk} - (\Sigma_{uv})_{jk}) \end{aligned}$$

and the proof follows. ■

|||| **Definition 1.74**

A complex random vector X is *circularly symmetrically distributed* if $e^{i\phi} X$ has the same distribution as X for all real ϕ . This is equivalent to $\Gamma_x = \mathbf{0}$.

The last remark in the definition follows from the identity

$$\Gamma_x = \Gamma_{e^{i\phi} x} = C(e^{i\phi} X, \overline{e^{i\phi} X}) = E(e^{2i\phi} X X^T) = e^{2i\phi} \Gamma_x$$

which can only be true for all real ϕ if $\Gamma_x = \mathbf{0}$.

For a circularly-symmetric complex distribution the dispersion matrix of the corresponding $2p$ -dimensional real distribution takes a particular form. We have

||| Corollary 1.75

Let X be circularly-symmetrically distributed. Using the notation from above we have

$$D \begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{vu} & \Sigma_{vv} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} Re(\Sigma) & -Im(\Sigma) \\ Im(\Sigma) & Re(\Sigma) \end{bmatrix}$$

||| Proof

If $\Gamma = \mathbf{0}$, we obviously have

$$\Sigma_{uu} = \Sigma_{vv} \quad \text{and} \quad \Sigma_{vu} = -\Sigma_{uv}$$

Then

$$\Sigma = 2\Sigma_{uu} - 2i\Sigma_{uv}$$

and

$$\Sigma_{uu} = \Sigma_{vv} = \frac{1}{2}Re(\Sigma)$$

$$\Sigma_{uv} = -\Sigma_{vu} = -\frac{1}{2}Im(\Sigma)$$

which concludes the proof.

■

1.6.2 The complex multivariate normal distribution

There are several ways of defining a complex normal (or Gaussian) distribution. The most direct definition is simply to state that a (multivariate) Gaussian distribution is the distribution of a complex random variable, where the real and the imaginary parts are real Gaussian variables with a joint real Gaussian distribution. In many (most) applications we shall however, restrict ourselves to consider circularly-symmetric normal distributions. Also for mathematical reasons it is convenient to include circular symmetry in the definition. If we are using the broader definition, the distribution (with mean zero) will not be determined by its dispersion matrix alone. We must also know the relation matrix. If however, we limit ourselves to the circularly-symmetric case (fig. 1.3) we - as in the real case – only need to know the dispersion matrix. Similarly most well-known results from the real case will have an intuitive complex counterpart.

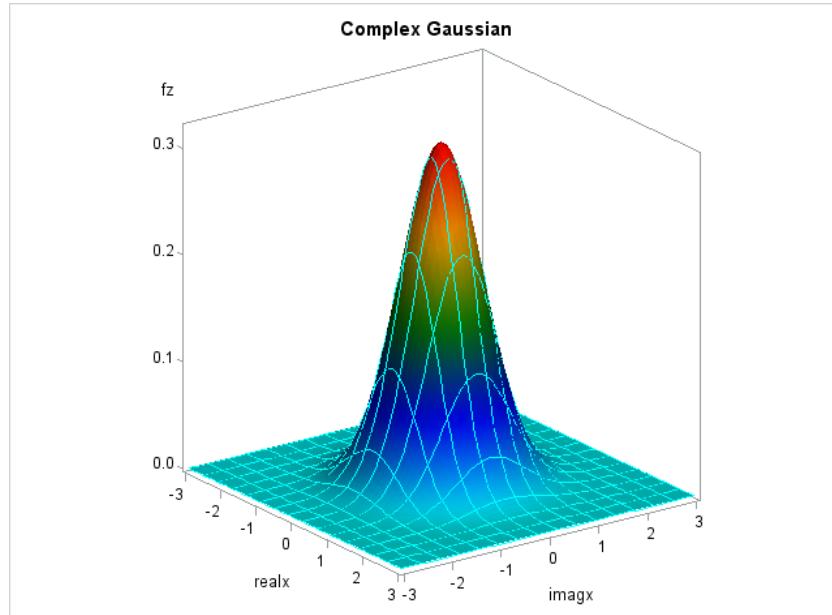


Figure 1.3 – The density function for a one-dimensional complex normal distribution.

We initially define the univariate complex normal distribution in

||| Definition 1.76

The complex random variable X has a (*circularly-symmetric univariate complex normal distribution*) with mean zero and variance 1 if, for $U = \operatorname{Re}(X)$ and $V = \operatorname{Im}(X)$, the distribution of $(U \ V)^T$ is the two-dimensional real normal distribution

$$\begin{bmatrix} U \\ V \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{1}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = N_2(\mathbf{0}, \frac{1}{2}I).$$

We write $X \sim N_{\mathbb{C}}(0, 1)$ and the frequency function is

$$f(x) = \frac{1}{\pi} e^{-\bar{x}x} = \frac{1}{\pi} e^{-(u^2+v^2)}, \quad x = u + iv \in \mathbb{C}$$

Further to this we say that $Y = \mu + \sigma X$, $\mu \in \mathbb{C}$, $\sigma \in \mathbb{R}_+$, has a (univariate) complex normal distribution with mean μ and variance σ^2 , and we write $Y \sim N_{\mathbb{C}}(\mu, \sigma^2)$.

The frequency function for Y is

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

and the corresponding two-dimensional real distribution is

$$N_2 \left(\begin{bmatrix} Re(\mu) \\ Im(\mu) \end{bmatrix}, \frac{\sigma^2}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) = N_2 \left(\begin{bmatrix} Re(\mu) \\ Im(\mu) \end{bmatrix}, \frac{\sigma^2}{2} I \right).$$

It is now possible to define the multivariate complex normal distribution

||| Definition 1.77

We say that the random variable $X \in \mathbb{C}^p$ has a (*circularly-symmetric*) **complex multivariate normal (or Gaussian) distribution**, if all linear combinations $a^H X$, $a \in \mathbb{C}^p$, follow a univariate, complex normal distribution.

||| Remark 1.78

In the sequel we shall only consider circularly-symmetric complex multivariate normal distributions, wherefore we shall omit the term “circularly-symmetric”.

||| Theorem 1.79

Let the $X \in \mathbb{C}^p$ have a complex multivariate normal distribution with mean μ and dispersion matrix (variance-covariance matrix) Σ , i.e.

$$X \sim N_{\mathbb{C}}(\mu, \Sigma).$$

If Σ has full rank then X has the frequency function

$$\begin{aligned} f(x) &= \frac{1}{\pi^p \det \Sigma} \exp \left(-(\overline{x - \mu})^T \Sigma^{-1} (x - \mu) \right) = \\ &= \frac{1}{\pi^p \det \Sigma} \exp \left(-(x - \mu)^H \Sigma^{-1} (x - \mu) \right) \\ &= \frac{1}{\pi^p \det \Sigma} \exp \left(-\text{tr}\{\Sigma^{-1}(x - \mu)(x - \mu)^H\} \right) \end{aligned}$$

|||| Proof omitted

Omitted. The classical reference to this is [Goodman \(1963\)](#). A reference using a more up-to-date mathematical notation is [Andersen et al. \(1995\)](#). [Gallager \(2008\)](#) is focussing upon equivalence of conditions leading to circular symmetry.

To elaborate a little bit further on the results in section [1.6.1](#), we shall consider the elements in the dispersion matrix Σ . With the notation from [Goodman \(1963\)](#) we may write the dispersion matrix on the form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & (\alpha_{12} + i\beta_{12})\sigma_1\sigma_2 & \cdots & (\alpha_{1p} + i\beta_{1p})\sigma_1\sigma_p \\ (\alpha_{21} + i\beta_{21})\sigma_2\sigma_1 & \sigma_2^2 & \cdots & (\alpha_{2p} + i\beta_{2p})\sigma_2\sigma_p \\ \vdots & & & \vdots \\ (\alpha_{p1} + i\beta_{p1})\sigma_p\sigma_1 & (\alpha_{p2} + i\beta_{p2})\sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{bmatrix}$$

$$= Re(\Sigma) + iIm(\Sigma)$$

where

$$\alpha_{jk} = \alpha_{kj} \quad \text{and} \quad \beta_{jk} = -\beta_{kj}$$

and

$$Re(\Sigma) = \begin{bmatrix} \sigma_1^2 & \alpha_{12}\sigma_1\sigma_2 & \cdots & \alpha_{1p}\sigma_1\sigma_p \\ \alpha_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \alpha_{2p}\sigma_2\sigma_p \\ \vdots & & & \vdots \\ \alpha_{p1}\sigma_p\sigma_1 & \alpha_{p2}\sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{bmatrix}$$

$$Im(\Sigma) = \begin{bmatrix} 0 & \beta_{12}\sigma_1\sigma_2 & \cdots & \beta_{1p}\sigma_1\sigma_p \\ \beta_{21}\sigma_2\sigma_1 & 0 & \cdots & \beta_{2p}\sigma_2\sigma_p \\ \vdots & & & \vdots \\ \beta_{p1}\sigma_p\sigma_1 & \beta_{p2}\sigma_p\sigma_2 & \cdots & 0 \end{bmatrix}$$

Obviously

$$(Re(\Sigma))^T = Re(\Sigma)$$

and

$$(Im(\Sigma))^T = -Im(\Sigma).$$

Introducing the real valued random variables U and V as in the previous section we again have

$$D \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} = D \begin{bmatrix} Re(\mathbf{X}) \\ Im(\mathbf{X}) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} Re(\boldsymbol{\Sigma}) & -Im(\boldsymbol{\Sigma}) \\ Im(\boldsymbol{\Sigma}) & Re(\boldsymbol{\Sigma}) \end{bmatrix}$$

and

$$V(U_j) = V(V_j) = \frac{1}{2}\sigma_j^2$$

$$Cov(U_j, V_j) = 0$$

$$Cov(U_j, U_k) = Cov(V_j, V_k) = \frac{1}{2}\alpha_{jk}\sigma_j\sigma_k \quad j \neq k$$

$$Cov(U_j, V_k) = -Cov(U_k, V_j) = \frac{1}{2}\beta_{jk}\sigma_j\sigma_k \quad j \neq k$$

which again underlines the consequences of imposing that the circularly-symmetric condition.

Most of the theorems on the multivariate normal distribution from the real case do have a complex counterpart. We summarize a few of the most important in

|||| **Theorem 1.80**

Let $\mathbf{X} \sim N_{\mathbb{C}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then – assuming that the implied products etc. exist –

$$\mathbf{AX} + \mathbf{b} \sim N_{\mathbb{C}}(A\boldsymbol{\mu} + \mathbf{b}, A\boldsymbol{\Sigma}A^H)$$

Let \mathbf{X} be partitioned as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{\mathbb{C}} \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right).$$

Then \mathbf{X}_1 and \mathbf{X}_2 are (stachastically) independent if and only if

$$\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^H = \mathbf{0}.$$

The marginal distribution of \mathbf{X}_j is again complex normal, i.e.

$$\mathbf{X}_j \sim N_{\mathbb{C}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_{jj})$$

Furthermore the conditional distribution of \mathbf{X}_1 given \mathbf{X}_2 is

$$(\mathbf{X}_1 | \mathbf{X}_2) \sim N_{\mathbb{C}}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

The sum of independent, complex normally distributed random variables is also complex normally distributed, i.e. we for independent random variables $\mathbf{X}_j \sim N_{\mathbb{C}}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j = 1, \dots, n$, have

$$\sum_j \mathbf{X}_j \sim N_{\mathbb{C}} \left(\sum_j \boldsymbol{\mu}_j, \sum_j \boldsymbol{\Sigma}_j \right)$$

|||| **Proof omitted**

Parameter estimation is also a straightforward generalization of the results from the real case.

|||| **Theorem 1.81**

We consider the independent, identically distributed random variables $X_j \sim N_C(\mu, \Sigma)$, $j = 1, \dots, n$. We also call the X_j 's independent observations from the distribution $N_C(\mu, \Sigma)$. We organize the observations in the data matrix

$$\mathbf{X} = [X_1, \dots, X_n]^H = \begin{bmatrix} \mathbf{X}_1^H \\ \vdots \\ \mathbf{X}_n^H \end{bmatrix} = \begin{bmatrix} \bar{X}_{11} & \cdots & \bar{X}_{1p} \\ \vdots & & \vdots \\ \bar{X}_{n1} & \cdots & \bar{X}_{np} \end{bmatrix},$$

Then the maximum likelihood estimator for μ is

$$\hat{\mu} = ave(\mathbf{X}) = \frac{1}{n} \sum_j \mathbf{X}_j$$

which is unbiased (for estimating μ), i.e. $E(\hat{\mu}) = \mu$. For known mean value μ , which we may assume is equal to $\mathbf{0}$, the maximum likelihood estimator for Σ is

$$\frac{1}{n} \sum_j \mathbf{X}_j \mathbf{X}_j^H = \frac{1}{n} [\mathbf{X}_1 \cdots \mathbf{X}_n] \begin{bmatrix} \mathbf{X}_1^H \\ \vdots \\ \mathbf{X}_n^H \end{bmatrix} = \frac{1}{n} \mathbf{X}^H \mathbf{X}$$

This is an unbiased estimator of Σ . For unknown μ we get the unbiased estimator

$$\hat{\Sigma} = S = \frac{1}{n-1} \sum_j (\mathbf{X}_j - ave(\mathbf{X})) (\mathbf{X}_j - ave(\mathbf{X}))^H.$$

The maximum likelihood estimator is

$$\hat{\Sigma}^* = \frac{n-1}{n} S$$

|||| **Proof omitted**

1.6.3 The complex multivariate Wishart distribution

For pedagogic reasons we define the complex counterpart to the real Wishart distribution indirectly, i.e. as the distribution of the estimated dispersion matrix.

More specifically:

||| Definition 1.82

We say that a complex positive definite $p \times p$ Hermitian matrix \mathbf{W} has a *complex Wishart distribution* $W_{\mathbb{C}}(p, n, \Sigma)$ if \mathbf{W} has the same distribution as

$$\sum_{j=1}^n \mathbf{X}_j \mathbf{X}_j^H = \mathbf{X}^H \mathbf{X}$$

where the independent random variables $\mathbf{X}_j \sim N_{\mathbb{C}}(\mathbf{0}, \Sigma)$, $j = 1, \dots, n$.

||| Theorem 1.83

For a Wishart distributed, Hermitian positive definite $p \times p$ random matrix $\mathbf{W} \sim W_{\mathbb{C}}(p, n, \Sigma)$, the frequency function is

$$f(\mathbf{w}) = \frac{1}{\Gamma_p(n)} \frac{1}{(\det \Sigma)^n} (\det \mathbf{w})^{n-p} \exp(-\text{tr}(\Sigma^{-1} \mathbf{w})), \quad \mathbf{w} \text{ positive definite},$$

where

$$\Gamma_p(n) = \pi^{p(p-1)/2} \prod_{j=1}^p \Gamma(n-j+1).$$

In evaluating integrals, the volume element becomes

$$d\mathbf{w} = (dw_{11} \cdots dw_{pp}) \prod_{j,k=1, j < k}^n dw_{jk,R} dw_{jk,I}$$

where R and I denote real and imaginary part.

||| Proof omitted

We now summarize a few useful results on the complex Wishart distribution. From the definition the following theorem follows immediately.

|||| **Theorem 1.84**

Let $\mathbf{W}_1 \sim W_C(p, n_1, \Sigma)$ and $\mathbf{W}_2 \sim W_C(p, n_2, \Sigma)$ be independent. Then their sum will also be Wishart distributed

$$\mathbf{W}_1 + \mathbf{W}_2 \sim W_C(p, n_1 + n_2, \Sigma)$$

|||| **Proof**

Follows immediately from the definition.

■

The mean value of a complex Wishart distribution is also an immediate consequence of the definition. We have

|||| **Theorem 1.85**

Let $\mathbf{W} \sim W_C(p, n, \Sigma)$. Then $E(\mathbf{W}) = n\Sigma$.

|||| **Proof**

With the notation from the definition we get

$$E(\mathbf{W}) = \sum_{j=1}^n E(X_j X_j^H) = \sum_{j=1}^n D(X_j) = n\Sigma,$$

which is the desired result.

■

For the univariate case we have

|||| Theorem 1.86

The univariate complex Wishart distribution is a χ^2 -distribution or gamma distribution, i.e.

$$W_C(1, n, \sigma^2) = \frac{\sigma^2}{2} \chi^2(2n) = G(n, \sigma^2)$$

having the frequency function

$$f(x) = \frac{1}{\Gamma(n)\sigma^{2n}} x^{n-1} \exp\left(-\frac{x}{\sigma^2}\right), \quad x > 0.$$

|||| Proof

We consider independent $X_j \sim N_C(0, \sigma^2)$, $j = 1, \dots, n$, and have that $U_j = \operatorname{Re}(X_j) \sim N(0, \frac{\sigma^2}{2})$ and $V_j = \operatorname{Im}(X_j) \sim N(0, \frac{\sigma^2}{2})$ are independent. Furthermore

$$\sum_{j=1}^n X_j X_j^H = \sum_{j=1}^n (U_j + iV_j)(U_j - iV_j) = \sum_{j=1}^n (U_j^2 + V_j^2)$$

and as a sum of squares of independent normally distributed random variables the result follows. ■

|||| Theorem 1.87

We consider

$$W = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \sim W_C(p, n, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix})$$

where the dimension of W_{11} is p_1 and that of W_{22} is p_2 , $p_1 + p_2 = p$. Then the marginal distributions of W_{ii} will be Wishart distributions, i.e. $W_{ii} \sim W_C(p_i, n, \Sigma_{ii})$. If $\Sigma_{12} = \Sigma_{21}^H = \mathbf{0}$, W_{11} and W_{22} will be independent.

|||| **Proof omitted**

|||| **Corollary 1.88**

The diagonal elements W_{jj} in a complex Wishart distributed matrix $\mathbf{W} \sim W_{\mathbb{C}}(p, n, \Sigma)$ are gamma distributed $\sim G(n, \sigma_j^2)$.

|||| **Proof**

Follows immediately from Theorems 1.86 and 1.87.

■

The usual results on loss of degrees of freedom when estimating mean value parameters apply. If we e.g. consider the situation given in the previous theorem, we have

|||| **Theorem 1.89**

Let the independent observations $X_j \sim N_{\mathbb{C}}(\boldsymbol{\mu}, \Sigma)$, $j = 1, \dots, n$. Then

$$\hat{\boldsymbol{\mu}} = ave(\mathbf{X}) = \frac{1}{n} \sum_j X_j \sim N_{\mathbb{C}}\left(\boldsymbol{\mu}, \frac{1}{n}\Sigma\right)$$

and

$$\begin{aligned} (\mathbf{n} - \mathbf{1}) \hat{\Sigma} &= \\ (\mathbf{n} - \mathbf{1}) \mathbf{S} &= \\ \sum_j (X_j - ave(\mathbf{X})) (X_j - ave(\mathbf{X}))^H &\sim W_{\mathbb{C}}(p, n - 1, \Sigma) \end{aligned}$$

Furthermore, the random variables $\hat{\boldsymbol{\mu}}$ and $\mathbf{S} = \hat{\Sigma}$ are independent.

|||| **Proof omitted**

In hypothesis testing we shall also need some distributions derived from the complex Wishart distribution.

|||| **Definition 1.90**

Let the random matrices W_1 and W_2 be independent and Wishart distributed

$$W_1 \sim W_{\mathbb{C}}(p, n_1, \Sigma) \text{ and } W_2 \sim W_{\mathbb{C}}(p, n_2, \Sigma)$$

Then we define a *complex multivariate Beta distribution* $Be_{\mathbb{C}}(p, n_1, n_2)$ as the distribution of

$$W_1(W_1 + W_2)^{-1}$$

The (*Anderson's*) *complex U distribution* (or *Wilks' complex Lambda distribution*) is defined as the distribution of

$$U = \frac{\det(W_1)}{\det(W_1 + W_2)}$$

and we write

$$U \sim U_{\mathbb{C}}(p, n_1, n_2)$$

We skip proving that the distributions are independent of Σ . A useful result in finding approximations of the CDF and PDF of the U distribution is

|||| **Theorem 1.91**

Let $U \sim U_{\mathbb{C}}(p, n_1, n_2)$. Then we have

$$E(U^h) = \prod_{j=1}^p \frac{\Gamma(n_1 + h - j + 1)\Gamma(n_1 + n_2 - j + 1)}{\Gamma(n_1 - j + 1)\Gamma(n_1 + h + n_2 - j + 1)}$$

|||| **Proof omitted**

|||| Theorem 1.92

Let $\mathbf{W}_1 \sim W_C(p, n_1, \Sigma)$, $n_1 \geq p$, and $\mathbf{W}_2 \sim W_C(p, n_2, \Sigma)$, and let

$$U = \frac{\det(\mathbf{W}_1)}{\det(\mathbf{W}_1 + \mathbf{W}_2)}$$

Then the distribution of U is equal to the distribution of a product of independent Beta-distributed random variables $B_j \sim Be(n_1 - j + 1, n_2)$, $j = 1, \dots, p$. Furthermore, U and $\mathbf{W}_1 + \mathbf{W}_2$ are independent random variables.

|||| Proof omitted

|||| Theorem 1.93

Let the situation be as in the previous theorem, and let $n_2 = 1$. Then

$$U \sim U_C(p, n_1, 1) \sim Be(n_1 - p + 1, p)$$

and

$$\frac{p}{n_1 - (p - 1)} \frac{U}{1 - U} \sim F(n_1 - p + 1, p),$$

where $F(n_1 - p + 1, p)$ is Fisher's F-distribution with $(n_1 - p + 1, p)$ degrees of freedom.

|||| Proof omitted

|||| Case 1.94

In change detection studies in remote sensing, synthetic aperture radar (SAR) images may be extremely useful due to the all weather mapping capability of radar, e.g. independent of cloud cover. If also the more advanced technique, polarimetric SAR, is used, additional information about the objects on the ground will be available. In e.g. agricultural studies, the images contain useful information on the development of different crops over time since the radar backscattering is sensitive to parameters that changes during the growing period, such as the dielectric properties



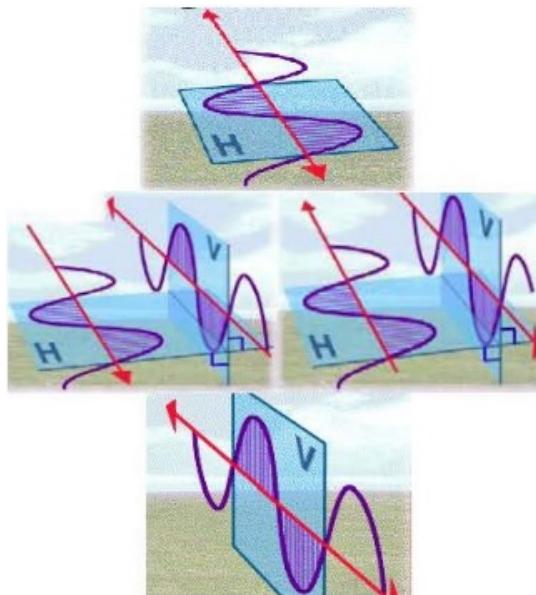
Figure 1.4 – Aerial photo of the test area around Foulum, Jutland.



Figure 1.5 – False color composite of a polarimetric, synthetic aperture radar image taken with the EMISAR radar on 17. April 1998 of the test area. Green, red, and blue are the values of the diagonal elements in the pixel values (see below), denoted HH, HV, and VV.

of the vegetation and the soil, the plant structure (i.e., the size, shape, and orientation distributions of the scatterers), the surface roughness, and the canopy structure (e.g., row direction and spacing and cover fraction), see e.g. [Conradsen et al. \(2003\)](#).

Polarimetric SAR transmits a signal $A(t) \cos(\omega t)$ and receives $kA(t) \cos(\omega t + \varphi)$ for different polarizations, i.e. the amplitude and phase of backscattered signals are measured in four combinations of the linear receive and transmit polarizations V and H, where H~Horizontal and V~Vertical. HV and VH are assumed equal due to ‘reciprocity’ of natural scenes, giving three different combinations:



First line is HH, second is HV=VH, and the last is VV. The is from <http://envisat.esa.int/polsarpro/tutorial.html>.

We assume that the measurement on a “basic” pixel no j is complex normally distributed with mean zero, i.e.

$$\mathbf{U}_j = \begin{bmatrix} S_{hh} \\ S_{hv} \\ S_{vv} \end{bmatrix} = \begin{bmatrix} amp_{hh} \cdot \exp(i \cdot phase_{hh}) \\ amp_{hv} \cdot \exp(i \cdot phase_{hv}) \\ amp_{vv} \cdot \exp(i \cdot phase_{vv}) \end{bmatrix} \sim N_{\mathbb{C}}(\mathbf{0}, \boldsymbol{\Sigma})$$

In order to reduce speckle, we estimate covariances by averaging over L pixels, $L = 2K + 1$, and obtain a so-called L-look image

$$\mathbf{C}_j = \frac{1}{L} \sum_{v=-K}^K \mathbf{U}_{j+v} \mathbf{U}_{j+v}^H = \begin{bmatrix} \langle S_{hh} S_{hh}^* \rangle & \langle S_{hh} S_{hv}^* \rangle & \langle S_{hh} S_{vv}^* \rangle \\ \langle S_{hv} S_{hh}^* \rangle & \langle S_{hv} S_{hv}^* \rangle & \langle S_{hv} S_{vv}^* \rangle \\ \langle S_{vv} S_{hh}^* \rangle & \langle S_{vv} S_{hv}^* \rangle & \langle S_{vv} S_{vv}^* \rangle \end{bmatrix} = \langle \mathbf{C} \rangle_{full}$$

The last two terms use terminology from signal processing where the bracket $\langle \rangle$ signifies averaging over observations, * indicates complex conjugate, and the subscript

'full' means that we have fully polarized data, i.e we measure all send and receive combinations. The resulting pixel value multiplied with L will – assuming independence of the involved pixel values - thus be complex Wishart distributed, i.e. $LC_j \sim W_C(3, L, \Sigma)$, where Σ is the unknown parameter characterizing the ground type of the pixel. In practice, the required independence is not met, and therefore the averaging is made over a larger window with varying weights that should ensure the desired L-look distribution, or alternatively, a smaller number of looks is used compared to the actual number of averaged pixels.

We consider a full polarimetry L-band EMISAR data from the 17. April 1998 of an area around Foulum, Jutland, see e.g [Christensen et al. \(1998\)](#) and shown in figure [1.4](#) and [1.5](#). The images have 5 m pixel spacing, 1024 lines and 1024 samples pr line. The number of looks is 13 corresponding to a theoretical distribution $W_C(3, 13, \Sigma)$. According to Corollary [1.88.](#), the diagonal elements in this distribution are gamma distributed with shape parameter 13, i.e. $\sim G(13, \sigma_j^2)$. Instead of using the terms $\langle S_{hh}S_{hh}^* \rangle$ etc., we just call the values of the diagonal elements HH , HV , and VV .

475	-97 - 51 <i>i</i>	156+197 <i>i</i>	67	-5-3 <i>i</i>	44+37 <i>i</i>	58	2+3 <i>i</i>	38+52 <i>i</i>
-97 + 51 <i>i</i>	65	-55 - 25 <i>i</i>	-5 + 3 <i>i</i>	8	1 - 7 <i>i</i>	2 - 3 <i>i</i>	6	7 + 5 <i>i</i>
156-197 <i>i</i>	-55 + 25 <i>i</i>	275	44-37 <i>i</i>	1+7· <i>i</i>	140	38-52 <i>i</i>	7-5 <i>i</i>	166
509	-78 - 49 <i>i</i>	137+146 <i>i</i>	110	-6-4 <i>i</i>	47+45 <i>i</i>	44	3+2 <i>i</i>	42+38 <i>i</i>
-78+49 <i>i</i>	102	-36-33 <i>i</i>	-6+4 <i>i</i>	9	3-1 <i>i</i>	3-2 <i>i</i>	6	10+9 <i>i</i>
137-146 <i>i</i>	-36+33 <i>i</i>	281	47-45 <i>i</i>	3+1· <i>i</i>	180	42-38 <i>i</i>	10-9 <i>i</i>	196
930	-128-31 <i>i</i>	304-14 <i>i</i>	464	-35+27 <i>i</i>	95-4 <i>i</i>	41	-6+4 <i>i</i>	12+22 <i>i</i>
-128+31 <i>i</i>	166	-77-30 <i>i</i>	-35-27 <i>i</i>	44	-11-13 <i>i</i>	-6-4 <i>i</i>	8	5+0 <i>i</i>
304+14 <i>i</i>	-77+30 <i>i</i>	346	95+4 <i>i</i>	-11+13 <i>i</i>	192	12-22 <i>i</i>	5+0 <i>i</i>	141

Pixel values ($\times 10000$) for 3×3 pixels from the EMISAR radar image. Each pixel value is a 3×3 complex, Hermitian matrix assumed to be the outcome of a complex Wishart distributed random matrix.

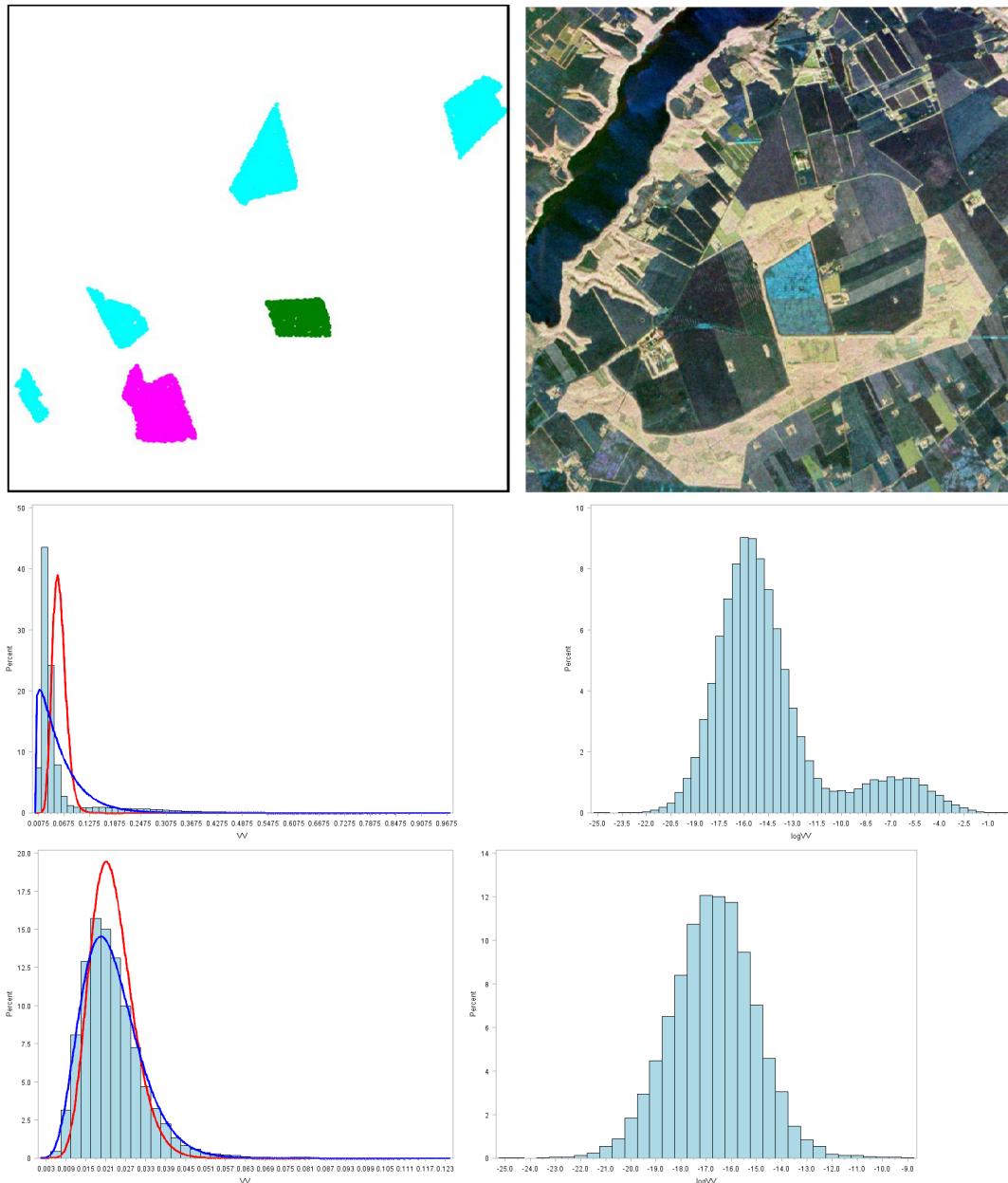
In the figure below we show the histograms for the third diagonal element VV for different grass fields, the areas colored in cyan or magenta. We assume that X is Gamma distributed $X \sim G(\alpha, \beta)$, i.e. with frequency function

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{1}{\beta}x\right), \quad x > 0,$$

and zero elsewhere. This distribution has mean $\alpha\beta$ and variance $\alpha\beta^2$. We show the maximum likelihood estimates of the parameters (α, β) and of β for fixed $\alpha = 13$ in the table below.

Parameters for fitted Gamma Distributions

Parameter	Symbol	Estimate, all pixels	Estimate, magenta pixels	Fixed shape, es- timate, all pixels	Fixed shape, estimate, magenta pixels
Scale	β	0.047171	0.003367	0.004402	0.001764
Shape	α	1.213151	6.811895	13	13
Mean		0.057225	0.022936	0.057225	0.022936
Std Dev		0.051955	0.008788	0.015871	0.006361



Row 1: Grass fields in cyan, green, and magenta. The Foulum scene where the

grass fields may be recognized. Row 2: Histogram of VV and two fitted gamma distributions for all grass pixels. In blue with estimated shape parameter, in red with shape parameter equal to 13. Histogram of 10 times $\log_{10}(VV)$ for all grass pixels. Row 3: The same as previous row but only for the “magenta” grass field.

The histograms to right show the empirical distributions of $10 \cdot \log_{10}(VV)$, where the bimodality in the case with all grass pixels is easier to see. If we threshold the distribution and only show the position of pixels causing the rightmost part of the distribution, it turns out that this part is caused by pixels from the grass field in the middle of the image (green). This field behave differently since the flight direction is 100% parallel to the plough furrows in the field. If we concentrate on e.g. the magenta field, the distribution looks very ‘regular’, and the Gamma distribution provides a very nice fit. The bimodality ‘pulls’ the distribution in the direction of the exponential that is a Gamma distribution with shape parameter 1. For the homogeneous “magenta” grass field, the shape parameter has increased to 6.8, but still only half of the ‘theoretical’ value 13. However, the best fit wit the ‘theoretical’ value 13 (shown in red) is not so bad!

The example illustrates that an inhomogeneity in the distribution of measurements in natural scenes (often) corresponds to inhomogeneities in the items measured, in the present case the multimodality corresponds to sampling from different fields.

1.7 On estimation of multidimensional parameters

In this section we present the most important results on estimation of multidimensional 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 1.95**

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

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

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

If the support of the frequency function $\{x \mid f(x; \theta) > 0\}$ does not depend on

θ , we may determine the maximum likelihood estimators by solving the *likelihood equations*

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

or

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

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

|||| Remark 1.96

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

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

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

|||| Example 1.97

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

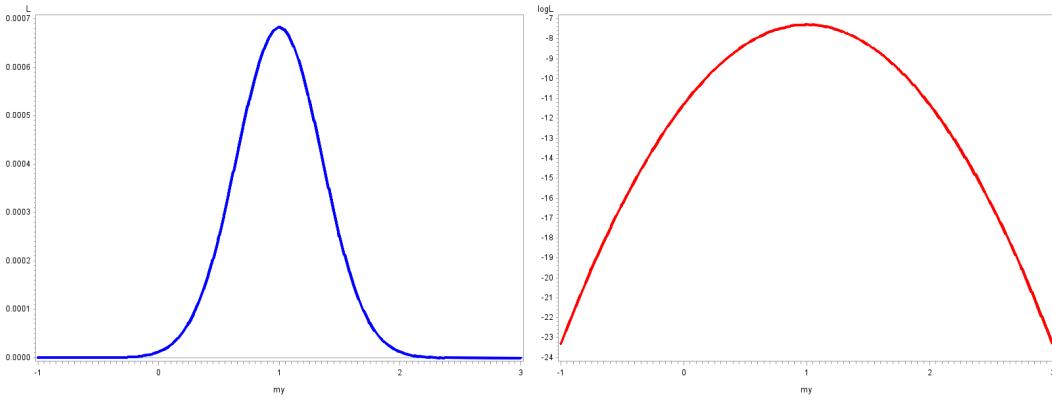
$$l_n(\mu) = \log (2\pi)^{-n/2} - \frac{1}{2} \sum_{i=1}^n (x_i - \mu)^2$$

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 1.98

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

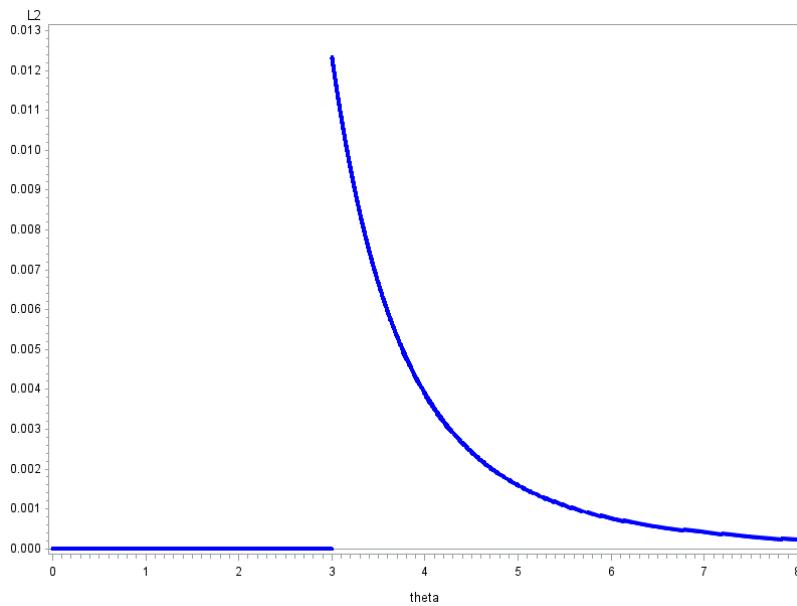
$$L(\theta) = \prod_{i=1}^n \frac{1}{\theta} I_{[0,\theta]}(x_i) = \theta^{-n} \prod_{i=1}^n I_{[0,\theta]}(x_i)$$

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 1.99

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 1.100

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 1.101

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 1.102

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 1.103**

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 1.104**

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 1.105**

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 1.106**

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

$$Y = x\theta + z\gamma + \varepsilon$$

where x and z are known (design) matrices, θ a vector of fixed (constant, unknown) parameters, and γ a vector of unknown random variables called random effects parameters. ε is the error term. We assume that γ and ε are normally distributed with

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

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

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

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 1.107

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

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

Taking the log we obtain

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

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 $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is assumed known. Consider the norm given by $\boldsymbol{\Sigma}^{-1}$ i.e.

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

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

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

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 \mathbb{R}^n . M is the image of \mathbb{R}^k corresponding to the linear projection \mathbf{x} . The dimension of M is $\text{rk}(\mathbf{x}) \leq k$. The situation is depicted in the following figure.

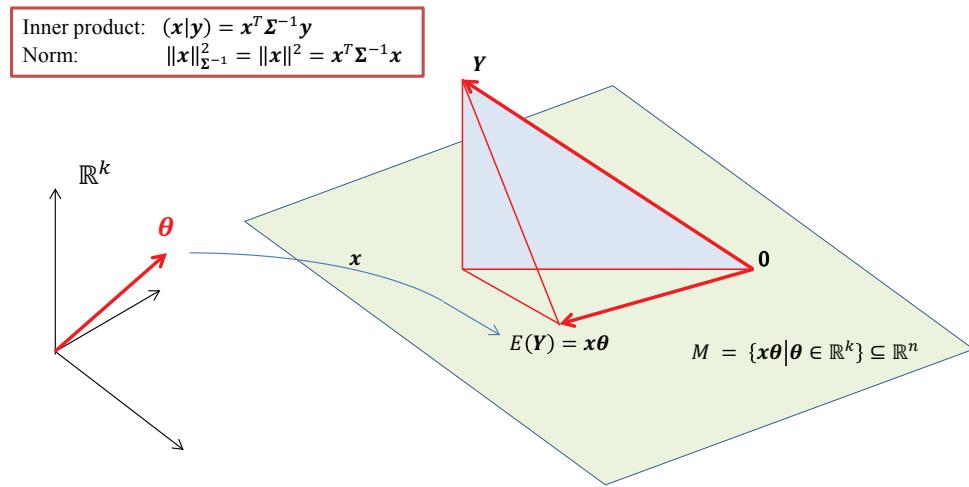


Figure 2.1 – Geometrical sketch of the general linear model.

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.

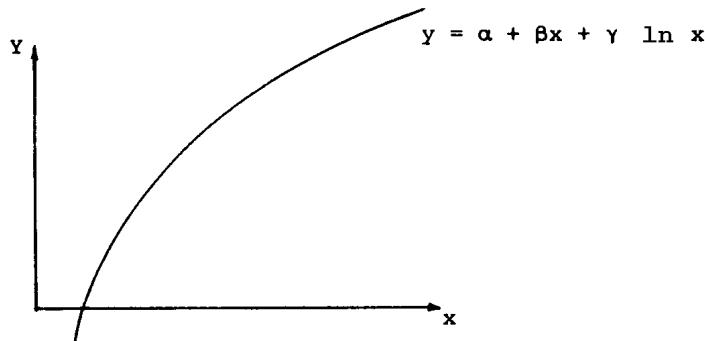
Another example is

||| Example 2.2

We now consider a situation, where

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

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



Even in this case we have a linear model which is

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

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

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

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 $Y \sim N_n(x\theta, \sigma^2\Sigma)$, 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)*

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

If x has full rank k , then

$$\hat{\theta} = (x^T \Sigma^{-1} x)^{-1} x^T \Sigma^{-1} Y,$$

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

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

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

|||| Proof

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

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

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

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

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

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

From the result p. 478 the value of $\mathbf{x}\boldsymbol{\theta}$, giving the minimum is equal to the orthogonal projection (with respect to $\boldsymbol{\Sigma}^{-1}$) of \mathbf{y} on M . From example A.50 p. 475 the optimal $\boldsymbol{\theta}$ is the solution to the equation

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

If $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ has full rank k , i.e. if \mathbf{x} has rank k (cf. p. 462) we therefore have

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

We have now shown the first half of the theorem.

From theorem 1.2 we find that

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

And from theorem 1.6 we find

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

■

The situation is illustrated in the following figure 2.2.

|||| Remark 2.4

We note that $\boldsymbol{\theta}$ is estimated by minimising the squared distance onto M . $\hat{\boldsymbol{\theta}}$ is therefore also a *least squares estimate* of $\boldsymbol{\theta}$. If we do not have the distributional assumption we will often be able to use the estimator $\hat{\boldsymbol{\theta}}$ in theorem 2.3 as an estimate of $\boldsymbol{\theta}$. It can be shown that the least squares estimator $\hat{\boldsymbol{\theta}}$ has the least generalised variance among all the estimators that are linear functions of the observations (the so-called *Gauss-Markov theorem*) cf. Kendall and Stuart (1967). 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

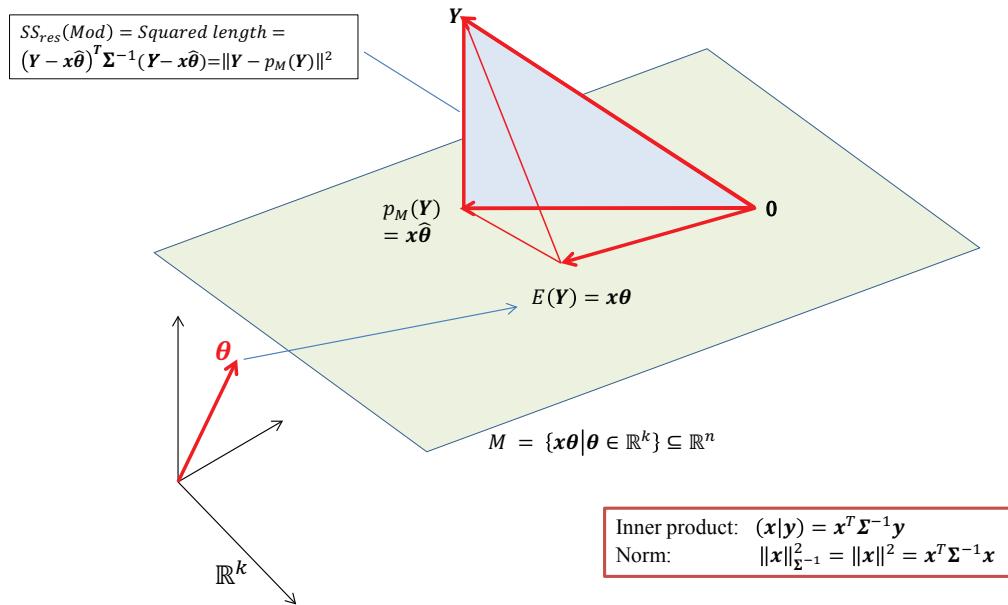


Figure 2.2 – Geometric sketch of the problem of estimation in the general linear model.

|||| Theorem 2.5

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

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

The unbiased estimator of σ^2 is

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

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

$$\hat{\sigma}^2 \sim \sigma^2 \chi^2(n - \text{rk}(\mathbf{x})) / (n - \text{rk}(\mathbf{x}))$$

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

||| Proof

The likelihood function is

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

and

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

now

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

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

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

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

$$P_M(\mathbf{Y} - \mathbf{x}\theta) = \mathbf{x}\hat{\theta} - \mathbf{x}\theta$$

and

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

are stochastically independent and that

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

From this we especially get

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

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

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

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

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

where $\hat{E}(Y_i) = (\mathbf{x} \hat{\theta})_i$.

||| Definition 2.7

The deviation

$$R_i = Y_i - \hat{E}(Y_i) = Y_i - (\mathbf{x} \hat{\theta})_i$$

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

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

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

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

and

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

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

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

Alternatively we may compute

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

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 $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ singular

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

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

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

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

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 \mathbb{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

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

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

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

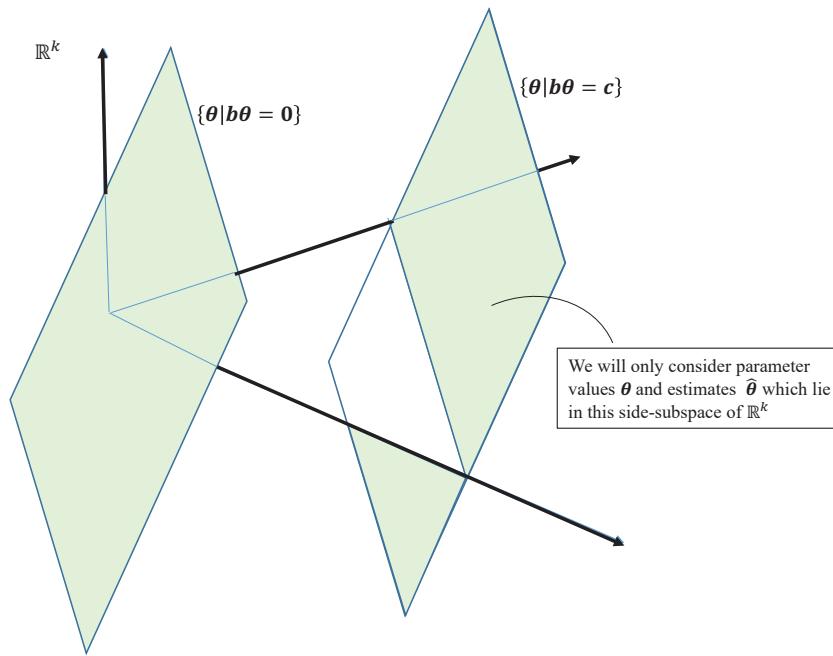


Figure 2.3 – Illustration of side-subspace

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

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

where the short notation should not cause confusion.

If we in the usual way compute

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

then we have a quantity which minimises

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

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

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

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

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

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

This expression is found immediately by using theorem 1.6.

As before the unbiased estimate of σ^2 is

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

Here we have $n - \text{rk}(x) = n - k + \text{rk}(b)$.

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

$$\begin{aligned} E(Y_{11}) &= E(Y_{12}) = \theta_1 \\ E(Y_{21}) &= E(Y_{22}) = \theta_2, \end{aligned}$$

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

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

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

$$\begin{aligned} E(Y_{11}) &= E(Y_{12}) = \mu + \alpha_1 \\ E(Y_{21}) &= E(Y_{22}) = \mu + \alpha_2 \end{aligned}$$

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

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

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

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

We now have that

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

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 $\mathbf{x}^T \Sigma^{-1} \mathbf{x}$ 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	

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

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

These are assumed to have the mean values

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

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

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

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

We find

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

and

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

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

Since $\mathbf{x}^T \mathbf{x}$ only has the rank 3, we are unable to invert it. Instead we can find a pseudo-inverse. We use the theorem A.18 p. 453 and get

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

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

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

where e.g.

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

Now, since

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

we have

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

From theorem A.16 the complete solution to the normal equations is therefore all vectors of the form

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

An arbitrary maximum likelihood estimator for θ is then of this form.

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

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

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

$$\text{rk}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}\right) = 4,$$

so we can use the result from p. 111. We find

$$\begin{aligned} \mathbf{x}^T \mathbf{x} + \mathbf{b}^T \mathbf{b} &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 3 & 1 & 0 \\ 2 & 1 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}. \end{aligned}$$

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

$$(\mathbf{x}^T \mathbf{x} + \mathbf{b}^T \mathbf{b})^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

We now get

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T \mathbf{x} + \mathbf{b}^T \mathbf{b})^{-1} \mathbf{x}^T \mathbf{y} = \begin{bmatrix} \bar{y}_{1..} \\ \bar{y}_{11..} - \bar{y}_{1..} \\ \bar{y}_{12..} - \bar{y}_{1..} \\ \bar{y}_{21..} \end{bmatrix}.$$

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

$$\begin{aligned} D(\hat{\theta}) &= \sigma^2 (\mathbf{x}^T \mathbf{x} + \mathbf{b}^T \mathbf{b})^{-1} \mathbf{x}^T \mathbf{x} (\mathbf{x}^T \mathbf{x} + \mathbf{b}^T \mathbf{b})^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}, \end{aligned}$$

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

$$\|\mathbf{y} - \mathbf{x} \hat{\theta}\|^2 = (\mathbf{y} - \mathbf{x} \hat{\theta})^T (\mathbf{y} - \mathbf{x} \hat{\theta}) = 1^2 + \dots + (\frac{2}{3})^2 = 28\frac{2}{3}.$$

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

$$\mathbf{H}^T \boldsymbol{\theta} = \xi.$$

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 θ is expressed in

|||| Theorem 2.11

Let $E(Y) = x\theta$ where Y is an n -dimensional random variable, x a known $n \times k$ matrix and θ a k -dimensional vector of unknown parameters satisfying the s linear constraints

$$H^T\theta = \xi,$$

where H is a known $k \times s$ matrix and ξ a known s -dimensional vector. Finally we suppose that $D(Y) = \sigma^2\Sigma$ where Σ is known. The least squares estimator $\tilde{\theta}$ for θ under the constraint $H^T\theta = \xi$ is a solution to the equations

$$\begin{bmatrix} x^T\Sigma^{-1}x & H \\ H^T & 0 \end{bmatrix} \begin{bmatrix} \theta \\ \lambda \end{bmatrix} = \begin{bmatrix} x^T\Sigma^{-1}y \\ \xi \end{bmatrix}.$$

|||| Proof

We must determine a θ that minimizes

$$\min_{H^T\theta=\xi} (Y - x\theta)^T\Sigma^{-1}(Y - x\theta).$$

we introduce the Lagrange multiplier λ and put

$$F(\theta, \lambda) = \frac{1}{2}(Y - x\theta)^T\Sigma^{-1}(Y - x\theta) + \lambda^T(H^T\theta - \xi).$$

Then

$$\begin{aligned} \frac{\partial F}{\partial \theta} &= -x^T\Sigma^{-1}y + x^T\Sigma^{-1}x\theta + H\lambda \\ \frac{\partial F}{\partial \lambda} &= H^T\theta - \xi. \end{aligned}$$

Those two derivatives are 0 in any extremum for $(Y - x\theta)^T\Sigma^{-1}(Y - x\theta)$ under the constraint $H^T\theta = \xi$. Using this, the result in the theorem follows immediately. ■

Next we consider the problem of estimating σ^2 in

||| Theorem 2.12

Letting

$$\begin{bmatrix} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} & \mathbf{H} \\ \mathbf{H}^T & \mathbf{0} \end{bmatrix}^- = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & \mathbf{C}_4 \end{bmatrix}.$$

be a pseudoinverse to the coefficient matrix in Theorem 2.11. Then

$$D(\tilde{\boldsymbol{\theta}}) = \sigma^2 \mathbf{C}_1,$$

and an unbiased estimation of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{f} (\mathbf{Y}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} - \tilde{\boldsymbol{\theta}}^T \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} - \boldsymbol{\xi}^T \tilde{\boldsymbol{\lambda}}),$$

where $(\tilde{\boldsymbol{\theta}}^T, \tilde{\boldsymbol{\lambda}}^T)^T$ is a solution to the equations in Theorem 2.11, and

$$f = n - \text{rk}(\mathbf{x}^T, \mathbf{H}) + \text{rk}(\mathbf{H}).$$

||| Proof

By introducing the pseudoinverse we get

$$\begin{aligned} \tilde{\boldsymbol{\theta}} &= \mathbf{C}_1 \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} + \mathbf{C}_2 \boldsymbol{\xi} \\ \tilde{\boldsymbol{\lambda}} &= \mathbf{C}_3 \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} + \mathbf{C}_4 \boldsymbol{\xi}. \end{aligned}$$

From this we immediately obtain

$$\begin{aligned} D(\tilde{\boldsymbol{\theta}}) &= \sigma^2 \mathbf{C}_1 \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} \mathbf{x} \mathbf{C}_1^T \\ &= \sigma^2 \mathbf{C}_1 \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \mathbf{C}_1^T \\ &= \sigma^2 \mathbf{C}_1 \end{aligned}$$

The last equality sign follows by using properties of pseudoinverse matrices.

By plugging in it is seen that

$$(\mathbf{Y} - \mathbf{x} \tilde{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x} \tilde{\boldsymbol{\theta}}) = (\mathbf{Y}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} - \tilde{\boldsymbol{\theta}}^T \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y} - \boldsymbol{\xi}^T \tilde{\boldsymbol{\lambda}}),$$

so now it just remains to be shown that the degrees of freedom are as postulated in the theorem. The solution to

$$\mathbf{H}^T \boldsymbol{\theta} = \boldsymbol{\xi},$$

can be written as

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \mathbf{B} \boldsymbol{\beta},$$

where $\boldsymbol{\theta}_0$ is a particular solution and \mathbf{B} is a $(k \times s)$ matrix ($\text{rk}(\mathbf{H}) = k - s$) satisfying

$$\mathbf{H}^T \mathbf{B} = \mathbf{0}.$$

Finally $\boldsymbol{\beta}$ is a s -dimensional vector of "free", new parameters. If we consider

$$\mathbf{Z} = \mathbf{Y} - \mathbf{x} \boldsymbol{\theta}_0,$$

we get

$$\begin{aligned} E(\mathbf{Z}) &= \mathbf{x} \boldsymbol{\theta} - \mathbf{x} \boldsymbol{\theta}_0 = \mathbf{x}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= \mathbf{x}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= \mathbf{x} \mathbf{B} \boldsymbol{\beta}. \end{aligned}$$

We may now consider the model

$$\mathbf{Z} = \mathbf{x} \mathbf{B} \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the error vector and solve this. By doing this we obtain the earlier stated estimates.

Letting

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^T \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \mathbf{B})^{-1} \mathbf{B}^T \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y},$$

we obtain

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \mathbf{B} \hat{\boldsymbol{\beta}},$$

and consequently

$$\begin{aligned} \mathbf{Y} - \mathbf{x} \tilde{\boldsymbol{\theta}} &= \mathbf{Z} + \mathbf{x} \boldsymbol{\theta}_0 - \mathbf{x} \boldsymbol{\theta}_0 - \mathbf{x} \mathbf{B} \hat{\boldsymbol{\beta}} \\ &= \mathbf{Z} - \mathbf{x} \mathbf{B} \hat{\boldsymbol{\beta}}. \end{aligned}$$

From the general theory it follows that the degrees of freedom are $n - \text{rk}(\mathbf{x} \mathbf{B})$. We have

$$\begin{aligned} \text{rk}(\mathbf{x} \mathbf{B}) &= \dim\{\mathbf{x} \mathbf{B} \boldsymbol{\beta} | \boldsymbol{\beta} \sim \mathbb{R}^s\} \\ &= \dim\{\mathbf{x} \boldsymbol{\gamma} | \mathbf{H}^T \boldsymbol{\gamma} = \mathbf{0}, \boldsymbol{\gamma} \sim \mathbb{R}^m\} \\ &= \text{rk}\left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}^T \end{array}\right) - \text{rk}(\mathbf{H}). \end{aligned}$$

The last equality sign follows from the relation

$$\dim S_1^* + \dim S_2^* = \text{rk}\left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}^T \end{array}\right),$$

where

$$\begin{aligned} S_1^* &= \left\{ \left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}^T \end{array} \right) \boldsymbol{\gamma} \mid \boldsymbol{\gamma} \sim N(H) \right\} \\ S_2^* &= \left\{ \left(\begin{array}{c} \mathbf{x} \\ \mathbf{H}^T \end{array} \right) \boldsymbol{\gamma} \mid \boldsymbol{\gamma} \sim N(H)^\perp \right\} \end{aligned}$$

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

$$\left[\begin{array}{cc|c} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} & \mathbf{H} & \\ \mathbf{H}^T & \mathbf{0} & \end{array} \right] = \left[\begin{array}{ccc|c} 2 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ \hline 1 & 1 & 1 & 0 \end{array} \right].$$

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

$$\begin{aligned}\tilde{\theta} &= \frac{1}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 106 \\ 148 \\ 94 \end{bmatrix} + \frac{1}{6} \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} 180 \\ &= \begin{bmatrix} -5 \\ 16 \\ -11 \end{bmatrix} + \begin{bmatrix} 60 \\ 60 \\ 60 \end{bmatrix} \\ &= \begin{bmatrix} 55 \\ 76 \\ 49 \end{bmatrix}.\end{aligned}$$

We observe that the sum of the estimates is 180.

The dispersion matrix is

$$D(\tilde{\theta}) = \sigma^2 C_1 = \frac{\sigma^2}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

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

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \Sigma).$$

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

$$\hat{\theta} = (\mathbf{x}^T \Sigma^{-1} \mathbf{x})^{-1} \mathbf{x}^T \Sigma^{-1} \mathbf{Y}$$

for θ , and σ^2 is estimated using

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

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:

$$(z_1, \dots, z_k) = \mathbf{z}^T$$

i.e. a new row in the \mathbf{x} -matrix, it is obvious that we will use

$$U = (z_1, \dots, z_k) \begin{bmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_k \end{bmatrix} = \mathbf{z}^T \hat{\theta}$$

as our predictor.

We have that $E(U) = E(Y)$ and that

$$\begin{aligned} V(U) &= \mathbf{z}^T D(\hat{\theta}) \mathbf{z} \\ &= \sigma^2 \mathbf{z}^T (\mathbf{x}^T \Sigma^{-1} \mathbf{x})^{-1} \mathbf{z} \\ &= \sigma^2 c, \end{aligned}$$

where

$$c = (z_1, \dots, z_k) (\mathbf{x}^T \Sigma^{-1} \mathbf{x})^{-1} \begin{bmatrix} z_1 \\ \vdots \\ z_k \end{bmatrix}.$$

We therefore immediately have

$$\frac{U - E(Y)}{\sigma\sqrt{c}} \sim N(0, 1),$$

and therefore also

$$\frac{U - E(Y)}{S\sqrt{c}} \sim t(n - k).$$

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.

■

|||| Remark 2.16

Often we have a situation where we seek the confidence interval of an existing observation x_i , where it is further the case that Σ is the identity matrix. In this case $c = h_{ii}$, see theorem 2.26 and section 3.2.

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 Y_q denoting a new (or coming) observation taken at (z_1, \dots, z_k) . If $Y_q \sim N(E(Y), c_1\sigma^2)$, and

if we now assume that the new (or future) observation is independent of those we already have then

$$U - Y_q \sim N(0, \sigma^2(c + c_1)),$$

i.e.

$$\frac{U - Y_q}{S\sqrt{c + c_1}} \sim t(n - k).$$

From this we can as before derive

|||| Theorem 2.17

Let us assume that a new observation taken at (z_1, \dots, z_k) has a variance $c_1\sigma^2$. Furthermore, it is independent of the earlier observations. In that case a $(1 - \alpha)$ -*prediction interval* for the new observation equals the interval

$$[u - t(n - k)_{1-\frac{\alpha}{2}} s\sqrt{c + c_1}, u + t(n - k)_{1-\frac{\alpha}{2}} s\sqrt{c + c_1}].$$

|||| Remark 2.18

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 2.19

We see that the correspondence to the interval for Y_q instead of the interval for $E(Y_q) = E(Y)$ just consists of the expression under the square root sign being larger by an amount equal to c_1 which is the variance of $\frac{Y_q}{\sigma}$.

|||| Example 2.20

We consider the following corresponding observations of an independent variable x and a dependent variable y :

x	1	2	3	4	5	6
y	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_6 which are normally distributed with mean values

$$E(Y|x) = \beta x^2$$

and variances

$$V(Y|x) = x^2\sigma^2$$

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:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \cdot \\ \vdots \\ \cdot \\ \varepsilon_6 \end{bmatrix} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$D(\boldsymbol{\varepsilon}) = \sigma^2 \begin{bmatrix} 1 & \cdot & \cdot & \cdot & \cdot & 0 \\ \cdot & 4 & & & & \cdot \\ \cdot & & 9 & & & \cdot \\ \cdot & & & 16 & & \cdot \\ \cdot & & & & 25 & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & 36 \end{bmatrix} = \sigma^2 \boldsymbol{\Sigma}.$$

We have that

$$\begin{aligned} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} &= (1, 4, 9, 16, 25, 36) \text{diag}(1, \frac{1}{4}, \dots, \frac{1}{36}) \begin{bmatrix} 1 \\ \vdots \\ 36 \end{bmatrix} \\ &= 91. \\ \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} &= 0.3 + 1.5 + 1.3 + 1.9 + 4.2 + 8.0 = 17.2. \end{aligned}$$

so

$$\hat{\beta} = \frac{17.2}{91} = 0.1890,$$

and

$$P_M(\mathbf{y}) = \begin{bmatrix} 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \cdot 0.1890 = \begin{bmatrix} 0.1890 \\ 0.7560 \\ 1.7010 \\ 3.0240 \\ 4.7250 \\ 6.8040 \end{bmatrix}.$$

The residuals are

$$\mathbf{y} - P_M(\mathbf{y}) = \begin{bmatrix} 0.1110 \\ 0.7440 \\ -0.4010 \\ -1.1240 \\ -0.5250 \\ 1.1960 \end{bmatrix},$$

so

$$\begin{aligned} \|\mathbf{y} - P_M(\mathbf{y})\|^2 &= (0.1110 \cdots 1.1960) \begin{bmatrix} \frac{1}{1} & & & \\ & \ddots & & \\ & & \frac{1}{36} & \\ & & & \end{bmatrix} \begin{bmatrix} 0.1110 \\ \vdots \\ 1.1960 \end{bmatrix} \\ &= 0.2983 \end{aligned}$$

i.e.

$$\hat{\sigma}^2 = s^2 = \frac{1}{6-1} 0.2983 = 0.0597 = 0.2443^2.$$

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

$$z = 100\hat{\beta} = 18.90$$

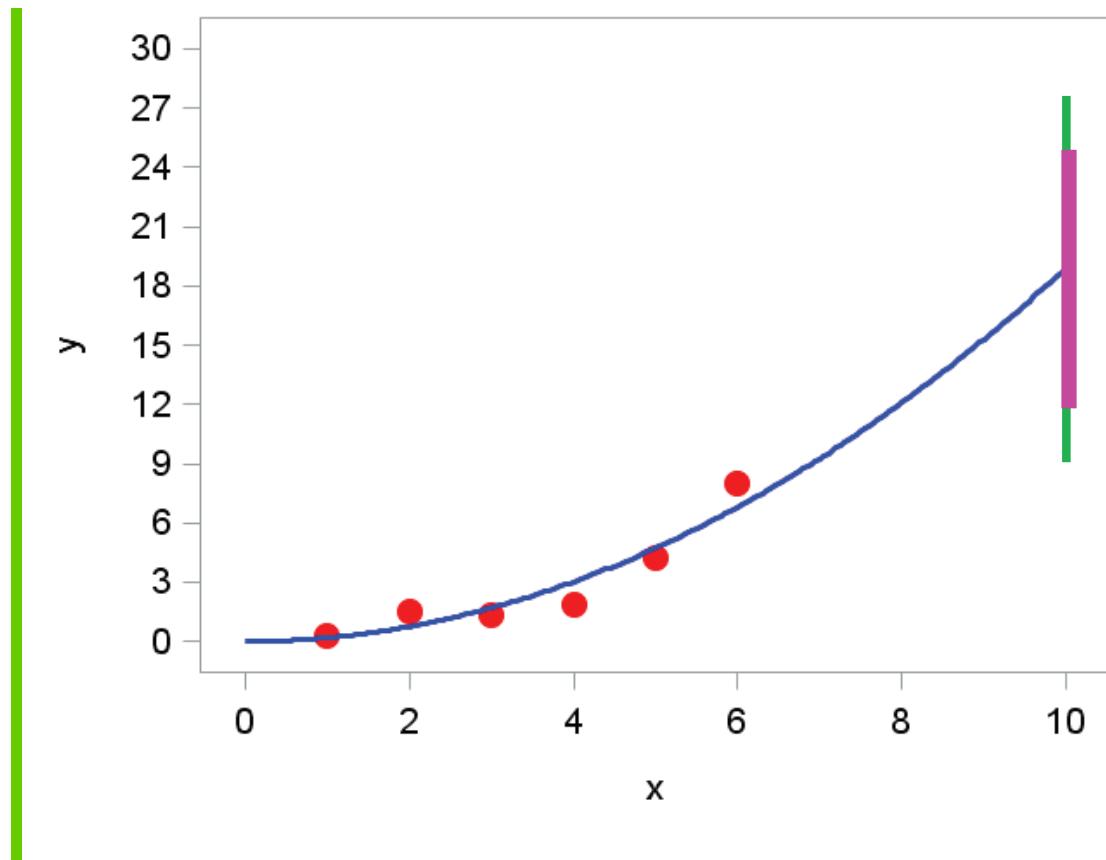
The confidence interval for the expected value at $x = 10$ is therefore given by

$$\begin{aligned} &18.90 \pm t(5)_{0.975} 0.2443 \sqrt{109.89} \\ &= 18.90 \pm 2.571 \cdot 0.2443 \sqrt{109.89} \\ &= 18.90 \pm 6.58. \end{aligned}$$

The corresponding prediction interval for the next observation is

$$\begin{aligned} &18.90 \pm t(5)_{0.975} \cdot 0.2443 \sqrt{109.89 + 100} \\ &= 18.90 \pm 9.10, \end{aligned}$$

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 confidence and prediction 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 $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$, 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 $p_M(\mathbf{Y})$, 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 $\frac{1}{n}\|\mathbf{y} - p_H(\mathbf{y})\|^2$. This is illustrated in figure 2.4 together with the test-statistic.

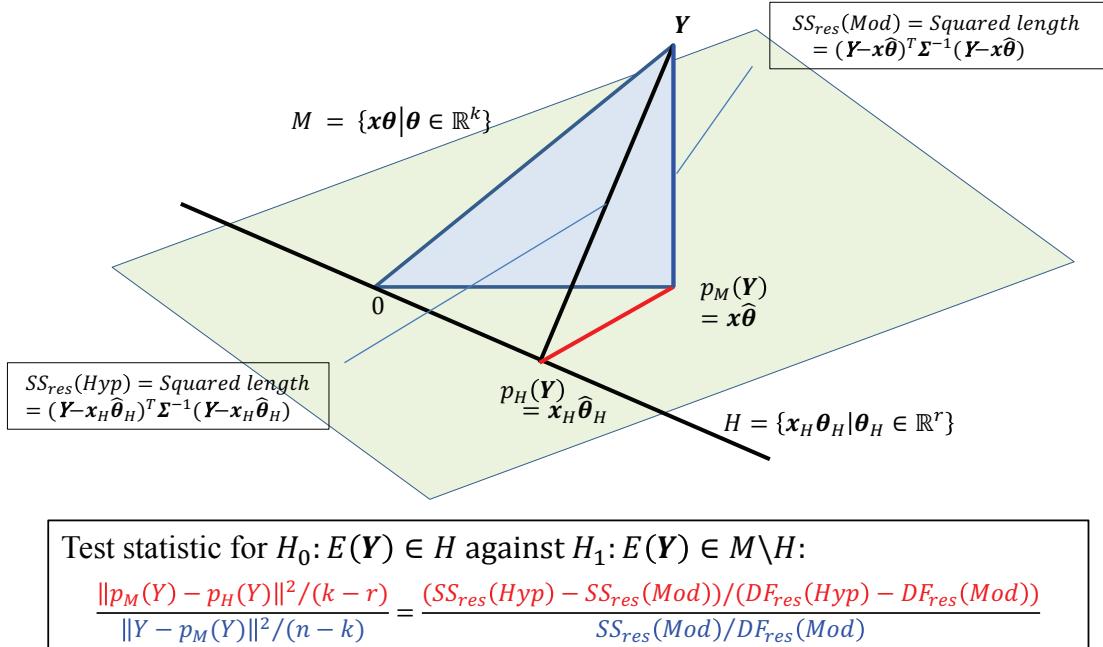


Figure 2.4

The likelihood function is

$$\begin{aligned} L(\boldsymbol{\mu}, \sigma^2) &= \frac{1}{\sqrt{2\pi}^n} \frac{1}{\sigma^n} \frac{1}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})\right) \\ &= k \cdot \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{y} - \boldsymbol{\mu}\|^2\right). \end{aligned}$$

With this notation we have

||| Theorem 2.21

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

$$C_\alpha = \{(\mathbf{y}_1, \dots, \mathbf{y}_n) \mid \frac{\|p_M(\mathbf{y}) - p_H(\mathbf{y})\|^2/(k-r)}{\|\mathbf{y} - p_M(\mathbf{y})\|^2/(n-k)} > F(k-r, n-k)_{1-\alpha}\}.$$

|||| 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

$$Q < q \Leftrightarrow \frac{\|\mathbf{y} - p_M(\mathbf{y})\|^2}{\|\mathbf{y} - p_H(\mathbf{y})\|^2} < k_1.$$

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

$$Q < q \Leftrightarrow \frac{\|p_M(\mathbf{y}) - p_H(\mathbf{y})\|^2 / (k - r)}{\|\mathbf{y} - p_M(\mathbf{y})\|^2 / (n - k)} > c. \quad (2-1)$$

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 2.22**

Often one will be in the situation that the sub-spaces M and H are parameterised, i.e.

$$\begin{aligned}\boldsymbol{\mu} \in M &\Leftrightarrow \exists \boldsymbol{\theta} \in \mathbb{R}^k (\boldsymbol{\mu} = \mathbf{x}\boldsymbol{\theta}) \\ \boldsymbol{\mu} \in H &\Leftrightarrow \exists \boldsymbol{\gamma} \in \mathbb{R}^r (\boldsymbol{\mu} = \mathbf{x}_0\boldsymbol{\gamma}),\end{aligned}$$

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

$$\begin{aligned}(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})\hat{\boldsymbol{\theta}} &= \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{y} \\ (\mathbf{x}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_0)\hat{\boldsymbol{\gamma}} &= \mathbf{x}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{y}\end{aligned}$$

with respect to $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$.

Sometimes we may arrive at a bit simpler at the test statistic. Let the hypothesis be of the form that one of the coordinates, say no i_0 , in

$$\boldsymbol{\theta} = [\theta_1 \ \dots \ \theta_{i_0} \ \dots \ \theta_k]^T$$

assumes a specific value c (often = 0), i.e. we are testing

$$H_0 : \theta_{i_0} = c \text{ against } H_1 : \theta_{i_0} \neq c$$

Since $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1})$, we have

$$\hat{\theta}_{i_0} \sim N \left(\theta_{i_0}, \sigma^2 \left\{ \left(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \right)^{-1} \right\}_{ii} \right) = N \left(\theta_{i_0}, V(\hat{\theta}_{i_0}) \right)$$

Furthermore

$$\hat{\sigma}^2 = \frac{1}{f} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) \sim \sigma^2 \chi^2(f)/f, \quad f = n - \text{rk}(\mathbf{x})$$

and thus also

$$\hat{V}(\hat{\theta}_{i_0}) = \hat{\sigma}^2 \left\{ \left(\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \right)^{-1} \right\}_{ii}$$

are independent of $\hat{\theta}_{i_0}$. Therefore we get the t-distributed random variable

$$\frac{\hat{\theta}_{i_0} - E(\hat{\theta}_{i_0})}{\sqrt{\hat{V}(\hat{\theta}_{i_0})}} \sim t(f), \quad f = n - \text{rk}(\mathbf{x})$$

If the hypothesis H_0 is true, we thus have

$$\frac{\hat{\theta}_{i_0} - c}{\sqrt{\hat{V}(\hat{\theta}_{i_0})}} \sim t(f), \quad f = n - \text{rk}(\mathbf{x})$$

This is used in setting up a test for the hypothesis in

|||| Theorem 2.23

Let the situation be as above. Then the critical region for testing H_0 against H_1 at significance level α is

$$C_\alpha = \left\{ (y_1, \dots, y_n) \mid \hat{\theta}_{i_0} < c - t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})} \text{ or } \hat{\theta}_{i_0} > c + t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})} \right\}$$

|||| Proof

An immediate consequence of the above.



|||| **Remark 2.24**

The estimated standard deviation $(\hat{V}(\hat{\theta}_{i_0}))^{1/2}$ of $\hat{\theta}_{i_0}$ is often provided by standard software as '*standard error of estimate*' or similar. It is thus straight forward to compute the critical limits. This result may also be used in setting up confidence limits for θ_{i_0} . More specifically, a $(1 - \alpha)$ confidence interval becomes

$$\left[\hat{\theta}_{i_0} - t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})}, \quad \hat{\theta}_{i_0} + t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})} \right]$$

|||| **Remark 2.25**

The test given in Theorem 2.23 is of course equivalent to the ordinary F-test presented in Theorem 2.21 for $k - r = 1$. This may be established directly using the fact that the square of a t-distributed random variable is F-distributed, mnemonically written

$$[t(f)]^2 = F(1, f).$$

The advantage by using Theorem 2.23 is that we with obvious modifications may test one-sided hypothesis like $H_0 : \theta_{i_0} \leq c$ against $H_1 : \theta_{i_0} > c$. This is not possible with the F-test.

We now state a useful theorem and corollary used in residual analysis in the general linear model. More precisely we show that in the not full rank case, the choice of generalized inverse will not have any influence on the residuals. For notational reasons and since we shall mostly consider the case where $\Sigma = I$, i.e. where $D(Y) = \sigma^2 \Sigma = \sigma^2 I$, we will assume this in the following theorem.

|||| Theorem 2.26

Let \mathbf{x} be an $n \times k$ matrix not necessarily of full rank. Then the socalled *hat matrix*

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

is independent of the choice of the generalised inverse $(\mathbf{x}^T \mathbf{x})^{-1}$. Furthermore it is idempotent and symmetric, and

$$\text{rk}(\mathbf{H}) = \text{tr}(\mathbf{H}) = \text{rk}(\mathbf{x})$$

The matrix \mathbf{M} is given by

$$\mathbf{M} = \mathbf{I} - \mathbf{H}$$

See also section 3.1.2.

|||| Proof

See Graybill (1976) (p. 32). If $(\mathbf{x}^T \mathbf{x})$ 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

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

projects on the orthogonal complement, i.e.

$$\mathbf{MH} = \mathbf{0}$$

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

$$D(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{HH}^T = \sigma^2 \mathbf{H}.$$

and

$$D(\mathbf{R}) = \sigma^2 \mathbf{MM}^T = \sigma^2 \mathbf{M}.$$

We have now proven

|||| Corollary 2.27

Let the situation be as above assuming that the design matrix is not of full rank. Then the predicted values, the residuals and thus the residual sum of squares will be independent of the choice of generalized inverse used in solving the normal equations.

|||| Remark 2.28

Using Pythagoras' theorem we see that there are two other ways of computing

$$\|p_M(\mathbf{Y}) - p_H(\mathbf{Y})\|^2 = (\mathbf{x}\hat{\boldsymbol{\theta}} - \mathbf{x}_0\hat{\boldsymbol{\gamma}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}\hat{\boldsymbol{\theta}} - \mathbf{x}_0\hat{\boldsymbol{\gamma}}) \quad (2-2)$$

besides the direct formula, namely as

$$\|p_M(\mathbf{Y})\|^2 - \|p_H(\mathbf{Y})\|^2 = (\mathbf{x}\hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}\hat{\boldsymbol{\theta}}) - (\mathbf{x}_0\hat{\boldsymbol{\gamma}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_0\hat{\boldsymbol{\gamma}}) \quad (2-3)$$

or as

$$\begin{aligned} & \|\mathbf{Y} - p_H(\mathbf{Y})\|^2 - \|\mathbf{Y} - p_M(\mathbf{Y})\|^2 \\ &= (\mathbf{Y} - \mathbf{x}_0\hat{\mathbf{Y}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x}_0\hat{\mathbf{Y}}) - (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) \end{aligned} \quad (2-4)$$

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 2.29 GLM without intercept

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

$$\|\mathbf{z}\|^2 = \mathbf{z}^T \mathbf{z} = \sum_{i=1}^n z_i^2$$

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

$$\begin{aligned} \text{SS(Model)} &= \|p_M(\mathbf{Y})\|^2 = (\mathbf{x}\hat{\theta})^T(\mathbf{x}\hat{\theta}) = \mathbf{Y}^T \mathbf{H} \mathbf{Y} \\ \text{SSRes(Model)} &= \|\mathbf{Y} - p_M(\mathbf{Y})\|^2 = (\mathbf{Y} - \mathbf{x}\hat{\theta})^T(\mathbf{Y} - \mathbf{x}\hat{\theta}) = \mathbf{Y}^T \mathbf{M} \mathbf{Y} \\ \text{SSTot(Uncorrected)} &= \|\mathbf{Y}\|^2 = \mathbf{Y}^T \mathbf{Y} = \sum_{i=1}^n Y_i^2 \end{aligned}$$

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).

Once again we consider the model from Example 2.8 (p. 107).

||| Example 2.30

We have the model

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_A \\ \theta_B \end{bmatrix} + \boldsymbol{\epsilon}.$$

We observe data where $y^T = (90, 30, 75)$. We wish to test the hypothesis

$$H_0 : \theta_B = 0 \quad \text{versus} \quad H_1 : \theta_B \neq 0.$$

We reformulate the hypothesis into

$$H_0 : E(\mathbf{Y}) = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \theta_A = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \gamma.$$

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

$$\|\mathbf{y} - \mathbf{x}_0 \hat{\gamma}\|^2 = (\mathbf{y} - \mathbf{x}_0 \hat{\gamma})^T (\mathbf{y} - \mathbf{x}_0 \hat{\gamma}) = 1620.$$

and since we had (p. 108)

$$\|\mathbf{y} - \mathbf{x} \hat{\theta}\|^2 = (\mathbf{y} - \mathbf{x} \hat{\theta})^T (\mathbf{y} - \mathbf{x} \hat{\theta}) = 150,$$

we get

$$\|\mathbf{x} \hat{\theta} - \mathbf{x}_0 \hat{\gamma}\|^2 = 1620 - 150 = 1470.$$

We may also compute this quantity as

$$\begin{aligned}\|\hat{\boldsymbol{\theta}}\|^2 - \|\mathbf{x}_0\hat{\boldsymbol{\gamma}}\|^2 &= (\hat{\boldsymbol{\theta}}^T \hat{\boldsymbol{\theta}}) - (\mathbf{x}_0^T \hat{\boldsymbol{\gamma}})(\mathbf{x}_0 \hat{\boldsymbol{\gamma}}) \\ &= 14475 - 13005 \\ &= 1470.\end{aligned}$$

From this the test statistic becomes

$$\frac{\|\hat{\boldsymbol{\theta}} - \mathbf{x}_0\hat{\boldsymbol{\gamma}}\|^2 / (2-1)}{\|\mathbf{y} - \hat{\boldsymbol{\theta}}\|^2 / (3-2)} = \frac{1470}{150} = 9.8 \sim F(1, 1)_{0.80},$$

and we accept the hypothesis at least for any $\alpha < 20\%$.

Explanation of the degrees of freedom:

$$\begin{aligned}\text{rk}(\mathbf{x}) &= \text{rk} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = 2 = k \\ \text{rk}(\mathbf{x}_0) &= \text{rk} \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} = 1 = r \\ n &= 3.\end{aligned}$$

In this case - as it is a single parameter we are testing - we could have used theorem 2.23 instead.

We had

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

From the theorem we need the degrees of freedom in the original model $f = n - \text{rk}(\mathbf{x}) = 3 - 2 = 1$, as well as the variance of the parameters

$$\hat{\sigma}^2 = \frac{1}{f} (\mathbf{Y} - \hat{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \hat{\boldsymbol{\theta}}) = \frac{150}{1} = 150$$

$$\hat{V}(\hat{\theta}_2) = \hat{\sigma}^2 \left\{ (\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x})^{-1} \right\}_{22} = 150 \left\{ \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix} \right\}_{22} = 150 \cdot \frac{5}{6} = 125$$

We can now insert

$$\hat{\theta}_{i_0} > c + t(f)_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{\theta}_{i_0})}$$

$$35 > 0 + t(1)_{0.975} \sqrt{125}$$

$$35 > 0 + 12.71 \sqrt{125}$$

$$35 > 142.1$$

Since 35 is not larger than the test value, we do not reject the hypothesis.

We will now look at the continuation of example 2.10 p. 113.

||| Example 2.31

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

$$E(Y) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \end{bmatrix} = \begin{bmatrix} \mu_1 + \theta_{11} \\ \mu_1 + \theta_{11} \\ \mu_1 + \theta_{12} \\ \mu_1 + \theta_{12} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

We now find

$$\mathbf{x}_1^T \mathbf{x}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix},$$

and

$$\mathbf{x}_1^T \mathbf{Y} = \begin{bmatrix} Y_{1..} \\ Y_{11..} \\ Y_{12..} \end{bmatrix}.$$

We see that $\mathbf{x}_1^T \mathbf{x}_1$ 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

$$\mathbf{b}^T \mathbf{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

we have

$$\mathbf{x}^T \mathbf{x} + \mathbf{b}^T \mathbf{b} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

This matrix is inverted on p. 113. 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 $(-9, +8, -8)^T$. The new residual vector is

$$\mathbf{y} - \mathbf{x}_1 \hat{\theta}_1 = (1, -1, -2, +2, -6, 0, -2)^T.$$

The norm of this vector is 50, and the number of degrees of freedom is $7-2=5$. We therefore find that

$$\begin{aligned} \|p_M(\mathbf{y}) - p_H(\mathbf{y})\|^2 &= \|\mathbf{y} - p_H(\mathbf{y})\|^2 - \|\mathbf{y} - p_M(\mathbf{y})\|^2 \\ &= 50 - 28.6 = 21.4. \end{aligned}$$

We now collect the calculations in the following analysis of variance table.

Variation	SS	f	S^2	Test
$M - H$	21.4	$3 - 2 = 1$	21.4	2.97
$O - M$	28.6	$7 - 3 = 4$	7.16	
$O - H$	50	$7 - 2 = 5$		

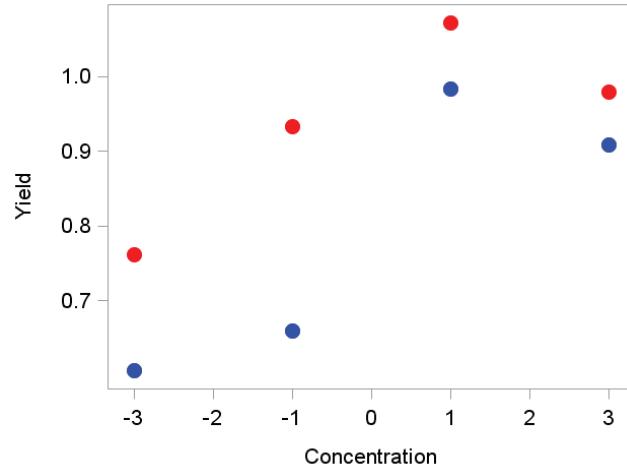
Since the observed value of the test statistic $2.97 < F(1,4)_{0.90}$ we will accept the hypothesis, and therefore assume that H_0 is true.

||| Example 2.32

We shall consider 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 Davies (1967) p. 314. The yield has been expressed by the logarithm of the weight of the mycelium after one week of growth.



The penicillin data. Red corresponds to cane sugar, blue to lactose.

We are now interested in investigating the influence of the two factors A and B on the yield. We assume that the observations are stochastically independent and normally distributed. They are called

Lactose: $Y_{11}, Y_{12}, Y_{13}, Y_{14}$

Cane sugar: $Y_{21}, Y_{22}, Y_{23}, Y_{24}$

We change the scale of the sugar concentration

$$x_{new} = \frac{x_{old} - 5\%}{1\%}$$

giving

$$2\% \rightarrow -3$$

$$\begin{aligned} 4\% &\rightarrow -1 \\ 6\% &\rightarrow 1 \\ 8\% &\rightarrow 3. \end{aligned}$$

We assume that the yield within the given limits can be expressed as polynomials of second degree, i.e. we for x_{ij} equal to the appropriate rescaled concentration, will have the expectations

$$\text{Model } M: \quad E(Y_{ij}) = \alpha_i + \beta_i x_{ij} + \gamma_i x_{ij}^2,$$

and we want to investigate whether we may use the same model for the two types, i.e.

$$\text{Hypothesis H: } E(Y_{ij}) = \alpha + \beta x_{ij} + \gamma x_{ij}^2,$$

Furthermore we assume that the observations are independent and normally distributed with the same variance σ^2 . Thus the model becomes

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \\ \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \\ \varepsilon_{24} \end{bmatrix}$$

or

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I_8)$$

We may find the estimates by direct computations, but since the design matrix has a block diagonal structure, we shall illustrate the use of this in the computations. For

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix}, \quad i = 1, 2,$$

$$\mathbf{z} = \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix},$$

and

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

we then have

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

and the model may be written

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{z} & \mathbf{0} \\ \mathbf{0} & \mathbf{z} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}.$$

Now

$$\mathbf{x}^T \mathbf{x} = \begin{bmatrix} \mathbf{z}^T \mathbf{z} & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{z}^T \mathbf{z} \end{bmatrix}$$

with

$$\mathbf{z}^T \mathbf{z} = \begin{bmatrix} 4 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix},$$

and $\mathbf{0}_3$ equal to the 3×3 null matrix. We note that

$$(\mathbf{z}^T \mathbf{z})^{-1} = \begin{bmatrix} \frac{41}{64} & 0 & -\frac{5}{64} \\ 0 & \frac{1}{20} & 0 \\ -\frac{5}{64} & 0 & \frac{1}{64} \end{bmatrix}.$$

The parameter estimates are thus

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y} = \begin{bmatrix} (\mathbf{z}^T \mathbf{z})^{-1} & \mathbf{0}_3 \\ \mathbf{0}_3 & (\mathbf{z}^T \mathbf{z})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{z}^T & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{z}^T \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} (\mathbf{z}^T \mathbf{z})^{-1} \mathbf{z}^T \mathbf{Y}_1 \\ (\mathbf{z}^T \mathbf{z})^{-1} \mathbf{z}^T \mathbf{Y}_2 \end{bmatrix},$$

and with the numerical values inserted we get

$$\begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{16}(-Y_{11} + 9Y_{12} + 9Y_{13} - Y_{14}) \\ \frac{1}{20}(-3Y_{11} - Y_{12} + Y_{13} + 3Y_{14}) \\ \frac{1}{16}(Y_{11} - Y_{12} - Y_{13} + Y_{14}) \\ \frac{1}{16}(-Y_{21} + 9Y_{22} + 9Y_{23} - Y_{24}) \\ \frac{1}{20}(-3Y_{21} - Y_{22} + Y_{23} + 3Y_{24}) \\ \frac{1}{16}(Y_{21} - Y_{22} - Y_{23} + Y_{24}) \end{bmatrix} = \begin{bmatrix} 0.830125 \\ 0.0615 \\ -0.008125 \\ 1.0190625 \\ 0.03965 \\ -0.0165625 \end{bmatrix}$$

The predicted values become

$$p_{H_1}(\mathbf{Y}) = \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{z}\hat{\theta}_1 \\ \mathbf{z}\hat{\theta}_2 \end{bmatrix} = \begin{bmatrix} 0.5725 \\ 0.7605 \\ 0.8835 \\ 0.9415 \\ 0.75105 \\ 0.96285 \\ 1.04215 \\ 0.98895 \end{bmatrix}$$

and the vector of residuals becomes

$$\mathbf{Y} - p_M(\mathbf{Y}) = \mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}} = \begin{bmatrix} 0.0335 \\ -0.1005 \\ 0.1005 \\ -0.0335 \\ 0.00995 \\ -0.02985 \\ 0.02985 \\ -0.00995 \end{bmatrix}$$

with squared length

$$SS_{res}(M) = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = 0.0335^2 + \dots + (-0.00995)^2 = 0.02442505$$

and $(8 - 6) = 2$ degrees of freedom.

The hypothesis corresponds to the model

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} = \begin{bmatrix} z \\ z \end{bmatrix} \boldsymbol{\vartheta} = \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix} \boldsymbol{\vartheta} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

Since

$$\begin{bmatrix} \mathbf{z} \\ \mathbf{z} \end{bmatrix}^T \begin{bmatrix} \mathbf{z} \\ \mathbf{z} \end{bmatrix} = [\mathbf{z}^T \quad \mathbf{z}^T] \begin{bmatrix} \mathbf{z} \\ \mathbf{z} \end{bmatrix} = 2\mathbf{z}^T \mathbf{z}$$

and

$$\begin{bmatrix} \mathbf{z} \\ \mathbf{z} \end{bmatrix}^T \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \mathbf{z}^T \mathbf{Y}_1 + \mathbf{z}^T \mathbf{Y}_2$$

we have

$$\hat{\boldsymbol{\vartheta}} = \frac{1}{2} \left\{ (\mathbf{z}^T \mathbf{z})^{-1} \mathbf{z}^T \mathbf{Y}_1 + (\mathbf{z}^T \mathbf{z})^{-1} \mathbf{z}^T \mathbf{Y}_2 \right\} = \frac{1}{2} \{ \hat{\theta}_1 + \hat{\theta}_2 \} = \begin{bmatrix} 0.92459375 \\ 0.050575 \\ -0.01234375 \end{bmatrix},$$

very naturally the average of the two $\hat{\theta}_i$ estimates. The vector of residuals becomes

$$\mathbf{Y} - p_H(\mathbf{Y}) = \mathbf{Y} - \begin{bmatrix} \mathbf{z} \\ \mathbf{z} \end{bmatrix} \hat{\boldsymbol{\vartheta}} = \begin{bmatrix} -0.055775 \\ -0.201675 \\ 0.021175 \\ -0.057225 \\ 0.099225 \\ 0.071325 \\ 0.109175 \\ 0.01375 \end{bmatrix}$$

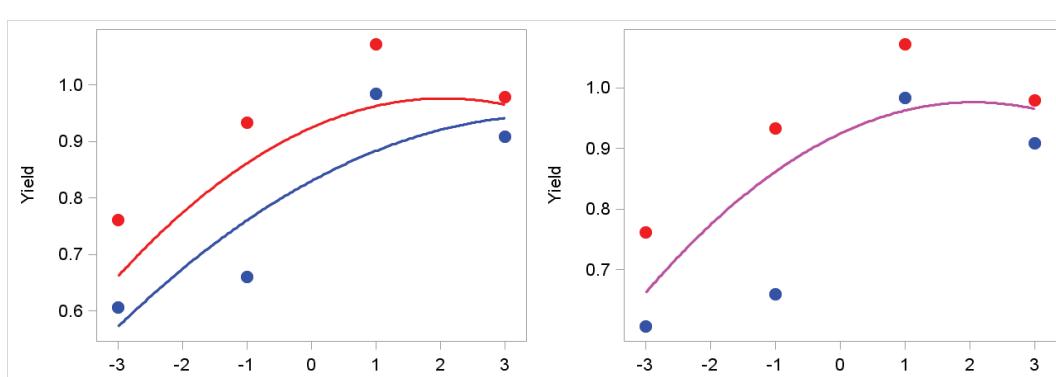
with squared length

$$SS_{res}(H) = 0.074548525$$

that has $8 - 3 = 5$ degrees of freedom. The test statistic for testing the hypothesis is therefore

$$\frac{(0.074548525 - 0.02442505)/(5-2)}{0.02442505/2} = 1.368089 \cong F(3, 2)_{0.5513}$$

i.e. the hypothesis will be accepted for all significance levels < 0.4487 .



The fitted polynomials for the model (left) and the hypothesis (right)

2.2.2 Specification of effects and sums of squares in PROC GLM

We initially consider models with intercept and their treatment in SAS, before moving to model specification and sums of squares.

||| Explanation 2.33 GLM with intercept

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

$$y = \alpha + \beta_1 x1 + \beta_2 x2 + \varepsilon.$$

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

$$\text{SS(Model)} = (\mathbf{x}\hat{\boldsymbol{\theta}} - \bar{Y}\mathbf{1})^T(\mathbf{x}\hat{\boldsymbol{\theta}} - \bar{Y}\mathbf{1}) = \mathbf{Y}^T \mathbf{H} \mathbf{Y} - n\bar{Y}^2$$

$$\text{SSRes(Model)} = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = \mathbf{Y}^T \mathbf{M} \mathbf{Y}$$

$$\text{SSTot(Corrected)} = \mathbf{Y}^T \mathbf{Y} - n\bar{Y}^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

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

$$F = \frac{\text{SS(Model)} / (\text{rk}(\mathbf{x}) - 1)}{\text{SSRes(Model)} / (n - \text{rk}(\mathbf{x}))}$$

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 .

In the sequel we quote parts of the SAS manual on the GLM procedure verbatim. For a complete exposition we refer to the manual.

"Each term in a model, called an *effect*, is a *variable* or a *combination of variables*. Effects are specified using variable names and operators. There are two kinds of variables:

1. *Class variables* (Classification, categorical, qualitative, discrete, or nominal)
2. *Continuous variables*

There are two primary operations

1. *Crossing*
2. *Nesting*

There are seven different types of effects used in the GLM procedure. in the following list, assume that A, B, C, D , and E are class variables and that $X1, X2$, and Y are continuous variables.

1. *Regressor effects* are specified by writing continuous variables by themselves: $X1 X2$
2. *Polynomial effects* are specified by joining two or more continuous variables with asterisks: $X1*X1 X1*X2$.
3. *Main effects* are specified by writing class variables by themselves: $A B C$.
4. *Crossed effects (interactions)* are specified by joining classification variables with asterisks: $A*B B*C A*B*C$.

5. *Nested effects* are specified by following a main effect or crossed effect with a classification variable or list of classification variables enclosed in parentheses. The main effect or crossed effect is nested within the effects listed in parentheses: $B(A)$ $C(B^*A)$ $D^*E(C^*B^*A)$. In this example $B(A)$ is read “ B nested within A ”.
6. *Continuous-by-class effects* are written by joining continuous variables and classification variables with asterisks: $X1^*A$.
7. *Continuous-nesting-class effects* consist of continuous variables followed by a classification variable interaction enclosed in parentheses: $X1(A) X1^*X2(A^*B)$.

We illustrate the use of these conventions in

||| Example 2.34

We again consider the penicillin data presented in example 2.32. We assume that the model M is valid and want to test the hypothesis H . We introduced the model design matrix

$$\mathbf{x} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix}$$

and the hypothesis design matrix

$$\mathbf{x}_0 = \begin{bmatrix} 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \\ 1 & -3 & 9 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 3 & 9 \end{bmatrix}$$

in order to find the necessary quantities. However, if we are using the above conventions, we just need the design matrix

lac	-3
lac	-1
lac	1
lac	3
can	-3
can	-1
can	1
can	9

Here, the first column corresponds to the first variable, let us say ‘type’. This is a classification variable assuming the values ‘lac’ and ‘can’. The second variable (~ second column), let us say ‘conc’ gives the sugar concentration and assumes the values $-3, -1, 1, 3$. If we declare ‘type’ as a classification variable:

```
class type;
```

the two model statements

```
model yield = type conc(type) conc * conc(type) / noint solution;
model yield = conc conc * conc / solution;
```

will provide all the parameter estimates and residual sums of squares for the model M and the hypothesis H described in example 2.32.

When analyzing effects in a general linear model, SAS considers four types of sums of squares, Type I, Type II, Type III, and Type IV. These different types are particularly relevant when we consider unbalanced ANOVA’s, i.e. ANOVA’s with a different number of observations for different treatment combinations. We shall only consider Type I and Type III sums of squares.

||| Definition 2.35

A *Type I Sum of Squares (SS)* is a sequential sum of squares where the effects are added into the model equation one at a time, and the SS is the resulting reduction in the residual (error) sum of squares. A *Type III Sum of Squares* is a partial sum of squares where the SS is computed by comparing the full model to the model without the effect considered.

In many ways, Type III Sums of Squares are preferable, and we shall use them as default. In balanced models the two types coincide, in unbalanced cases they

differ.

In general the Type III SS must be computed by introducing dummy variables for the different treatments (number of levels minus 1 for each treatment). Interactions are represented by the product of the respective dummy variables. In this setting the test statistics are computed in the usual way. We shall illustrate this with an example. For a deeper discussion on different types of sum of squares see the SAS manual or e.g. [Milliken and Johnson \(1993\)](#).

||| Example 2.36

We consider coded values of the strength of cement samples. In the experiment two different mixers and two different crushers were used. The data are presented in the table below.

		Crusher	
		C1	C2
Mixer	M1	1,2	-2,1,2
	M2	5,4	4

8 measurements of strength of cement samples obtained using two mixers and two crushers.

For a person not familiar with SAS, it is not obvious how to make an input routine for the data in the table above. Therefore a sample program for doing this is given below

```

data cement;
infile cards missover;
input mixer $ crusher $ strength @;
do until (strength =.);
  output;
  input strength @;
end;
datalines;
M1 C1 1 2
M1 C2 -2 1 2
M2 C1 5 4
M2 C2 4
run;
```

The table below shows how the data look (using the print procedure) when organized observation upon observation, the format needed for most SAS procedures.

Obs	strength	mixer	crusher
1	1	M1	C1
2	2	M1	C1
3	-2	M1	C2
4	1	M1	C2
5	2	M1	C2
6	5	M2	C1
7	4	M2	C1
8	4	M2	C2

Using the GLM procedure we may now compute statistics in a usual two-sided analysis of variance. The SAS statements are

```
proc glm data = cement;
class mixer crusher;
model strength = mixer crusher mixer*crusher;
run;
```

The output from this program will include the sums of squares presented in the table below. We shall now show how the Type I and Type III sums of squares may be computed by repeated application of 'PROC GLM'.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	25.20833333	8.40277778	3.48	0.1300
Error	4	9.66666667	2.41666667		
Corrected Total	7	34.87500000			

Source	DF	Type I SS	Mean Square	F Value	Pr > F
mixer	1	23.40833333	23.40833333	9.69	0.0358
crusher	1	1.60952381	1.60952381	0.67	0.4603
mixer*crusher	1	0.19047619	0.19047619	0.08	0.7928

Source	DF	Type III SS	Mean Square	F Value	Pr > F
mixer	1	19.04761905	19.04761905	7.88	0.0484
crusher	1	1.19047619	1.19047619	0.49	0.5215
mixer*crusher	1	0.19047619	0.19047619	0.08	0.7928

The sums of squares provided by PROC GLM.

We shall first consider the Type I case. The different quantities may be found as differences between residual sums of squares in the following models:

H_1 : `class` mixer crusher; `model` strength=mixer;

H_2 : `class` mixer crusher; `model` strength=mixer crusher;

H_3 : `class` mixer crusher; `model` strength= mixer crusher mixer*crusher;

We get the values

$$\begin{aligned} SS_{tot} &= 34.87500000 \quad dif = 23.40833333 \\ SS_{res}(H_1) &= 11.46666667 \quad dif = 1.60952381 \\ SS_{res}(H_2) &= 9.85714286 \quad dif = 0.19047619 \\ SS_{res}(H_3) &= 9.66666667 \end{aligned}$$

where we have used that $SS_{mod}(H_1) = SS_{tot} - SS_{res}(H_1)$, and we immediately recognize the Type I SS.

In order to find the Type III SS, we introduce dummy variables as described earlier. They are presented in the table below.

strength	idum	mdum	cdum	mcdum
1	1	1	1	1
2	1	1	1	1
-2	1	1	-1	-1
1	1	1	-1	-1
2	1	1	-1	-1
5	1	-1	1	-1
4	1	-1	1	-1
4	1	-1	-1	1

The dummy variables idum (the intercept), mdum (mixer), cdum (crusher), and mcdum (the interaction mixer*crusher). The last column is the product of the two previous columns.

Using this data matrix we may apply `PROC GLM` six times and obtain the following residuals and differences between residuals:

`model` strength = idum mdum cdum mcdum;
`model` strength = idum mdum cdum;
 SS_{III} (mixer * crusher)

$$\begin{aligned} SS_{res} &= 9.66666667 \\ \underline{SS_{res}} &= 9.85714286 \\ &= 0.19047619 \end{aligned}$$

`model` strength = idum cdum mcdum mdum;
`model` strength = idum cdum mcdum;
 SS_{III} (mixer)

$$\begin{aligned} SS_{res} &= 9.66666667 \\ \underline{SS_{res}} &= 28.71428571 \\ &= 19.04761904 \end{aligned}$$

`model` strength = idum mdum mcdum cdum;
`model` strength = idum mdum mcdum;
 SS_{III} (crusher)

$$\begin{aligned} SS_{res} &= 9.66666667 \\ \underline{SS_{res}} &= 10.85714286 \\ &= 1.19047619 \end{aligned}$$

Note that we do **not** have a 'class' statement here! The models are fitted directly to the data in the table above.

2.2.3 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

$$H_0 \supseteq H_1 \supseteq H_2 \supseteq \cdots \supseteq H_m, \quad m \leq k.$$

The approach is illustrated in figure 2.5.

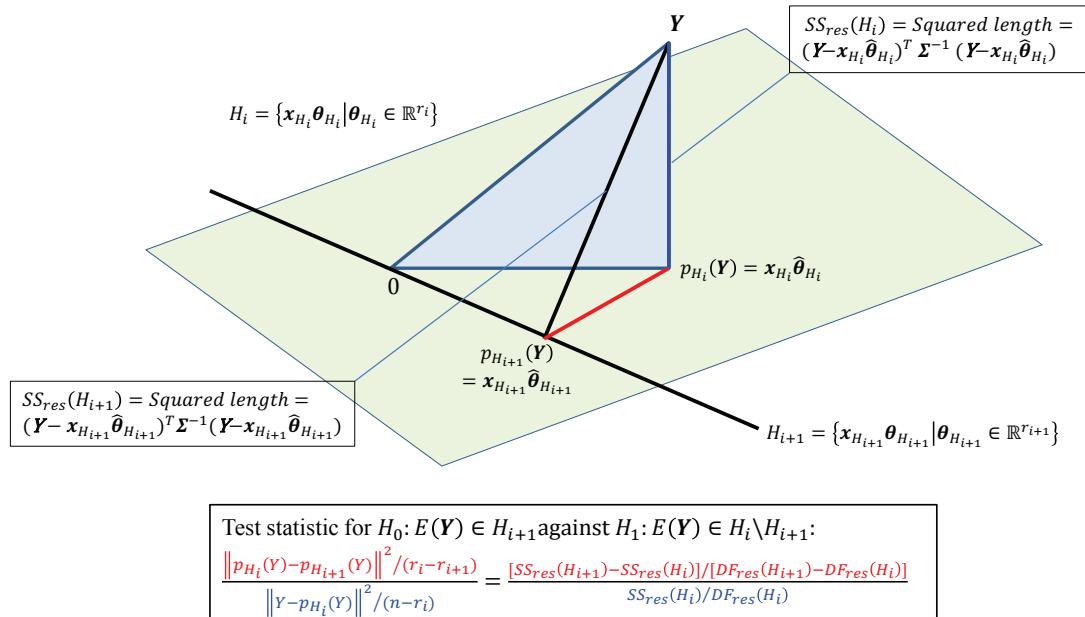


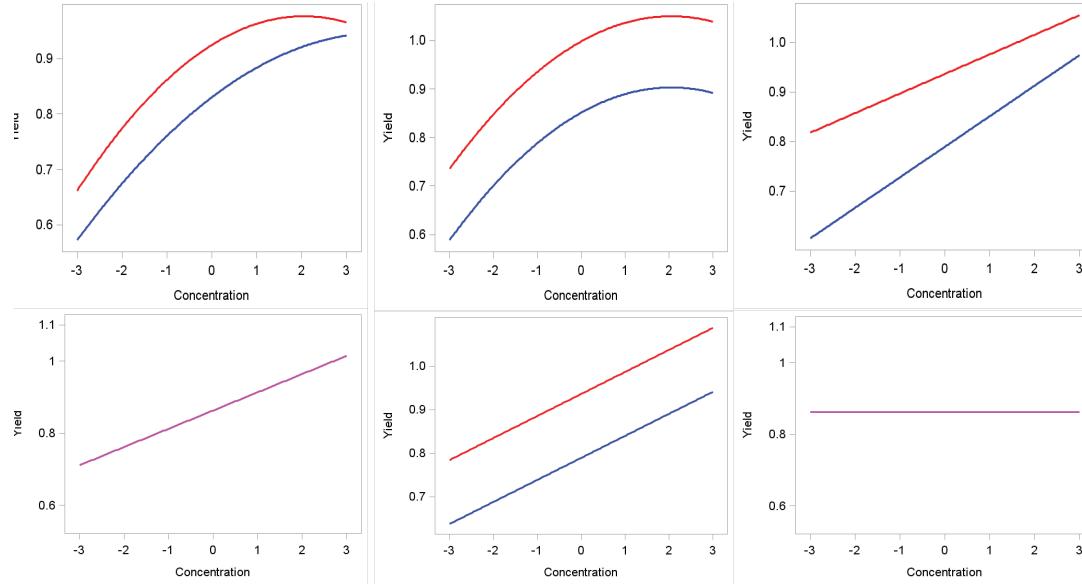
Figure 2.5

To demonstrate we consider a continuation of the analysis of the penicillin data in

||| Example 2.37

We again consider the penicillin data analyzed in examples 2.32 and 2.34. We shall now investigate possible simplifications of modelling the dependence of yield on sugar type and sugar concentration. In the previous examples we simplified the model by going from having two different polynomials to only having one. Here we shall consider a different path of simplifications. We shall rename the model M as H_1 , and have

$$H_1: E(Y_{ij}) = \alpha_i + \beta_i x_{ij} + \gamma_i x_{ij}^2, \quad i = 1, 2, \quad j = 1, 2, 3, 4,$$



A sequence of possible simplifications of the model for describing how penicillin yield depends on sugar type and sugar concentration.

We shall now successively investigate simpler models by testing whether we may assume that

1. $\gamma_1 = \gamma_2 = \gamma$, i.e. assuming a common second degree term. If that is accepted then we may test whether
2. $\gamma = 0$, corresponding to that a description by affine functions is sufficient. If that is accepted then test whether
3. $\beta_1 = \beta_2 = \beta$, i.e. whether the marginal effect by increasing the concentration is the same for the two types of sugar, and if this is accepted test whether
4. $\alpha_1 = \alpha_2 = \alpha$, i.e. the affine fittings of the yield for the two types of sugar are equal. If this is the case, test whether
5. $\beta = 0$. Accept of this means that the concentration has no influence at all.

The successive simplifications given above corresponds to a sequence of models (and parameter vectors)

$$\begin{aligned}
 H_1 : E(Y_{ij}) &= \alpha_i + \beta_i x_{ij} + \gamma_i x_{ij}^2 & \boldsymbol{\theta}^T &= [\alpha_1 \quad \beta_1 \quad \gamma_1 \quad \alpha_2 \quad \beta_2 \quad \gamma_2] \\
 H_2 : E(Y_{ij}) &= \alpha_i + \beta_i x_{ij} + \gamma x_{ij}^2 & \boldsymbol{\theta}^T &= [\alpha_1 \quad \beta_1 \quad \alpha_2 \quad \beta_2 \quad \gamma] \\
 H_3 : E(Y_{ij}) &= \alpha_i + \beta_i x_{ij} & \boldsymbol{\theta}^T &= [\alpha_1 \quad \beta_1 \quad \alpha_2 \quad \beta_2] \\
 H_4 : E(Y_{ij}) &= \alpha_i + \beta x_{ij} & \boldsymbol{\theta}^T &= [\alpha_1 \quad \alpha_2 \quad \beta] \\
 H_5 : E(Y_{ij}) &= \alpha + \beta x_{ij} & \boldsymbol{\theta}^T &= [\alpha \quad \beta] \\
 H_6 : E(Y_{ij}) &= \alpha & \boldsymbol{\theta}^T &= [\alpha]
 \end{aligned}$$

The design matrices corresponding to these models are presented in the table below

yield	H_1						H_2					H_3				H_4			H_5		H_6		
	1	2	3	4	5	6	1	2	3	4	5	1	2	3	4	1	2	3	1	2	1	2	1
0.606	1	-3	9				1	-3		9		1	-3			1	-3		1	-3		1	
0.660	1	-1	1				1	-1		1		1	-1			1	-1		1	-1		1	
0.984	1	1	1				1	1		1		1	1			1	1		1	1		1	
0.908	1	3	9				1	3		9		1	3			1	3		1	3		1	
0.761		1	-3	9				1	-3	9				1	-3		1	-3		1	-3		1
0.933		1	-1	1				1	-1	1				1	-1		1	-1		1	-1		1
1.072		1	1	1				1	1	1				1	1		1	1		1	1		1
0.979		1	3	9				1	3	9				1	3		1	3		1	3		1

The dependent variable 'yield' and the 6 design matrices corresponding to 6, 5, 4, 3, 2, and 1 parameter(s) respectively. Missing entries are zeros.

As we saw in example 2.34, we do not need to introduce all these matrices in order to estimate the necessary parameters.

yield	type	conc
0.606	lac	-3
0.660	lac	-1
0.984	lac	1
0.908	lac	3
0.761	can	-3
0.933	can	-1
1.072	can	1
0.979	can	3

The data matrix for the penicillin case.

With the variables 'yield', 'type', and 'conc', the following statements will estimate all necessary parameters in the 6 models.

```

 $H_1$  : class type; model yield = type conc(type) conc * conc(type) / noint solution;
 $H_2$  : class type; model yield = type conc(type) conc * conc / noint solution;
 $H_3$  : class type; model yield = type conc(type) / noint solution;
 $H_4$  : class type; model yield = type conc / noint solution;
 $H_5$  : class type; model yield = conc / solution;
 $H_6$  : class type; model yield = ;

```

Now, in order to test a hypothesis like

$$H_0: \boldsymbol{\theta} \in H_{i+1} \text{ against } \boldsymbol{\theta} \in H_i \setminus H_{i+1}$$

we may compute the residual sums of squares

$$SS_{res}(H_i) \text{ and } SS_{res}(H_{i+1})$$

and obtain the test statistic as

$$F = \frac{\{SS_{res}(H_{i+1}) - SS_{res}(H_i)\} / \{DF_{res}(H_{i+1}) - DF_{res}(H_i)\}}{SS_{res}(H_i) / DF_{res}(H_i)}$$

The critical values are with significance level p

$$\left\{ Y \mid F > F(DF_{res}(H_{i+1}) - DF_{res}(H_i), DF_{res}(H_i))_{1-p} \right\}$$

Model H_i	Dim(H_i)	$SS_{res}(H_i)$	$DF_{res}(H_i)$	Variation $SS_{res}(H_i)$ $- SS_{res}(H_{i-1})$	DF	Test statistic	p-value
H_6	1	0.19636487	7	0.10231322	1	$\frac{0.10231322}{0.09405165/6} = 6.5270$	0.0432
H_5	2	0.09405165	6	0.04307113	1	$\frac{0.04307113}{0.05098052/5} = 4.2243$	0.0950
H_4	3	0.05098052	5	0.00477422	1	$\frac{0.00477422}{0.04620630/4} = 0.4133$	0.5553
H_3	4	0.04620630	4	0.01950313	1	$\frac{0.01950313}{0.02670317/3} = 2.1911$	0.2354
H_2	5	0.02670317	3	0.00227812	1	$\frac{0.00227812}{0.02442505/2} = 0.1865$	0.7079
H_1	6	0.02442505	2	0.02442505	2		
Obs - H_6				0.19636487	7		

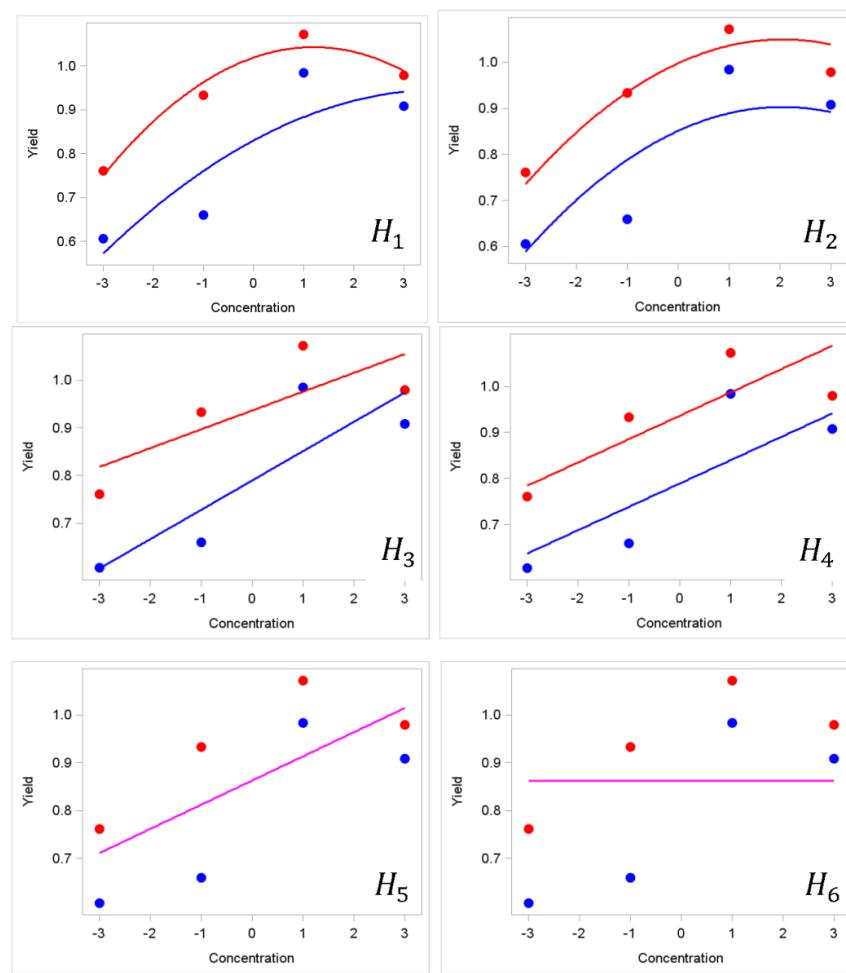
The 6 models, their residual sums of squares and test statistics

We have collected the relevant quantities for testing the 5 possible hypotheses in the table above. It follows that the p-value for testing $\theta \in H_2$ against $\theta \in H_1 \setminus H_2$ is 0.7179. Therefore the hypothesis is accepted and we assume a common second degree term. Given this, we test $\theta \in H_3$ against $\theta \in H_2 \setminus H_3$. Again we have a p-value $(0.2354) > 0.05$, and we accept. We continue with H_4 and H_5 and they are both accepted given the previous model. But assuming H_5 , we can not accept H_6 since the p-value is smaller than 0.05. The resulting model is thus

$$H_5 : E(Y_{ij}) = \alpha + \beta x_{ij}$$

i.e the penicillin yield may be described as an affine function of the sugar concentration independent of sugar type.

Now, obviously the number of observations is rather limited, so the conclusion is not strongly supported by empirical evidence. A visual inspection of the different fits presented in the figure below may indicate that model H_2 may be a more realistic fit.



The different models fitted to the observed yields.

2.3 Repeated Measurements Models - RMM

In this section we shall consider a special case - *repeated measurements models* - of an extension of the general linear model, namely the so-called mixed models. The standard form of the linear mixed model is obtained by extending the usual general linear model $\mathbf{Y} = \mathbf{x}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\theta}$ is a vector of fixed-effects parameters, with a term $\mathbf{z}\mathbf{G}$ containing the random effects \mathbf{G} . The random effects are unobservable random variables that are assumed to be uncorrelated with the error term $\boldsymbol{\varepsilon}$ giving

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

$$\mathbf{G} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$$

$$\varepsilon \sim N(\mathbf{0}, \Delta)$$

$$C(G, \varepsilon) = \mathbf{0}$$

We then obtain

$$Y \sim N(x\theta, z\Gamma z^T + \Delta)$$

This model encompasses a huge number of extremely relevant special cases, and the interested reader must be referred to the vast literature on this. The software procedure of choice in SAS is **proc mixed**.

2.3.1 Introduction

We shall introduce the models via a small example on assessing indoor climate.

||| Example 2.38

Over several years the concept of thermal comfort was investigated and brought on a scientifically tractable form (the comfort equation) through experiments in the environmental chamber at the Laboratory of Heating and Air Conditioning, the Technical University of Denmark. For an early presentation of the work, see e.g. P.O. Fanger (1972). The present data are a minor subset of the outcome of an experiment where the skin temperature at 14 localizations on the body are measured over a period of 2.5 hours on 16 males and 16 females. The test subjects are assumed to be in ‘thermal comfort’, i.e. they are regularly asked whether they feel comfortable, and the air temperature, humidity etc. are modified according to the needs expressed by them. The indoor climate parameters can be controlled accurately and changed quickly. The question of interest is whether there is a development over time in the measurements. In table 1 we show the measurements on the right foot for three male and three female test subjects.

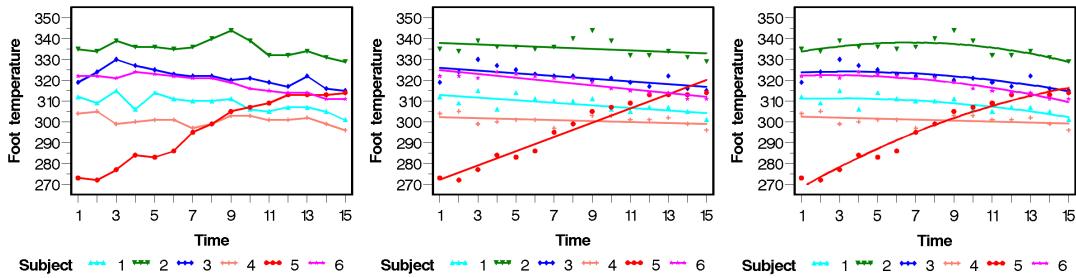
An immediate suggestion for modelling those data are to estimate regression functions for each subject. The outcome of such analyses are shown in figure 1. We have assumed that the data are realizations of independent random variables

$$\begin{aligned} Y_{1,1}, \dots, Y_{1,15} &\sim N(\alpha_1 + \beta_1 t + \gamma_1 t^2, \sigma^2) \\ &\vdots \\ Y_{6,1}, \dots, Y_{6,15} &\sim N(\alpha_6 + \beta_6 t + \gamma_6 t^2, \sigma^2) \end{aligned}$$

with an obvious modification if we assume a linear instead of a quadratic relation.

Gen- der	Sub- ject	Temperature measurement at time														
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	312	309	315	306	314	311	310	310	311	306	305	307	307	305	301
1	2	335	334	339	336	336	335	336	340	344	339	332	332	334	331	329
1	3	319	324	330	327	325	323	322	322	320	321	319	317	322	316	315
2	4	304	305	299	300	301	301	297	299	303	303	301	301	302	299	296
2	5	273	272	277	284	283	286	295	299	305	307	309	313	313	313	314
2	6	322	322	321	324	323	322	321	321	319	316	315	314	314	311	311

Skin temperature in 0.1°C on right foot at 15 equally spaced time points for 6 test subjects



Graphs depicting the measurements from the table. The first graph shows the measurements joined by line segments, the second a regression line for each subject, and the third quadratic regressions.

On the graphs we see that the points do not vary randomly around the regression curves indicating that observations from the same subject seem to be correlated, i.e. $\text{Cov}(Y_{i,j}, Y_{i,k}) \neq 0$, and thus violating the assumptions behind the regression analyses. The problem is due to the fact that we have measured the temperature repeatedly on the same subject. A more accurate model will therefore include autocorrelations between measurements on the same subjects, but of course still assume that measurements from different subjects are independent.

We shall formulate these considerations in the next section.

2.3.2 A model for repeated measurements

We consider an experiment where we want to assess the effect of different treatments by measuring the outcome Y at n different time points for a series of subjects that each have been assigned to one of the treatments. The layout is depicted in the table 2.1.

Treatment j	Subject i(j)	Time		
		t_1	\dots	t_n
1	1(1)	$Y_{1(1)1}$	\dots	$Y_{1(1)n}$
	\vdots	\vdots	\dots	\vdots
	$k_1(1)$	$Y_{k_1(1)1}$	\dots	$Y_{k_1(1)n}$
\vdots	\vdots	\vdots	\dots	\vdots
m	1(m)	$Y_{1(m)1}$	\dots	$Y_{1(m)n}$
	\vdots	\vdots	\dots	\vdots
	$k_m(m)$	$Y_{k_m(m)1}$	\dots	$Y_{k_m(m)n}$

Table 2.1 – Lay out for repeated measurements regression models.

We may consider this as a so-called factorial layout with 3 factors:

Treatments $j = 1, \dots, m$

Subjects $i(j)$, $i = 1, \dots, k_j$, $j = 1, \dots, m$ nested & within treatments

Time t_1, \dots, t_n

For subject no $i(j)$, $i = 1, \dots, k_j$, $j = 1, \dots, m$ we define

$$\mathbf{Y}_{i(j)} = \begin{bmatrix} Y_{i(j)1} \\ \vdots \\ Y_{i(j)n} \end{bmatrix}$$

Since we are observing repeated measurements on the same subjects we can not assume that all observations are independent. Allowing for autocorrelations between measurements on the same subject, the basic model for subject no i , $i = 1, \dots, k$ is in this more general setting

$$\mathbf{Y}_{i(j)} \sim N_n(\boldsymbol{\mu}_{i(j)}, \boldsymbol{\Sigma}_{i(j)})$$

where as well the mean value parameter $\boldsymbol{\mu}_{i(j)}$ as the dispersion matrix $\boldsymbol{\Sigma}_{i(j)}$ are unknown. The structure of $\boldsymbol{\Sigma}_{i(j)}$ is determined by the autocorrelations. However, we shall in general apply the reasonable assumption that measurements on different subjects are modelled as independent observations.

|||| **Remark 2.39**

With concepts introduced in chapter 4, it follows that we have a Multivariate General Linear Model with 2 factors.

Treatments $j = 1, \dots, m$

Subjects $i(j)$, $i = 1, \dots, k_j$, $j = 1, \dots, m$ within treatments

But this way of considering the observations does not immediately allow us to model the time dependency of the measurements.

2.3.3 Some covariance structures

For a wide class of dispersion matrices we may actually use the F-tests obtained from an ordinary least squares analysis of the three way factorial. The general form of such covariance structures are known as the *Huyn-Feldt conditions*.

|||| **Definition 2.40**

A dispersion matrix satisfies the *Huyn-Feldt condition* iff it has the form

$$\Sigma = \lambda \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \gamma_1 & \cdots & \gamma_n \end{bmatrix}$$

See e.g. [Milliken and Johnson \(1993\)](#).

A special case of this is matrices of the form

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

This form is called *compound symmetry*.

Another common form – and in repeated measures designs perhaps more applicable – structure is the *autoregressive model* (AR(1)) corresponding to time points $1, \dots, n$:

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^n \\ \rho & 1 & \rho & \cdots & \rho^{n-1} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^n & \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}$$

This form does **not** generally satisfy the Huyn-Feldt condition.

2.3.4 Subject dependent time points

Often there will be missing observations by accident or simply by design (different time points for some of the subjects) so that the model for subject i , ($i = 1, \dots, k$, where $k = k_1(1) + \dots + k_m(m)$ is the total number of subjects) becomes

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

To exemplify matters we may e.g. assume that the mean for each subject is a second degree polynomial in time, i.e. the basic model for subject no. i is

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \\ \gamma_i \end{bmatrix} + \boldsymbol{\varepsilon}_i.$$

We assume that \mathbf{Y}_i is normally distributed with mean $\mathbf{x}_i \boldsymbol{\theta}_i$ and variance-covariance matrix \mathbf{R}_i , and that $\mathbf{Y}_1, \dots, \mathbf{Y}_k$ are independent. This gives the combined model

$$\begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_k \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{x}_k \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta}_1 \\ \vdots \\ \boldsymbol{\theta}_k \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_k \end{bmatrix}$$

or

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

with

$$D \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_k \end{bmatrix} = \begin{bmatrix} \mathbf{R}_1 & \cdots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \cdots & \mathbf{R}_k \end{bmatrix} = \mathbf{R}$$

The random variable \mathbf{Y} is thus multivariate normally distributed

$$\mathbf{Y} \sim N_K(\mathbf{x}\boldsymbol{\theta}, \mathbf{R})$$

where K is the total number of observations.

The balanced covariance models from section 2.3.3 are no longer directly applicable, but must be modified. In continuation of the above exemplification, we e.g. assume that the dispersion matrix of $\boldsymbol{\varepsilon}_i$ is the generalization of the first order autoregressive model to the case with unevenly distributed time points, i.e.

$$D(\boldsymbol{\varepsilon}_i) = \mathbf{R}_i = \sigma^2 \begin{bmatrix} 1 & \cdots & \rho^{|t_{i1}-t_{in_i}|} \\ \vdots & & \vdots \\ \rho^{|t_{i1}-t_{in_i}|} & \cdots & 1 \end{bmatrix}$$

In [PROC MIXED](#) this structure (type) is called sp(pow)(time).

In the resulting model $\mathbf{Y} \sim N_K(\mathbf{x}\boldsymbol{\theta}, \mathbf{R})$, we shall estimate the parameters by the maximum likelihood method (ML) or the restricted maximum likelihood method (REML).

If we use ML we may obtain tests by means of the usual chi-square approximation to -2 times the log of the likelihood ratio test statistic. The degrees of freedom in this approximation are equal to the decrease in the number of 'free' parameters from the more elaborate to the simpler model.

If we use the SAS procedure [PROC MIXED](#), the data must be written in so-called univariate mode, i.e. one record for each observation giving the values of subject no., treatment, time, observation etc. This is made more obvious in the following example 2.41.

||| Example 2.41

We now continue example 2.38. The data in table A are presented in the so-called multivariate mode, and we must bring them on univariate form. This is presented in table B

Obs	Subject	Gender	time	temp
1	1	1	1	312
2	1	1	2	309
3	1	1	3	315
:	:	:	:	:
13	1	1	13	307
14	1	1	14	305
15	1	1	15	301
16	2	1	1	335
17	2	1	2	334
18	2	1	3	339
:	:	:	:	:
88	6	2	13	314
89	6	2	14	311
90	6	2	15	311

Table A: The same data as in example 2.38, namely skin temperature in 0.1 °C on right foot at 15 equally spaced time points for 6 test subjects, but presented in univariate mode.

A reasonable model for the autocorrelation structure is the (first order) autoregression AR(1). If we want to estimate gender dependent quadratic regression functions (model M) this may be done by the following SAS-code:

```
proc mixed data=comfortu method=ml covtest;
  class gender subject;
  model temp = gender time(gender) time*time(gender) / noint s;
  repeated / type = AR(1)
    sub=subject(gender)
    rcorr=1 2 3 4 5 6 ;
run;
```

The model statement for the gender independent regression H_1 and for no time dependence H_2 are

```
model temp = time time*time / s;
model temp = / s;
```

If we want to compare the repeated measurements analysis with ordinary least squares we may basically use **PROC GLM** with the same class and model statements. The estimates of the fixed effect parameters using the two model types are presented in table B. The estimates are similar, but certainly different! We have shown the data and the common quadratic regression in the figure A.

Fixed effect parameters	Independent residuals		Repeated measurements	
	Estimate	Std. Err.	Estimate	Std. Err.
$\alpha (1)$	322.14	6.4272	320.85	7.7120
$\alpha (2)$	295.99	6.4272	298.23	7.7120
$\beta (1)$	0.8726	1.8485	1.2618	1.0888
$\beta (2)$	1.8315	1.8485	1.4662	1.0888
$\gamma (1)$	-0.0880	0.1123	-0.1102	0.0619
$\gamma (2)$	-0.0668	0.1123	-0.0587	0.0619

Table B: Comparison between ordinary least squares estimates and repeated measurement estimates of the fixed effects in model M .

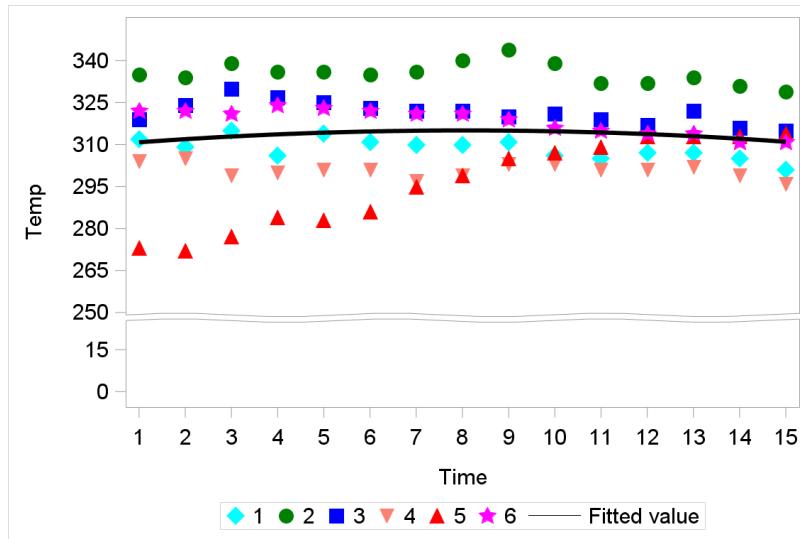


Figure A: The data and the quadratic regression corresponding to hypothesis H_1

In table C we have summarized the statistics involved in testing for a simple structure. It follows that if we assume model M , the hypotheses corresponding to H_1 and H_2 are not rejected at reasonable levels of significance.

Model	Fixed effects mean values	$\hat{\sigma}^2$	Nb. fix. eff. par.	-2 "log likelihood"	Diff. d	DF	$P \{ \chi^2 > d \}$
M	$\alpha (j) + \beta (j) t + \gamma(j)t^2$	164.59	6	482.0246			
H_1	$\alpha + \beta t + \gamma t^2$	238.85	3	486.9687	4.9441	3	0.1759
H_2	α	240.55	1	490.4989	3.5302	2	0.1712

Table C: Maximum likelihood fits with autocorrelated residuals (repeated measurements analysis) and statistics for successively testing a simpler model structure.

If we assume H_2 , the parameter estimates become

$$\begin{array}{ll} \hat{\alpha} = 311.31 & Stderr(\hat{\alpha}) = 5.8653 \\ \hat{\sigma}^2 = 240.55 & Stderr(\hat{\sigma}^2) = 121.15 \\ \hat{\rho} = 0.9767 & Stderr(\hat{\rho}) = 0.01218 \end{array}$$

and the estimated correlation matrix for any subject becomes

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85	0.83	0.81	0.79	0.77	0.75	0.74	0.72
2	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85	0.83	0.81	0.79	0.77	0.75	0.74
3	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85	0.83	0.81	0.79	0.77	0.75
4	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85	0.83	0.81	0.79	0.77
5	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85	0.83	0.81	0.79
6	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85	0.83	0.81
7	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85	0.83
8	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87	0.85
9	0.83	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89	0.87
10	0.81	0.83	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91	0.89
11	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93	0.91
12	0.77	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95	0.93
13	0.75	0.77	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98	0.95
14	0.74	0.75	0.77	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1	0.98
15	0.72	0.74	0.75	0.77	0.79	0.81	0.83	0.85	0.87	0.89	0.91	0.93	0.95	0.98	1

Table D: the estimated correlation matrix for any of the six subjects.

As mentioned above, the repeated measurements on “subjects” will not necessarily involve measurements in time. A common situation is that the measurements are taken at different locations (~physical coordinates) on the same item, and that those measurements are spatially correlated. The general term in the **spatial power** covariance model `sp(pow)(list)` is of the form

$$\sigma^2 \rho^{dist(location\ i, location\ j)}$$

where `list` contains the variables giving the coordinates of the measurement locations. If we do not know the exact locations we may e.g. use an unstructured model or a compound symmetry model.

||| Case 2.42

Color is an important factor in assessing food quality, hence measuring color of food products is important in the food industry. The data analyzed in this example come from a larger study where a traditional colorimeter (Minolta CR-300) is compared with a multispectral imaging system (VideometerLab (MSI)). The methods were compared using samples from 12 different meat products as presented in Table 7. The experimental set up was organized as shown in Table 8 with 5 different samples from each product and 4 plus 4 measurements taken at randomly chosen locations on each sample giving a total of 480 measurements. Each measurement gave the simultaneous values of the CIELAB color coordinates L^* (lightness), and the two chromatic components a^* (the green–red color component) and b^* (blue–yellow color component). We shall only consider the lightness, called L in the sequel. For further details and an alternative analysis, see [Trinderup et al. \(2015\)](#).

1: Pork loin	5: Beef loin	9: Sausage B
2: Round of pork	6: Round of beef	10: Cooked ham
3: Veal loin	7: Turkey breast	11: Cooked turkey
4: Round of veal	8: Sausage A	12: Fried meat balls

Table A: The meat products used in the analysis.

Product (Meat type)		Method	
		Colorimeter	Vision System
1: Pork Loin	Sample 1	Position 1 … Position 4	Position 5 … Position 8
	⋮ Sample 5	Position 1 … Position 4	Position 5 … Position 8
⋮			
12: Fried meat balls	Sample 1	Position 1 … Position 4	Position 5 … Position 8
	⋮ Sample 5	Position 1 … Position 4	Position 5 … Position 8

Table B: The experimental set up for comparing the two methods for measuring color.



Figure A: Two samples of round of beef. Each row represents one sample, cut into two with the cut surface upwards. The pattern of fat and connective tissue are easily recognized in each half. One half is used for measurements with the colorimeter, the other for measurements with a vision system.

Obs	Product	Sample	Method	Position	L	a	b
1	1	1	1	1	55.4200	5.9600	6.4300
2	1	1	1	2	54.6800	5.5300	5.8400
3	1	1	1	3	52.5600	5.6000	5.9100
4	1	1	1	4	54.7200	6.7900	7.3000
5	1	1	2	5	60.3096	13.9680	14.6268
6	1	1	2	6	58.0329	14.4670	14.4590
7	1	1	2	7	61.1319	14.1886	15.7660
8	1	1	2	8	58.4357	17.3479	16.8884
9	1	2	1	1	58.6300	7.3200	8.0200
10	1	2	1	2	57.2300	5.7700	6.1800
:							
471	12	4	2	7	73.1220	2.7108	13.9351
472	12	4	2	8	72.8798	2.2685	14.2577
473	12	5	1	1	70.5600	4.6600	11.2600
474	12	5	1	2	70.3300	4.4700	11.8100
475	12	5	1	3	70.2800	4.2300	11.7600
476	12	5	1	4	70.3200	4.6800	11.4400
477	12	5	2	5	73.5656	2.3999	14.1760
478	12	5	2	6	73.4646	2.5265	13.7267
479	12	5	2	7	73.1231	2.9103	13.6446
480	12	5	2	8	72.5810	2.9879	14.1671

Table C: The first and the last 10 observations of the color measurement dataset.

When comparing two methods, it may be fruitful to plot corresponding values from each measurement against each other. For each sample we have computed the average for each method over locations, and plotted those averages against each other. The result is shown in Figure B. If the methods were equivalent the points should fall on the identity line. This is however, clearly not the case. For the darker meat products the colorimeter gives larger values of the lightness, and for the lighter products the vision system provides the larger values. We shall investigate this further in the sequel.

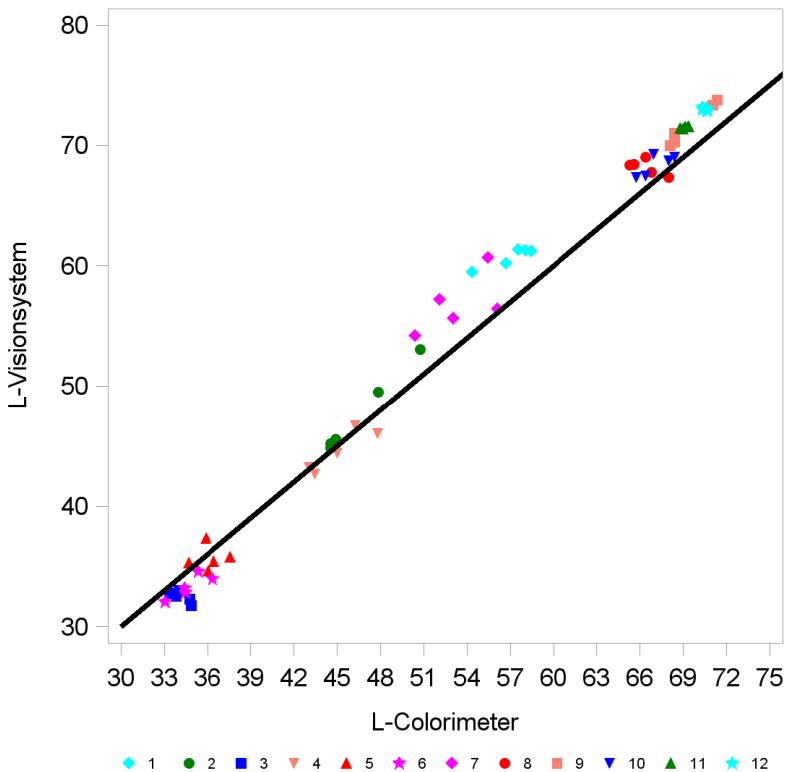


Figure B: The sample averages for the two methods plotted against each other. The 12 products are indicated in the legend. The black line is the identity line $y = x$.

We have four factors, product, sample, method, and position, influencing the outcome of the measurements. In setting up a proper model for the general observation Y_{ijkh} , we therefore introduce the effects

- | | | |
|--------------------------|----------------------|--|
| effect m_i | $i = 1, 2,$ | method, |
| effect p_j | $j = 1, \dots, 12,$ | product, |
| effect $s(p)_{k(j)}$ | $k = 1, 2, 3, 4, 5,$ | sample within product, |
| effect $q(mps)_{h(ijk)}$ | $h = 1, 2, 3, 4,$ | position within sample, product and method |

Additional effects are the interaction terms between the crossed factors, i.e. mp_{ij} and $ms(p)_{k(j)}$. We only have 4 different locations for each combination of sample,

product and method, so therefore we let the index h run from 1 to 4 and not from 1 to 8. This results in the model

$$Y_{ijkh} = \mu + m_i + p_j + mp_{ij} + s(p)_{k(j)} + ms(p)_{k(j)} + q(mps)_{h(ijk)} + \varepsilon_{ijkh}$$

where ε_{ijkh} is the error term consisting of independent $N(0, \sigma^2)$ -distributed random variables. Since the samples and the locations are due to random variation, it is more natural to model effects including those factors by independent (normally distributed) random variables giving a so-called **variance component model**. In order to distinguish this type of model from the above, we shall use capital letters for random variables

$$Y_{ijkh} = \mu + m_i + p_j + mp_{ij} + S(P)_{k(j)} + MS(P)_{k(j)} + Q(MPS)_{h(ijk)} + \varepsilon_{ijkh}$$

$$V(Y_{ijkh}) = \sigma_{S(P)}^2 + \sigma_{MS(P)}^2 + \sigma_{Q(MPS)}^2 + \sigma^2$$

and e.g.

$$\text{Cov}(Y_{ijk1}, Y_{ijk2}) = \sigma_{S(P)}^2 + \sigma_{Q(MPS)}^2$$

giving the correlation between measurements on the same sample

$$\rho(Y_{ijk1}, Y_{ijk2}) = \frac{\sigma_{S(P)}^2 + \sigma_{Q(MPS)}^2}{\sigma_{S(P)}^2 + \sigma_{MS(P)}^2 + \sigma_{Q(MPS)}^2 + \sigma^2}$$

Such a model (and more elaborate versions of that) may be analyzed by e.g. **PROC GLM** or **PROC MIXED** by adding a statement like “random sample(product);” after the model specification. This line of analysis is used in [Trinderup et al. \(2015\)](#).

We shall not go into further details with this but instead focus upon the fact that measurements taken at different positions on the same sample represent repeated measurements. We re-write the model as

$$Y_{ijkh} = \mu + m_i + p_j + mp_{ij} + \gamma_{ijkh}$$

where the error term γ_{ijkh} no longer consists of independent random variables, but of correlated variables due to the repeated character of the measurements. We shall consider three different models for each sample:

1. A compound symmetry model common for all 8 positions.
2. A compound symmetry model estimated for each method.

3. A variance component model estimated for each method

The different statistics may be obtained by the SAS statements below.

```
proc mixed data = camlab;
class product sample method position;
model L = method product method*product;
repeated / type = cs
    subject = sample(product)
    group = product
    r = 1 60
    rcorr = 1 60;
run;
```

Other settings used:

type = vc

group = product*method

2.96	1.25	1.25	1.25	1.25	1.25	1.25	1.25
1.25	2.96	1.25	1.25	1.25	1.25	1.25	1.25
1.25	1.25	2.96	1.25	1.25	1.25	1.25	1.25
1.25	1.25	1.25	2.96	1.25	1.25	1.25	1.25
1.25	1.25	1.25	1.25	2.96	1.25	1.25	1.25
1.25	1.25	1.25	1.25	1.25	2.96	1.25	1.25
1.25	1.25	1.25	1.25	1.25	1.25	2.96	1.25
1.25	1.25	1.25	1.25	1.25	1.25	1.25	2.96

1	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
0.42	1	0.42	0.42	0.42	0.42	0.42	0.42	0.42
0.42	0.42	1	0.42	0.42	0.42	0.42	0.42	0.42
0.42	0.42	0.42	1	0.42	0.42	0.42	0.42	0.42
0.42	0.42	0.42	0.42	1	0.42	0.42	0.42	0.42
0.42	0.42	0.42	0.42	0.42	1	0.42	0.42	0.42
0.42	0.42	0.42	0.42	0.42	0.42	1	0.42	0.42
0.42	0.42	0.42	0.42	0.42	0.42	0.42	1	0.42
0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	1

4.53	2.04	2.04	2.04				
2.04	4.53	2.04	2.04				
2.04	2.04	4.53	2.04				
2.04	2.04	2.04	4.53				
				1.39	0.44	0.44	0.44
				0.44	1.39	0.44	0.44
				0.44	0.44	1.39	0.44
				0.44	0.44	0.44	1.39

1	0.45	0.45	0.45					
0.45	1	0.45	0.45					
0.45	0.45	1	0.45					
0.45	0.45	0.45	1					
				1	0.32	0.32	0.32	
				0.32	1	0.32	0.32	
				0.32	0.32	1	0.32	
				0.32	0.32	0.32	1	

4.21							
	4.21						
		4.21					
			1.32				
				1.32			
					1.32		
						1.32	

1								
	1							
		1						
			1					
				1				
					1			
						1		
							1	
								1

Table D: The estimated R_{11} -matrix and corresponding correlation matrix for the 8 measurements on sample 1 from product 1 (pork loin) using the compound symmetry model (the two first, I & II) and for the variance component model (the third). The compound symmetry model has been estimated for all 8 measurements simultaneously (model I) and for each method (model II). The variance component model has been estimated for each method.

From the different covariance structures it follows that the variance of colorimeter data are larger than the variance for the vision system. This is in good accordance with the fact that the colorimeter give a spot measurement whereas the vision system integrates over an area. Despite the noticeable differences in the estimated variances and correlations, the three models do not give remarkably different results when it comes to testing for the fixed effects. The resulting Type III test statistics are given in Table 11. It follows that there – of course – are substantial differences between the lightness of the different meat products, but also that there are significant differences between the methods. The significant interaction shows that the differences are product dependent.

Effect	Num DF	Den DF	Model I		Model II		Model III	
			F Value	Pr > F	F Value	Pr > F	F Value	Pr > F
Method	1	48	58.95	<.0001	17.65	0.0001	38.88	<.0001
Product	11	48	16471.60	<.0001	6055.43	<.0001	5165.90	<.0001
Product*Method	11	48	15.86	<.0001	16.31	<.0001	13.13	<.0001

Table E: The Type III test values for the fixed effects parameters in the three models I-III.

For model II different the least squares means of effects and differences of effects are shown in Table 12.. The are found by adding statements like “`lsmeans method*product/diff;`” We see that for the darker products (veal and beef) the colorimeter estimates a larger lightness, and for the brighter products a smaller brightness. Some of those differences are highly significant, others are not.

	LS estimate of m_1, m_2 and p_1, \dots, p_{12}		LS estimate of $m_1 - m_2$ and $pm_{i1} - pm_{i2}$			
	Mean	S.Err.Mean	Diff.	S.Err.Dif.	t Value	Pr > t
Colorimeter	54.15	0.20	-1.18	0.28	-4.20	0.0001
	55.32	0.20				
Pork loin	58.85	0.41	-3.67	0.82	-4.49	<.0001
Round of pork	47.08	1.01	-1.11	2.01	-0.55	0.5825
Veal loin	33.31	0.18	1.66	0.36	4.65	<.0001
Round of veal	44.86	0.59	0.52	1.18	0.44	0.6599
Beef loin	35.91	0.32	0.47	0.65	0.72	0.4746
Round of beef	34.02	0.35	1.40	0.71	1.98	0.0537
Turkey breast	55.12	0.76	-3.40	1.52	-2.24	0.0295
Sausage A	67.29	0.28	-1.76	0.56	-3.13	0.0030
Sausage B	70.57	0.53	-2.22	1.06	-2.09	0.0420
Cooked ham	67.73	0.32	-1.24	0.64	-1.95	0.0572
Cooked turkey	70.31	0.06	-2.32	0.12	-19.14	<.0001
Fried meat balls	71.77	0.05	-2.47	0.10	-24.83	<.0001

Table F: Least squares estimates of some fixed effects.

||| Case 2.43

The data in this example come from a larger study on investigating changes in total gross tumor volume and density for 10 head and neck cancer patients during image-guided radiation therapy, see [Bjerre \(2013\)](#). The tumor sizes were estimated from a PET-CT planning scan and 2-4 replanning CT scans during the course of the treatment. The gross tumor volume was delineated on the planning PET-CT and subsequently propagated from the planning CT to each re-planning CT using manual contouring by an experienced radiation oncologist. The initial tumor volumes were between 83 and 311 cm³.

Obviously it is of clinical interest to monitor the development over time of the tumor size, and identifying possible patterns in this development may likewise be very relevant. The distribution of choice for size measurements resulting from biological growth or shrinking is the lognormal distribution. Therefore we shall consider the logarithms of the volumes in order to ensure that the distributions of the residuals from linear models are closer to Gaussians. The data are presented in table A and graphically in figure A.

In the sequel we shall only touch upon the data analytic aspects of the analyses. With respect to the clinical implications, we shall refer to the original paper found in chapter 6 in [Bjerre \(2013\)](#).

Obs	Patient	time	logvol	Obs	patient	time	logvol
1	1	0	5.3991	21	5	32	5.0168
2	1	20	4.6850	22	5	39	4.7257
3	1	22	4.9369	23	6	0	5.2481
4	1	27	4.9022	24	6	29	4.7592
5	1	34	4.8646	25	6	41	4.8260
6	2	0	5.7388	26	7	0	4.8530
7	2	19	5.4193	27	7	19	4.7308
8	2	25	5.4953	28	7	25	4.2900
9	2	31	5.4389	29	8	0	4.7609
10	3	0	4.4132	30	8	14	4.6895
11	3	20	4.2901	31	8	26	4.2037
12	3	26	3.7871	32	9	0	5.3941
13	3	33	3.4303	33	9	19	4.9042
14	4	0	4.9114	34	9	25	4.8085
15	4	20	4.7814	35	9	32	4.7490
16	4	25	4.7475	36	10	0	5.6314
17	4	26	4.6542	37	10	18	5.2399
18	4	32	4.5005	38	10	21	5.1176
19	5	0	4.7768	39	10	32	5.1192
20	5	25	4.8298				

Table A: Log(tumor size) for 10 patients at different time points.

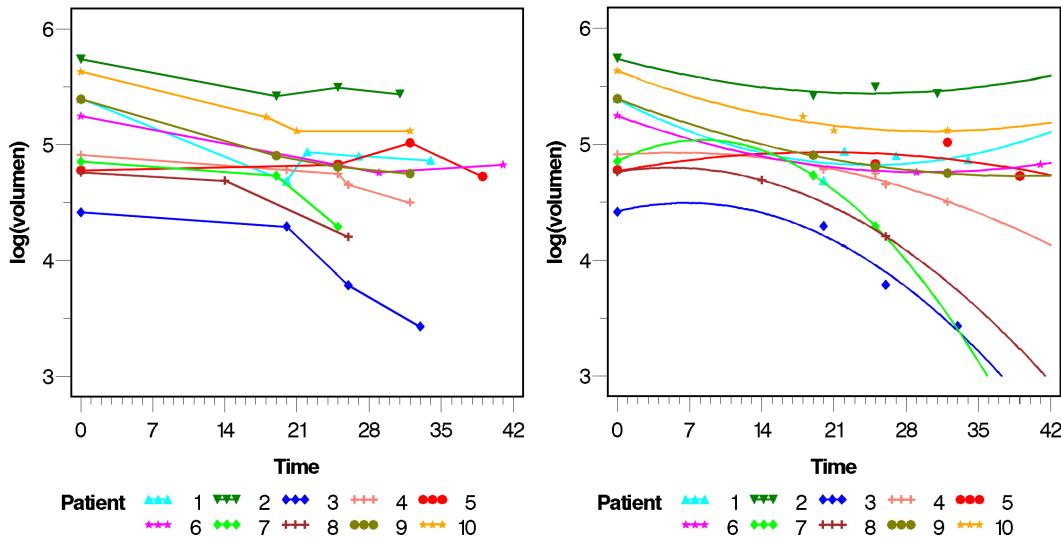


Figure A: The data points connected with piecewise linear functions, and fitted individually by quadratic regression curves.

As preliminary model for the mean values we consider individual quadratic mean value functions for each patient

$$M : E(Y_{ij}) = \alpha_i + \beta_i t_{ij} + \gamma_i t_{ij}^2, \quad j = 1, \dots, n_i, \quad i = 1, \dots, 10$$

In order to emphasize that the choice of model for the covariance structure really matters, we shall estimate the parameters as well by using ordinary least squares in the general linear model as by using the maximum likelihood method in the repeated measurements model. The SAS code for the latter approach is

```
proc mixed data=headneck method=ml covtest ;
  class patient ;
  model logvol = patient time(patient) time*time(patient)/noint s;
  repeated / type= sp(pow)(time)
            subject=patient
            rcorr=1;
  run;
```

In figure A we have also shown the individually fitted quadratic functions for the 10 patients. However, for 3 of the patients we only have 3 data points and therefore obtain a perfect fit, and in total we have 30 parameters (plus uncertainty parameters)

and only 39 observations. But it is doable to estimate all those parameters, and therefore this model (M) is considered. In the sequel we focus upon the models H_1 - H_4 obtained by increasing simplification of the previous models (See tables B and C). We may perform a formal successive testing of hypothesis and alternatives

$$H_0 : E(\mathbf{Y}) \in H_v \text{ against } H_1 : E(\mathbf{Y}) \in H_{v-1} \setminus H_v$$

This is done as well b.m.o. the usual F-test statistic as in Example 2.37 (corresponding to $\rho = 0$ in the repeated measurements model), as b.m.o. the asymptotic results for the likelihood ratio test statistic. The latter basically says that $-2 \log$ (likelihood ratio test statistic Q) under the null hypothesis will be asymptotically chi squared distributed with degrees of freedom equal to the decrease in number of 'free' parameters from the more elaborate to the simpler model, cf. earlier remarks. To make this more specific, we assume that model H_2 is true, i.e.

$$H_2 : E(Y_{ij}) = \alpha_i + \beta_i t_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, 10$$

and we want to investigate whether we may simplify this to

$$H_3 : E(Y_{ij}) = \alpha_i + \beta t_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, 10$$

Clearly we have 20 mean value parameters respectively 11 in the two models. From the iteration history in [PROC MIXED](#) we have at the respective maxima

$$H_2 : -2 \log \text{likelihood function} = -2l(\hat{\theta}_{H_2}) = -60.7821$$

$$H_3 : -2 \log \text{likelihood function} = -2l(\hat{\theta}_{H_3}) = -37.0526$$

The difference between those two values is 23.7295 which under the hypotheses that H_3 is true is an outcome of a $\chi^2(20 - 11)$ -distribution, i.e. $\chi^2(9)$ -distribution. Since

$$P\{\chi^2(9) > 23.7295\} = 0.0047$$

we reject the hypothesis.

Model	Fixed effects mean values	No. fix. eff. par.	Residual SS	DF	F-statistic f	$P\{F > f\}$
M	$\alpha_i + \beta_i t_{ij} + \gamma_i t_{ij}^2$	30	0.108494	9		
H_1	$\alpha_i + \beta_i t_{ij} + \gamma t_{ij}^2$	21	0.480786	18	3.431452	0.0403
H_2	$\alpha_i + \beta_i t_{ij}$	20	0.480851	19	0.002434	0.9615
H_3	$\alpha_i + \beta t_{ij}$	11	0.886207	28	1.779661	0.1388
H_4	α_i	10	2.047867	29	36.703027	0.0000

Table B: Least squares fits with independent residuals and statistics for successively testing a simpler model structure.

Model	Fixed effects mean values	No. fix. eff. par.	-2 "log likelihood"	Succ. diff. sd	DF	$P \{ \chi^2 > sd \}$
M	$\alpha_i + \beta_i t_{ij} + \gamma_i t_{ij}^2$	30	-118.8345			
H_1	$\alpha_i + \beta_i t_{ij} + \gamma t_{ij}^2$	21	-60.7923	58.0422	9	0.0000
H_2	$\alpha_i + \beta_i t_{ij}$	20	-60.7821	0.0102	1	0.9196
H_3	$\alpha_i + \beta t_{ij}$	11	-37.0526	23.7295	9	0.0047
H_4	α_i	10	-4.5017	32.3509	1	0.0000

Table C: Maximum likelihood fits with autocorrelated residuals (repeated measurements analysis) and statistics for successively testing a simpler model structure.

As mentioned, we discard M as a feasible model, and in our interpretation of the results, we take our starting point in model H_1 . It follows from the tables that we accept H_2 with both methods but only the classical GLM approach accepts H_3 . Since we clearly are in a repeated measurements situation, we put the stronger emphasis on the result from that analysis, and thus we conclude that model H_2 , i.e. 10 individual regression lines, is the simplest model adequately describing the development of the logarithms of the tumor volumes.

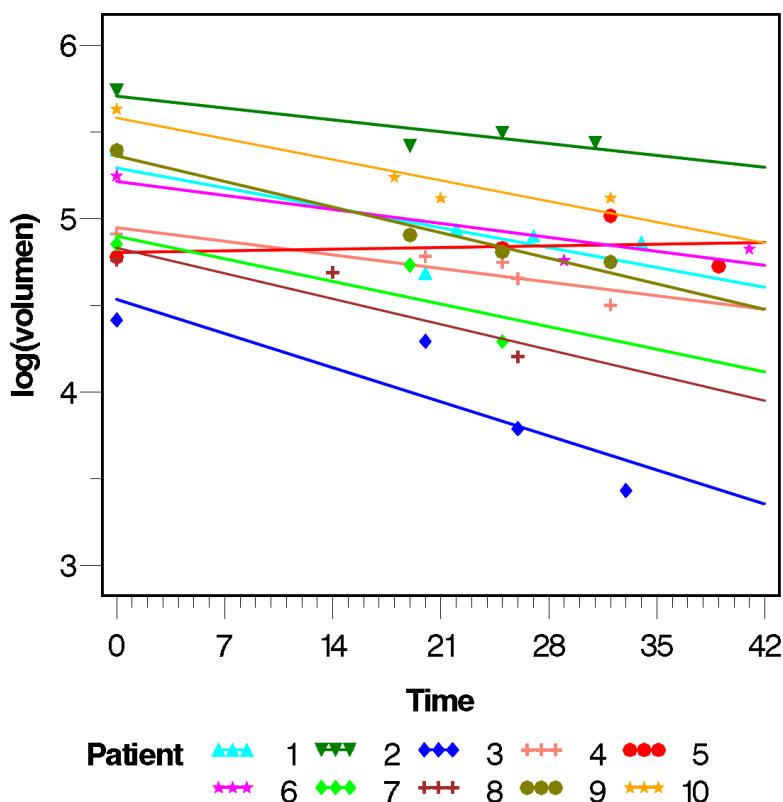


Figure B: The data points fitted by linear regression curves for each patient.

Furthermore we have that

$$\hat{\sigma}^2 = 0.01234 \quad Stderr(\hat{\sigma}^2) = 0.002799$$

$$\hat{\rho} = -0.2383 \quad Stderr(\hat{\rho}) = 0.8442$$

and the estimated correlation matrix for e.g. patient no. 1 becomes

Time	0	20	22	27	34
0	1	0	0	0	0
20	0	1	0.0568	-0.0000	0
22	0	0.0568	1	-0.0008	0
27	0	-0.0000	-0.0008	1	-0
34	0	0	0	-0	1

Table D: The estimated correlation matrix for patient no. 1

The fixed effect parameters obtained by either method are shown in table 4. It follows that the main difference between the two situations is that the estimated errors on the parameters are smaller in the repeated measurements model, also explaining why we have stronger significance in that case and thus end up with different models for the time dependence, namely individual regression lines versus regressions with a common slope and only different intercepts.

Patient	Independent residuals				Repeated measurements			
	$\hat{\alpha}_i$	$S.e.(\hat{\alpha}_i)$	$\hat{\beta}_i$	$S.e.(\hat{\beta}_i)$	$\hat{\alpha}_i$	$S.e.(\hat{\alpha}_i)$	$\hat{\beta}_i$	$S.e.(\hat{\beta}_i)$
1	5.29	0.15	-0.016	0.006	5.30	0.10	-0.016	0.004
2	5.71	0.15	-0.010	0.007	5.71	0.11	-0.010	0.005
3	4.53	0.15	-0.028	0.006	4.53	0.11	-0.028	0.005
4	4.95	0.15	-0.011	0.006	4.95	0.11	-0.011	0.004
5	4.80	0.15	0.001	0.005	4.80	0.11	0.001	0.004
6	5.21	0.15	-0.012	0.005	5.21	0.11	-0.012	0.004
7	4.90	0.16	-0.019	0.009	4.90	0.11	-0.019	0.006
8	4.83	0.15	-0.021	0.009	4.83	0.10	-0.021	0.006
9	5.36	0.15	-0.021	0.007	5.36	0.10	-0.021	0.005
10	5.58	0.15	-0.017	0.007	5.58	0.10	-0.017	0.005

Table E: Estimates in the GLM model (independent residuals) and the repeated measurements model (autocorrelated residuals). The common slope estimated by OLS (model H_3) is -0.014.

|||| 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 many applications are often concerned with regression situations we will describe the results in this language. Some results will depend on the presence of an intercept term and are not true for the general linear model as such.

The sections on logistic regression and non-linear regression are short and are only meant to indicate the topics.

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 reasonable 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 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

$$D(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is known and σ^2 is (usually) unknown.

The estimators are found in the usual way by solving the normal equations

$$\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} \boldsymbol{\beta} = \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{Y},$$

or if $\boldsymbol{\Sigma} = \mathbf{I}$

$$\mathbf{x}^T \mathbf{x} \hat{\boldsymbol{\beta}} = \mathbf{x}^T \mathbf{Y}.$$

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

$$E(Y) = \alpha + \beta_1 t + \beta_2 \log t,$$

cf. example 2.2 or fig. 3.1. With n observations this model in matrix form becomes

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \log t_1 \\ \vdots & \vdots & \vdots \\ 1 & t_n & \log t_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

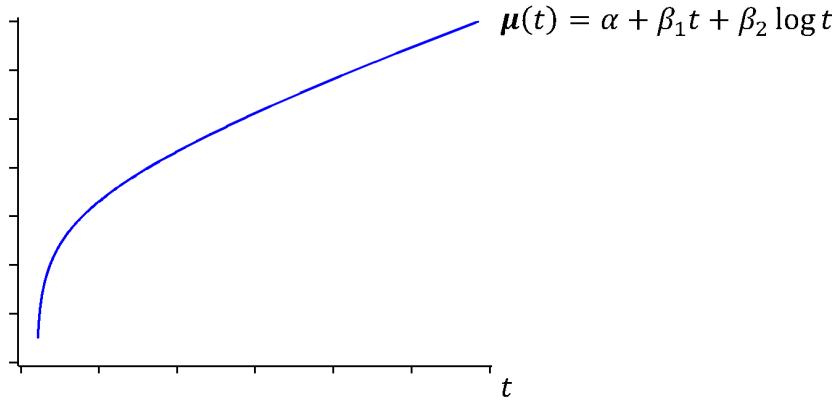


Figure 3.1 – The mean value function $\mu(t)$ is linear in the parameters α , β_1 , and β_2 (but obviously non-linear in t)

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

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & & \vdots \\ x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

3.1.2 Basic properties of regression residuals

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

$$\begin{aligned} \mathbf{H} &= \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \\ \mathbf{M} &= \mathbf{I} - \mathbf{H} \end{aligned}$$

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

$$\text{D}(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{H} \mathbf{H}^T = \sigma^2 \mathbf{H}$$

and

$$\text{D}(\mathbf{R}) = \sigma^2 \mathbf{M} \mathbf{M}^T = \sigma^2 \mathbf{M}.$$

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

$$\text{V}(\hat{Y}_i) = h_{ii}\sigma^2 = \left(\mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \right)_{ii} \sigma^2$$

and the residual variance becomes

$$\text{V}(R_i) = (1 - h_{ii})\sigma^2 = \left(\mathbf{I} - \mathbf{x}(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \right)_{ii} \sigma^2.$$

Furthermore it follows that *the residuals and the predicted values are uncorrelated* since

$$\text{Corr}(\hat{\mathbf{Y}}, \mathbf{R}) = \text{Corr}(\mathbf{H}\mathbf{Y}, \mathbf{M}\mathbf{Y}) = \sigma^2 \mathbf{H} \mathbf{M} = \mathbf{0}.$$

And finally we find that the sum of the residuals is 0. This follows from

$$\mathbf{1}^T \mathbf{R} = \mathbf{1}^T \mathbf{M} \mathbf{Y} = 0$$

since $\mathbf{1}^T$ is the first row in \mathbf{x}^T and we have

$$\mathbf{x}^T \mathbf{M} = \mathbf{x}^T - \mathbf{x}^T \mathbf{x} (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T = \mathbf{0}.$$

3.1.3 The Anatomy of Regression

In this section we consider properties of estimators in regression analysis. The title of the section is inspired by presentations in econometrics. Many of the

results are valid for the general linear model, but some depend on the inclusion of an intercept in the model, i.e. that the design matrix must include a column of ones. A substantial part plays on the relation between theoretical correlation measurements and their empirical analogues that are important empirical measures in interpreting and formulating regression models. The most important examples are the multiple correlation coefficients and the partial correlation coefficients, which here are supplemented with the semi-partial correlations. These parameters have a theoretical background from multivariate distributions and are defined through the expressions derived from conditional distributions in the multivariate normal distribution, cf. section 1.3.

In estimating variance parameters we shall mostly use maximum likelihood (ML) estimators and not the unbiased estimators, cf. Theorems 1.31 and 2.5. The reason for this is that when we are manipulating ML estimators, the outcome will be a ML estimator of the manipulated parameters, or – to put it into formulas – $\widehat{f(\theta)} = f(\hat{\theta})$. If we are looking at distributions of test statistics like $\hat{\theta} / \{\hat{V}(\hat{\theta})\}^{0.5}$, cf Theorem 2.23, we are using the unbiased estimator of the variance since this produces test statistics whose distribution is a standard t- or F-distribution.

The first problem we shall consider is an older theorem – originally due to [Frisch and Waugh \(1933\)](#) – widely used in econometrics. We consider a common regression situation

$$\mathbf{Y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad D(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$$

here \mathbf{x} and $\boldsymbol{\beta}$ are partitioned

$$\mathbf{Y} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\varepsilon} = \mathbf{x}_1 \boldsymbol{\beta}_1 + \mathbf{x}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

The ordinary least squares estimator (OLS) for $\boldsymbol{\beta}$ is obtained by solving the normal equations

$$\begin{bmatrix} \mathbf{x}_1^T \mathbf{x}_1 & \mathbf{x}_1^T \mathbf{x}_2 \\ \mathbf{x}_2^T \mathbf{x}_1 & \mathbf{x}_2^T \mathbf{x}_2 \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \end{bmatrix} \mathbf{Y}$$

||| Remark 3.1 Normal Equation

Given an equation $Ax = b$, the normal equation is $A^T Ax = A^T b$, which minimises the squared difference. The name derives from $b - Ax$ being normal to A .

The hat (influence) and residual (annihilation) matrices (= prediction and residual forming matrices) corresponding to x_1 are

$$H_1 = x_1(x_1^T x_1)^{-1} x_1^T$$

$$M_1 = I - H_1$$

||| Theorem 3.2 Frisch-Waugh-Lovell

Let the situation be as above. Then the OLS estimate of β_2 is

$$\hat{\beta}_2 = (x_2^T M_1 x_2)^{-1} x_2^T M_1 Y = ((M_1 x_2)^T M_1 x_2)^{-1} (M_1 x_2)^T M_1 Y$$

This is the OLS solution to the regression problem

$$M_1 Y = M_1 x_2 \beta_2 + \epsilon$$

or, using a tilde for residuals after regression on x_1 ,

$$\tilde{Y} = \tilde{x}_2 \beta_2 + \epsilon$$

Furthermore, the residual vectors

$$\begin{aligned}\hat{\epsilon} &= Y - x_1 \hat{\beta}_1 - x_2 \hat{\beta}_2 \\ \tilde{\epsilon} &= M_1 Y - M_1 x_2 \hat{\beta}_2 = \tilde{Y} - \tilde{x}_2 \hat{\beta}_2\end{aligned}$$

are equal, i.e. $\hat{\epsilon} = \tilde{\epsilon}$.

|||| Proof

We have

$$\begin{aligned} \mathbf{x}_1^T \mathbf{x}_1 \hat{\beta}_1 + \mathbf{x}_1^T \mathbf{x}_2 \hat{\beta}_2 &= \mathbf{x}_1^T \mathbf{Y} \\ \mathbf{x}_2^T \mathbf{x}_1 \hat{\beta}_1 + \mathbf{x}_2^T \mathbf{x}_2 \hat{\beta}_2 &= \mathbf{x}_2^T \mathbf{Y} \end{aligned}$$

Solving the first equation for $\hat{\beta}_1$ gives

$$\hat{\beta}_1 = (\mathbf{x}_1^T \mathbf{x}_1)^{-1} \mathbf{x}_1^T (\mathbf{Y} - \mathbf{x}_2 \hat{\beta}_2)$$

or

$$\mathbf{x}_1 \hat{\beta}_1 = \mathbf{H}_1 (\mathbf{Y} - \mathbf{x}_2 \hat{\beta}_2)$$

Inserting this in the other equation gives

$$\mathbf{x}_2^T \mathbf{H}_1 (\mathbf{Y} - \mathbf{x}_2 \hat{\beta}_2) + \mathbf{x}_2^T \mathbf{x}_2 \hat{\beta}_2 = \mathbf{x}_2^T \mathbf{Y}$$

Rearranging gives

$$\mathbf{x}_2^T \mathbf{M}_1 \mathbf{x}_2 \hat{\beta}_2 = \mathbf{x}_2^T \mathbf{M}_1 \mathbf{Y},$$

the desired result. The part on the residuals follows immediately by inserting the expression for $\mathbf{x}_1 \hat{\beta}_1$ in the expression for $\hat{\varepsilon}$.

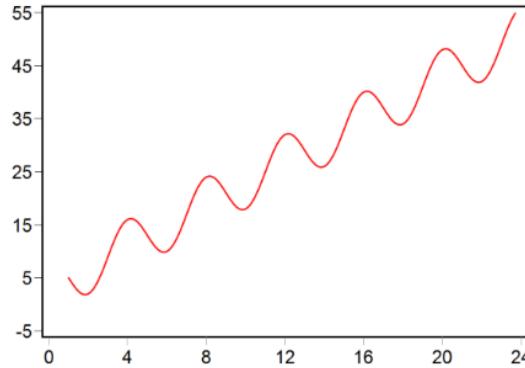
■

Here $\mathbf{M}_1 \mathbf{Y}$ is the residual vector after regressing \mathbf{Y} on \mathbf{x}_1 and $\mathbf{M}_1 \mathbf{x}_2$ are the columns of the residuals of the columns in \mathbf{x}_2 after regressing those on \mathbf{x}_1 . Thus we may obtain the least squares estimator of the last component of the parameter vector by regressing the residuals of \mathbf{Y} on the residuals of \mathbf{x}_2 , both sets of residuals are obtained after fitting with \mathbf{x}_1 . If β_2 is scalar, we may write

$$\hat{\beta}_2 = \frac{\tilde{\mathbf{Y}}^T \tilde{\mathbf{x}}_2}{\tilde{\mathbf{x}}_2^T \tilde{\mathbf{x}}_2} = \frac{\widehat{\mathbf{C}}(\tilde{\mathbf{Y}}, \tilde{\mathbf{x}}_2)}{\widehat{\mathbf{V}}(\tilde{\mathbf{x}}_2)}$$

||| Example 3.3

We consider (simulated) quarterly (econometric) data with a pronounced yearly cycle like in the figure below.



The data are – for $n=24$ and parameter vector $[3 \ 10 \ 2]^T$ – generated by the model

$$\begin{aligned} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} 1 & \cos(\frac{\pi}{2} t_1) & t_1 \\ \vdots & \vdots & \vdots \\ 1 & \cos(\frac{\pi}{2} t_n) & t_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \left[\begin{bmatrix} 1 & \cos(\frac{\pi}{2} t_1) \\ \vdots & \vdots \\ 1 & \cos(\frac{\pi}{2} t_n) \end{bmatrix} \begin{bmatrix} t_1 \\ \vdots \\ t_n \end{bmatrix} \right] \begin{bmatrix} \alpha \\ \beta \\ \beta_2 \end{bmatrix} + \varepsilon \\ &= [x_1 \ x_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon \end{aligned}$$

Y	const	$\cos(\pi t/2)$	t	Y	const	$\cos(\pi t/2)$	t
16.73	1	1	1	38.85	1	1	13
8.62	1	0	2	31.91	1	0	14
-3.09	1	-1	3	21.79	1	-1	15
8.89	1	0	4	34.40	1	0	16
19.50	1	1	5	47.66	1	1	17
16.13	1	0	6	40.18	1	0	18
4.07	1	-1	7	32.87	1	-1	19
19.52	1	0	8	41.89	1	0	20
29.69	1	1	9	53.10	1	1	21
23.83	1	0	10	48.56	1	0	22
14.88	1	-1	11	37.73	1	-1	23
23.34	1	0	12	51.19	1	0	24

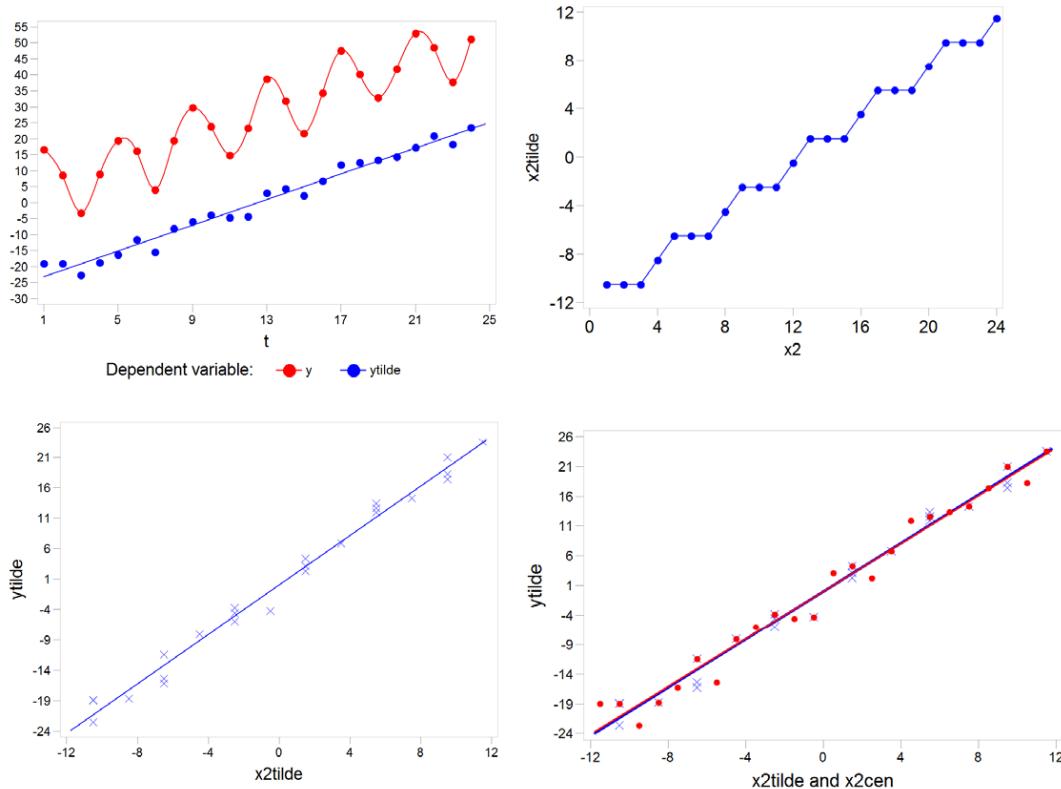
Having data like that makes it difficult to distinguish between a long term trend and the variation due to the cyclical behavior of the observations. An increase from

one quarter to the next quarter may be attributed to either a long term growth or just due to a fact like winter data are lower than spring data. In econometrics, a standard approach in analyzing data showing such strong periodicity is to de-seasonalize the data by estimating the periodic component and remove that from the data.

If we estimate all parameters in the model by OLS, we obtain

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2.1953891 \\ 10.137735 \\ 2.0319128 \end{bmatrix}$$

If we estimate the trend from the data corrected for seasonal variation, we obtain the slope 2.01, only deviating 1% from the OLS estimate. If we want to obtain the OLS estimate, we must also look at the residual of x_2 after regression on x_1 . Obviously, the ability of the cosine to predict the line is limited, and we see from Figure B below that - besides the centering - there are only minute differences between x_2 and \tilde{x}_2 . According to Frisch-Waugh-Lovell's theorem, the OLS estimate of β_2 is obtained by regressing \tilde{Y} on \tilde{x}_2 , not x_2 . The result of this is given in Figure C, and the two regression lines are compared in Figure D. We see that they are barely distinguishable.



A (top left): The raw data Y_i (red) as function of quarter number. Intermediate values are spline interpolated. In blue, the data corrected for seasonal variation, i.e. the residuals \tilde{Y}_i from regression on x_1 . The slope of the regression line is 2.0107102.
 B (top right): The residuals \tilde{x}_2 of the quarter numbers x_2 after regression on x_1 . C

(bottom left): The residuals \tilde{Y}_i as function of the residuals \tilde{x}_2 . The slope of the line is the OLS estimate $\hat{\beta}_2$. D (bottom right): The residuals \tilde{Y}_i as function of the residuals \tilde{x}_2 (blue x) and as function of the centered x_2 -values (red dot).

3.1.4 Estimation of partial and multiple correlations

We consider a normally distributed $k+2$ dimensional random variable

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{X} \end{bmatrix} = \begin{bmatrix} Y \\ V \\ X_1 \\ \vdots \\ X_k \end{bmatrix} \sim N_{2+k} \left(\begin{bmatrix} \boldsymbol{\mu}_w \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wx} \\ \boldsymbol{\Sigma}_{xw} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right).$$

with

$$\mathbf{D} \left(\begin{bmatrix} \mathbf{W} \\ \mathbf{X} \end{bmatrix} \right) = \begin{bmatrix} \boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wx} \\ \boldsymbol{\Sigma}_{xw} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} \sigma_y^2 & \sigma_{yv} \\ \sigma_{vy} & \sigma_v^2 \\ \sigma_{xy} & \sigma_{xv} \end{bmatrix} & \begin{bmatrix} \sigma_{yx} \\ \sigma_{vx} \end{bmatrix} \\ \begin{bmatrix} \sigma_{yx} & \sigma_{vx} \\ \Sigma_{xx} \end{bmatrix} \end{bmatrix}.$$

We assume that we have n independent observations from this distribution, i.e.

$$\begin{bmatrix} \mathbf{W}_i \\ \mathbf{X}_i \end{bmatrix} = \begin{bmatrix} Y_i \\ V_i \\ X_{i1} \\ \vdots \\ X_{ik} \end{bmatrix}, \quad i = 1, \dots, n$$

We organize the observations in the data matrix

$$[\mathbf{W} \quad \mathbf{X}] = \begin{bmatrix} \mathbf{W}_1^T & \mathbf{X}_1^T \\ \vdots & \vdots \\ \mathbf{W}_n^T & \mathbf{X}_n^T \end{bmatrix} = \begin{bmatrix} Y_1 & V_1 & X_{11} & \cdots & X_{1k} \\ \vdots & & & & \vdots \\ Y_n & V_n & X_{n1} & \cdots & X_{nk} \end{bmatrix}$$

The ML estimator for the mean is the average

$$\begin{bmatrix} \hat{\mu}_w \\ \hat{\mu}_x \end{bmatrix} = \frac{1}{n} [\mathbf{W} \quad \mathbf{X}]^T \mathbf{1} = \begin{bmatrix} \bar{W} \\ \bar{X} \end{bmatrix} = \begin{bmatrix} \bar{Y} \\ \bar{V} \\ \bar{X}_1 \\ \vdots \\ \bar{X}_k \end{bmatrix}$$

and the ML estimator for the dispersion matrix is given by

$$\begin{aligned} n\hat{\Sigma} &= n \begin{bmatrix} \hat{\Sigma}_{ww} & \hat{\Sigma}_{wx} \\ \hat{\Sigma}_{xw} & \hat{\Sigma}_{xx} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{W}^T \\ \mathbf{X}^T \end{bmatrix} [\mathbf{W} \quad \mathbf{X}] - n \begin{bmatrix} \bar{W} \\ \bar{X} \end{bmatrix} [\bar{W}^T \quad \bar{X}^T] \\ &= \begin{bmatrix} \mathbf{W}^T \mathbf{W} - n \bar{W} \bar{W}^T & \mathbf{W}^T \mathbf{X} - n \bar{W} \bar{X}^T \\ \mathbf{X}^T \mathbf{W} - n \bar{X} \bar{W}^T & \mathbf{X}^T \mathbf{X} - n \bar{X} \bar{X}^T \end{bmatrix} \\ &= n \begin{bmatrix} \hat{D}(\mathbf{W}_j) & \hat{C}(\mathbf{W}_j, \mathbf{X}_j) \\ \hat{C}(\mathbf{X}_j, \mathbf{W}_j) & \hat{D}(\mathbf{X}_j) \end{bmatrix} \end{aligned}$$

For an arbitrary j , the estimated conditional dispersion matrix for \mathbf{W}_j given \mathbf{X}_j is

$$\begin{aligned} \hat{D}(\mathbf{W}_j \mid \mathbf{X}_j = \mathbf{x}_j) &= \hat{\Sigma}_{ww} - \hat{\Sigma}_{wx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xw} \\ &= \hat{D}(\mathbf{W}_j) - \hat{C}(\mathbf{W}_j, \mathbf{X}_j) \left\{ \hat{D}(\mathbf{X}_j) \right\}^{-1} \hat{C}(\mathbf{X}_j, \mathbf{W}_j). \end{aligned}$$

From this expression we may determine the **estimated (or) partial correlation** between Y_j and V_j given X_j as the correlation in this conditional distribution using the definition given in section 1.3.1:

$$\hat{\rho}_{yv|x} = \frac{\hat{\sigma}_{yv} - \hat{\sigma}_{yx} \hat{\Sigma}_{xx}^{-1} \hat{\sigma}_{xv}}{\hat{\sigma}_{y|x} \hat{\sigma}_{v|x}}$$

where

$$\hat{\sigma}_{y|x}^2 = \hat{\sigma}_y^2 - \hat{\sigma}_{yx} \hat{\Sigma}_{xx}^{-1} \hat{\sigma}_{xy}$$

$$\widehat{\sigma}_{v|x}^2 = \widehat{\sigma}_v^2 - \widehat{\sigma}_{vx} \widehat{\Sigma}_{xx}^{-1} \widehat{\sigma}_{xv}$$

We shall now consider another approach leading to determination of the empirical partial correlation.

|||| Theorem 3.4

Let the situation be as above. Then the empirical partial correlation between Y_j and V_j is the same as the sample correlation between the residuals of Y_j and V_j , $j = 1, \dots, n$, after regressing those on \mathbf{X} .

|||| Proof

Let Z be either of the variables Y or V and \mathbf{Z} the vector of observed values $[Z_1 \ \dots \ Z_n]^T$. We consider the regression model M

$$\begin{bmatrix} Z_1 \\ \vdots \\ Z_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \alpha_z \\ \beta_z \end{bmatrix} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & \mathbf{X}_1^T \\ \vdots & \vdots \\ 1 & \mathbf{X}_n^T \end{bmatrix}$$

or

$$\mathbf{Z} = [\mathbf{1} \ \mathbf{X}] \begin{bmatrix} \alpha_z \\ \beta_z \end{bmatrix} + \boldsymbol{\varepsilon}$$

i.e.

$$Y = [\mathbf{1} \ \mathbf{X}] \begin{bmatrix} \alpha_y \\ \beta_y \end{bmatrix} + \boldsymbol{\varepsilon}_y$$

$$V = [\mathbf{1} \ \mathbf{X}] \begin{bmatrix} \alpha_v \\ \beta_v \end{bmatrix} + \boldsymbol{\varepsilon}_v$$

The normal equations for estimating the parameters α and β are

$$\begin{bmatrix} \mathbf{1}^T \\ \mathbf{X}^T \end{bmatrix} [\mathbf{1} \ \mathbf{X}] \begin{bmatrix} \widehat{\alpha}_z \\ \widehat{\beta}_z \end{bmatrix} = \begin{bmatrix} \mathbf{1}^T \\ \mathbf{X}^T \end{bmatrix} \mathbf{Z}$$

or

$$\begin{bmatrix} n & n\bar{X}^T \\ n\bar{X} & \mathbf{X}^T \mathbf{X} \end{bmatrix} \begin{bmatrix} \widehat{\alpha}_z \\ \widehat{\beta}_z \end{bmatrix} = \begin{bmatrix} n\bar{Z} \\ \mathbf{X}^T \mathbf{Z} \end{bmatrix},$$

which gives

$$\hat{\alpha}_z + \bar{X}^T \hat{\beta}_z = \bar{Z}$$

$$n\hat{\alpha}_z \bar{X} + X^T X \hat{\beta}_z = X^T Z$$

Inserting the first equation in the second gives

$$\left\{ X^T X - n \bar{X} \bar{X}^T \right\} \hat{\beta}_z = X^T Z - n \bar{X} \bar{Z}$$

or

$$\hat{\beta}_z = \left\{ \hat{D}(X_v) \right\}^{-1} \hat{C}(X_v, Z_v).$$

The residual vectors are $R_z^T = [R_{1z} \ \cdots \ R_{nz}]^T$ where

$$R_y = Y - [\mathbf{1} \ \ X] \begin{bmatrix} \hat{\alpha}_u \\ \hat{\beta}_u \end{bmatrix}$$

$$R_v = V - [\mathbf{1} \ \ X] \begin{bmatrix} \hat{\alpha}_v \\ \hat{\beta}_v \end{bmatrix}$$

with coordinate no i

$$R_{yi} = Y_i - [1 \ \ X_i^T] \begin{bmatrix} \hat{\alpha}_y \\ \hat{\beta}_y \end{bmatrix}$$

$$= Y_i - \bar{Y} - \left\{ X_i^T - \bar{X}^T \right\} \hat{\beta}_u$$

$$R_{vi} = V_i - \bar{V} - \left\{ X_i^T - \bar{X}^T \right\} \hat{\beta}_v$$

In order to determine the empirical correlation between R_{yi} and R_{vi} , the covariance may be determined as n times the inner product $R_u^T R_v$ between the two residual vectors, i.e.

$$\sum R_{yi} R_{vi} = \sum (Y_i - \bar{Y})(V_i - \bar{V}) - \sum (Y_i - \bar{Y}) \left\{ X_i^T - \bar{X}^T \right\} \hat{\beta}_v$$

$$- \sum \hat{\beta}_y^T (X_i - \bar{X})(V_i - \bar{V}) + \sum \hat{\beta}_y^T \{ X_i - \bar{X} \} \left\{ X_i^T - \bar{X}^T \right\} \hat{\beta}_v$$

or

$$\frac{1}{n} \sum R_{yi} R_{vi} = \hat{C}(Y_j, V_j) - \hat{C}(Y_j, X_j) \hat{\beta}_v - \hat{\beta}_u^T \hat{C}(X_j, V_j) + \hat{\beta}_u^T \hat{D}(X_j) \hat{\beta}_v$$

Now, the three last terms are all equal (with proper signs)

$$\widehat{C}(Y_j, X_j) \left\{ \widehat{D}(X_j) \right\}^{-1} \widehat{C}(X_j, V_j)$$

and thus

$$\frac{1}{n} \sum R_{yi} R_{vi} = \widehat{C}(Y_j, V_j) - \widehat{C}(Y_j, X_j) \left\{ \widehat{D}(X_j) \right\}^{-1} \widehat{C}(X_j, V_j) = \widehat{C}(Y_j, V_j | X_j),$$

i.e. the empirical conditional covariance between Y_j and V_j given X_j .

The squared length $R_z^T R_z$ of either residual vector is

$$\begin{aligned} \sum R_{zi}^2 &= \sum (Z_i - \bar{Z})^2 + \sum \widehat{\beta}_z^T \{X_i - \bar{X}\} \left\{ X_i^T - \bar{X}^T \right\} \widehat{\beta}_z - \sum (Z_i - \bar{Z}) \left\{ X_i^T - \bar{X}^T \right\} \widehat{\beta}_z \\ &\quad - \sum \widehat{\beta}_z^T (X_i - \bar{X})(Z_i - \bar{Z}) \end{aligned}$$

or

$$\frac{1}{n} \sum R_{zi}^2 = \widehat{V}(Z_j) + \widehat{C}(Z_j, X_j) \left\{ \widehat{D}(X_j) \right\}^{-1} \widehat{C}(X_j, Z_j) = \widehat{V}(Z_j | X_j)$$

i.e. the empirical conditional variance of Z_j given X_j . With this, we have established the theorem.

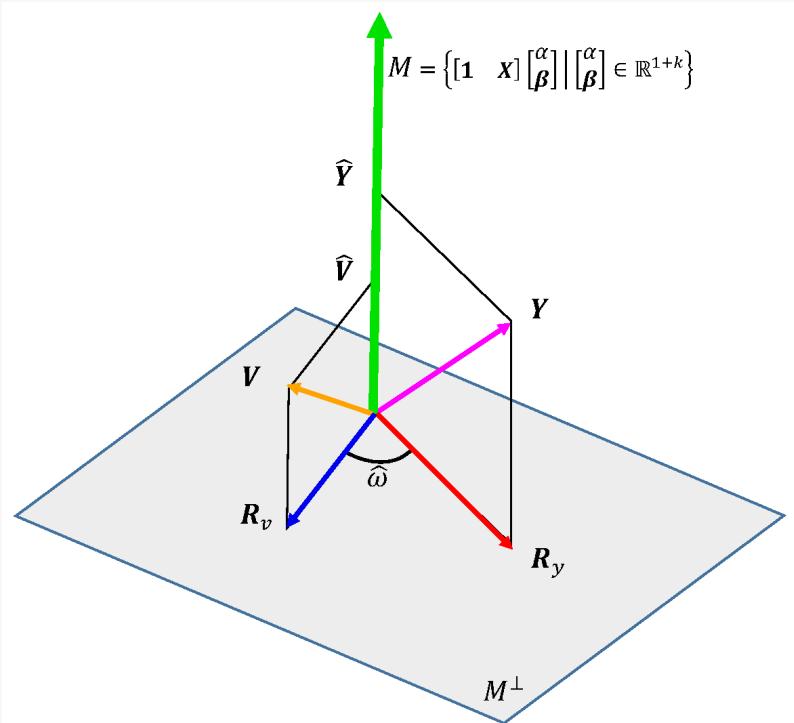
■

||| Remark 3.5 Partial correlation

With the notation from above, the empirical partial correlation becomes

$$\hat{\rho}_{yv|x} = \frac{\sum R_{yi}R_{vi}}{\left\{ \sum R_{yi}^2 \sum R_{vi}^2 \right\}^{1/2}} = \frac{\mathbf{R}_y^T \mathbf{R}_v}{\left\{ \mathbf{R}_y^T \mathbf{R}_y \right\}^{0.5} \left\{ \mathbf{R}_v^T \mathbf{R}_v \right\}^{0.5}} = \cos(\hat{\omega})$$

where $\hat{\omega}$ is the angle between the two residual vectors, see the figure below.



Geometric interpretation of partial correlation.

|||| Remark 3.6 Squared multiple correlation

For the **squared multiple correlation coefficient** we obtain (skipping the index on X so that X is the k dimensional random variable not to be mixed up with the $n \times k$ dimensional data matrix)

$$\rho_{y|x} = \frac{\sqrt{V(Y|X)}}{\sqrt{V(Y)}} = \frac{\sqrt{\sigma_{yx}\Sigma_{xx}^{-1}\sigma_{xy}}}{\sigma_y}$$

giving

$$\rho_{y|x}^2 = \frac{V(Y) - V(Y|X)}{V(Y)} = \frac{\sigma_y^2 - \sigma_{y|x}^2}{\sigma_y^2}$$

where

$$\sigma_{y|x}^2 = \sigma_y^2 - \sigma_{yx}\Sigma_{xx}^{-1}\sigma_{xy}$$

The empirical version becomes

$$\begin{aligned} \hat{\rho}_{y|x}^2 &= \frac{\hat{V}(Y) - \hat{V}(Y|X)}{\hat{V}(Y)} \\ &= \frac{\sum (Y_i - \bar{Y})^2 - \mathbf{R}_y^T \mathbf{R}_y}{\sum (Y_i - \bar{Y})^2} \\ &= \frac{(\mathbf{Y} - \mathbf{1}\bar{Y})^T (\mathbf{Y} - \mathbf{1}\bar{Y}) - \mathbf{R}_y^T \mathbf{R}_y}{(\mathbf{Y} - \mathbf{1}\bar{Y})^T (\mathbf{Y} - \mathbf{1}\bar{Y})} \\ &= \frac{SS_{Tot(y)} - SS_{Res(y|x)}}{SS_{Tot(y)}} \end{aligned}$$

3.1.5 The partial F-test and partial correlations

We now consider the regression model M

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} & V_1 \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} & V_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} + \varepsilon$$

and want to investigate whether we may accept the hypothesis $H: \gamma = 0$. This

may be done using standard techniques by first estimating the full model and then the model without γ . The test statistic then becomes the difference between the residual sums of squares divided by the residual sum of squares for the full model – all appropriately normed with the relevant degrees of freedom, cf Theorem 2.21. This will give an F-statistic with $(1, n - k - 2)$ degrees of freedom. We shall now investigate how this test statistic relates to the terminology introduced in the previous section.

We use Frisch-Waugh-Lovell's theorem (3.2) and see that γ can be estimated from the regression of the residuals \mathbf{R}_y on the residuals \mathbf{R}_v , i.e.

$$\mathbf{R}_y = \mathbf{R}_v\gamma + \boldsymbol{\epsilon}$$

or

$$\hat{\gamma} = \frac{\sum R_{yi}R_{vi}}{\sum R_{iv}^2} = \frac{\mathbf{R}_y^T \mathbf{R}_v}{\mathbf{R}_v^T \mathbf{R}_v} = \frac{\{\mathbf{R}_y^T \mathbf{R}_y\}^{0.5}}{\{\mathbf{R}_v^T \mathbf{R}_v\}^{0.5}} \hat{\rho}_{yv|x} = \frac{\hat{\sigma}_{y|x}}{\hat{\sigma}_{v|x}} \hat{\rho}_{yv|x}$$

The *unbiased* estimator of the variance is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - k - 2} (\mathbf{R}_y - \mathbf{R}_v\hat{\gamma})^T (\mathbf{R}_y - \mathbf{R}_v\hat{\gamma}) \\ &= \frac{1}{n - k - 2} \left\{ \mathbf{R}_y^T \mathbf{R}_y - \frac{\{\mathbf{R}_y^T \mathbf{R}_v\}^2}{\mathbf{R}_v^T \mathbf{R}_v} \right\} = \frac{1}{n - k - 2} \frac{(\mathbf{R}_y^T \mathbf{R}_y)(\mathbf{R}_v^T \mathbf{R}_v) - (\mathbf{R}_y^T \mathbf{R}_v)^2}{\mathbf{R}_v^T \mathbf{R}_v}\end{aligned}$$

and the estimated uncertainty on $\hat{\gamma}$ becomes

$$\hat{V}(\hat{\gamma}) = \hat{\sigma}^2 \frac{1}{\mathbf{R}_v^T \mathbf{R}_v}$$

The F-statistic for testing whether $\gamma = 0$ is equal to the squared t-test statistic for the same hypothesis and is equal to

$$\frac{\hat{\gamma}^2}{\hat{V}(\hat{\gamma})} = (n - 2 - k) \frac{(\mathbf{R}_y^T \mathbf{R}_v)^2}{(\mathbf{R}_y^T \mathbf{R}_y)(\mathbf{R}_v^T \mathbf{R}_v) - (\mathbf{R}_y^T \mathbf{R}_v)^2}$$

On the other hand, the test statistic for assessing whether $\rho_{yv|x}$ is equal to 0 is

$$\frac{\hat{\rho}_{yv|x}}{\sqrt{1 - \hat{\rho}_{yv|x}^2}} \sqrt{n - 2 - k},$$

which may be compared to quantiles in a t-distribution with $n - 2 - k$ degrees of freedom, cf. Theorem 1.37. The corresponding F-test statistic is the square of this and equals

$$(n - 2 - k) \frac{\hat{\rho}_{yv|x}^2}{1 - \hat{\rho}_{yv|x}^2} = (n - 2 - k) \frac{(\mathbf{R}_y^T \mathbf{R}_v)^2}{(\mathbf{R}_y^T \mathbf{R}_y)(\mathbf{R}_v^T \mathbf{R}_v) - (\mathbf{R}_y^T \mathbf{R}_v)^2},$$

i.e. the same as before. We have thus established

|||| Theorem 3.7

The F-test statistic for testing whether a single parameter in a regression model is equal to 0 is equivalent to the test statistic for testing whether the partial correlation between the dependent variable and the independent variable considered conditioned on all other independent variables is equal to 0. Sometimes this F-statistic is called the partial F-test.

3.1.6 The semi-partial correlation coefficient

We introduce the semi-partial correlation between V and U given X in

|||| Definition 3.8

With the same notation as in the previous sections, we define the **semi-partial correlation** between Y and $(V \text{ given } X = [X_1 \ \dots \ X_k]^T)$ as the correlation between Y and the residual of V after conditioning (regressing) on X , i.e.

$$\rho_{y(v|x)} = \text{Cor}(Y, V - E(V | X)).$$

There is a close relation to the partial correlation coefficient as we have

|||| **Theorem 3.9**

The semi-partial correlation is related to the partial correlation through

$$\rho_{y(v|x)} = \rho_{yv|x} \frac{\sigma_{y|x}}{\sigma_y}$$

If \mathbf{X} is one-dimensional, we have

$$\rho_{y(v|x)} = \frac{\rho_{yv} - \rho_{yx}\rho_{vx}}{\sqrt{(1 - \rho_{yx}^2)(1 - \rho_{vx}^2)}} = \rho_{yv|x} \sqrt{(1 - \rho_{yx}^2)}$$

|||| **Proof**

$$\text{Cov}(Y, V - E(V | \mathbf{X})) = \text{Cov}(Y - E(Y | \mathbf{X}), V - E(V | \mathbf{X})) + \text{Cov}(E(Y|\mathbf{X}), V - E(V | \mathbf{X}))$$

$$= \rho_{yv|x} \sigma_{y|x} \sigma_{v|x} + \text{Cov}(\sigma_{yx} \Sigma_{xx}^{-1} \mathbf{X}, V - \sigma_{vx} \Sigma_{xx}^{-1} \mathbf{X})$$

The last term is

$$\sigma_{yx} \Sigma_{xx}^{-1} \sigma_{xv} - \sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xx} \Sigma_{xx}^{-1} \sigma_{xv} = 0$$

Therefore

$$\rho_{y(v|x)} = \text{Cor}(Y, V - E(V | \mathbf{X})) = \rho_{yv|x} \sigma_{y|x} \sigma_{v|x} \frac{1}{\sigma_y \sigma_{v|x}} = \rho_{yv|x} \frac{\sigma_{y|x}}{\sigma_y}$$

Since the partial correlation between Y and V given a one dimensional X is

$$\rho_{yv|x} = \frac{\rho_{yv} - \rho_{yx}\rho_{vx}}{\sqrt{(1 - \rho_{yx}^2)(1 - \rho_{vx}^2)}}$$

and the conditional variance of Y given X is

$$V(Y | X) = \sigma_{y|x}^2 = \sigma_y^2(1 - \rho_{yx}^2),$$

the last statement follows immediately.

|||| **Remark 3.10**

The empirical version becomes

$$\hat{\rho}_{y(v|x)} = \hat{\rho}_{yv|x} \frac{\hat{\sigma}_{y|x}}{\hat{\sigma}_y} = \frac{\mathbf{R}_y^T \mathbf{R}_v}{\{\mathbf{R}_v^T \mathbf{R}_v\}^{0.5} \left\{ (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}})^T (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}}) \right\}^{0.5}}$$

The regression coefficient expressed by the semi-partial correlation is (and repeating the expression involving the partial correlation)

$$\hat{\gamma} = \frac{\hat{\sigma}_{y|x}}{\hat{\sigma}_{v|x}} \hat{\rho}_{yv|x} = \frac{\hat{\sigma}_y}{\hat{\sigma}_{v|x}} \hat{\rho}_{y(v|x)}$$

Now we consider the difference between the squared multiple correlation between \mathbf{Y} and $[\mathbf{X}^T \mathbf{V}]^T$ and between \mathbf{Y} and \mathbf{X} , ie.

$$\hat{\rho}_{y|xv}^2 - \hat{\rho}_{y|x}^2 = \frac{\widehat{\text{V}}(\mathbf{Y}) - \widehat{\text{V}}(\mathbf{Y}|\mathbf{X}\mathbf{V}) - \widehat{\text{V}}(\mathbf{Y}) + \widehat{\text{V}}(\mathbf{Y}|\mathbf{X})}{\widehat{\text{V}}(\mathbf{Y})} = \frac{\widehat{\text{V}}(\mathbf{Y}|\mathbf{X}) - \widehat{\text{V}}(\mathbf{Y}|\mathbf{X}\mathbf{V})}{\widehat{\text{V}}(\mathbf{Y})}$$

According to Frisch-Waugh-Lovell's theorem (3.2) we for the maximum likelihood estimators get

$$n \left\{ \widehat{\text{V}}(\mathbf{Y}|\mathbf{X}) - \widehat{\text{V}}(\mathbf{Y}|\mathbf{X}\mathbf{V}) \right\} = \mathbf{R}_y^T \mathbf{R}_y - \left\{ \mathbf{R}_y^T \mathbf{R}_y - \frac{\{\mathbf{R}_u^T \mathbf{R}_v\}^2}{\mathbf{R}_v^T \mathbf{R}_v} \right\}$$

obtaining

$$\hat{\rho}_{y|xv}^2 - \hat{\rho}_{y|x}^2 = \frac{\{\mathbf{R}_y^T \mathbf{R}_v\}^2}{\{\mathbf{R}_v^T \mathbf{R}_v\} \left\{ (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}})^T (\mathbf{Y} - \mathbf{1}\bar{\mathbf{Y}}) \right\}} = \hat{\rho}_{y(x|v)}^2$$

With this we have proven

|||| Theorem 3.11

The squared semi-partial correlation between U and $(V \text{ given } X)$ is equal to the increase in the squared multiple correlation of U given X by including V as regressor, i.e. regressing on X and V , i.e.

$$\hat{\rho}_{y(v|x)}^2 = \hat{\rho}_{y|xv}^2 - \hat{\rho}_{y|x}^2$$

We shall now briefly introduce a concept that is important in interpreting regression expressions, name variables not correlated with the independent variable, but still important in contributing to the overall predictability of the independent variable Y .

|||| Definition 3.12

The variable V is a *suppressor variable* if

$$\hat{\rho}_{y|xv}^2 > \hat{\rho}_{y|x}^2 + \hat{\rho}_{yv}^2 \iff \hat{\rho}_{y(v|x)}^2 > \hat{\rho}_{yv}^2$$

We shall not go into a detailed discussion on the implications of being a suppressor variable but just indicate the relations. For a deeper discussion see e.g. [Pandey and Elliott \(2010\)](#) and [Velicer \(1978\)](#). Shortly, we may say that a suppressor variable in general has a low correlation with the dependent variable but is strongly correlated with some of the independent variables (predictor variables) that again are correlated with the dependent variable. The suppressor variable will ‘explain’ the irrelevant part of the variation in the independent variables, and therefore, including them in the regression, will lead to an improved overall prediction. The word suppression is due to that the suppressor variable suppresses (compensates for) irrelevant variation in some predictor variables.

3.2 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.

3.2.1 Analysis of residuals

An important way of checking whether a model adequately describes the observations are an analysis of the distribution of the residuals from the model.

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 in figure 3.2 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 sketch 1 always is acceptable.

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)

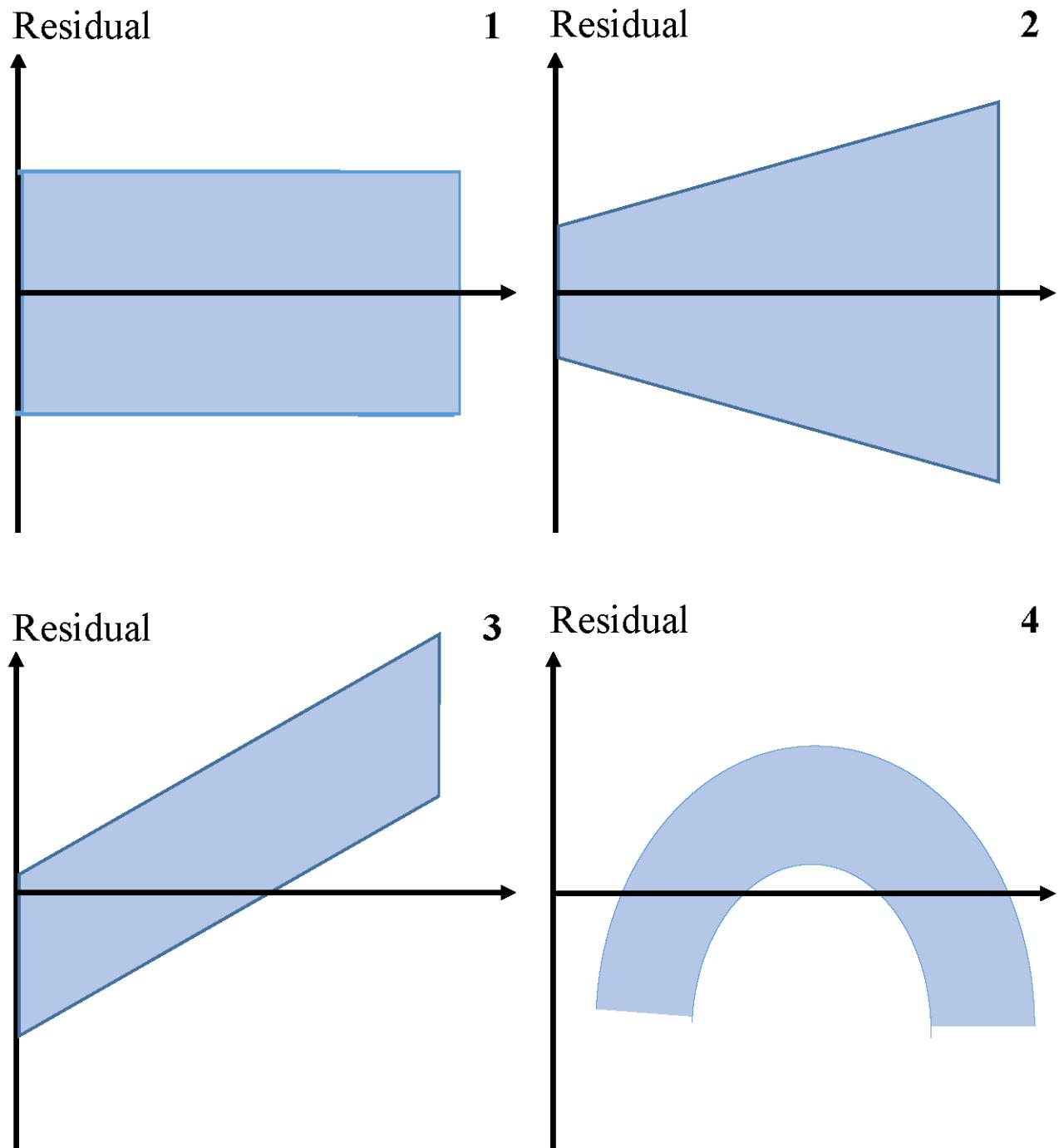


Figure 3.2 – Residual plots. The shaded areas indicate the regions where the residuals are located. Interpretations for different choices of abscissa are given in the text.

- 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
- 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 3.13

When analyzing diagnostic plots, the results in the table below should be taken into account.

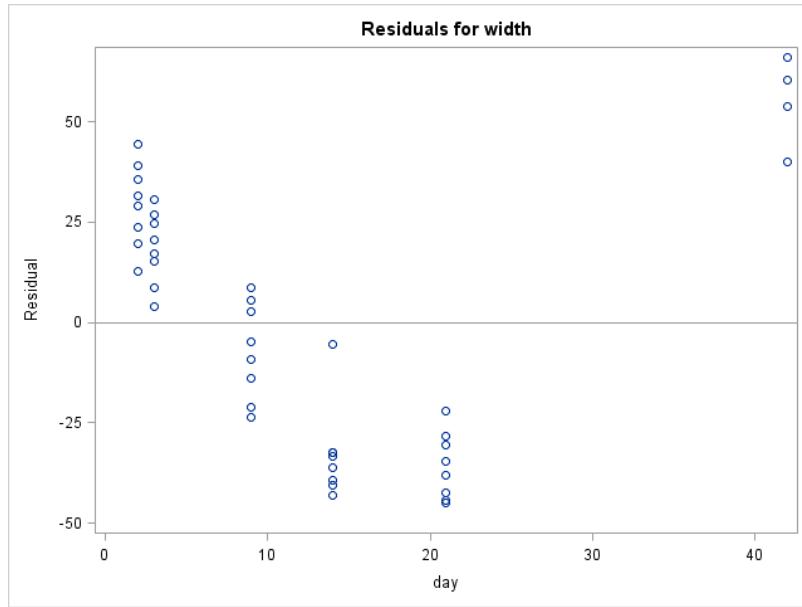
Plotting	Estimated slope	Estimated intercept	Estimated R-Square
(\hat{Y}_i, Y_i)	1	0	R^2
(\hat{Y}_i, R_i)	0	0	0
(Y_i, R_i)	$1 - R^2$	$-(1 - R^2)\bar{Y}$	$1 - R^2$

It follows that a plot of predicted values versus observed values should be distributed around a line with slope 1. A plot of the predicted value versus the residual should vary around a horizontal line, whereas a plot of the observed values versus residuals will be distributed around a line with slope $(1 - R)^2$. Maybe the last statement is a bit counter intuitive but think of a situation where $R^2=0$. This means that the predicted value is constantly equal to the average \bar{Y} . Therefore the residual is equal to the observation minus a constant meaning that all points fall on a straight line with slope 1. Therefore a plot of observations versus residuals is less useful.

We will illustrate the analysis of residuals with a simple example

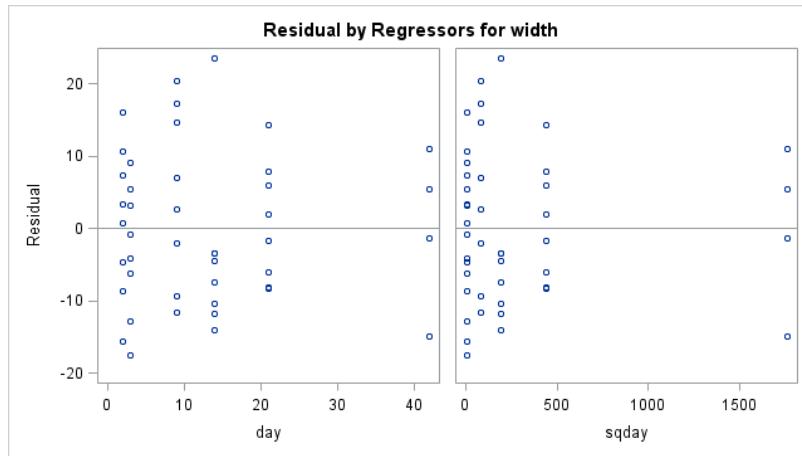
||| Example 3.14

We want to estimate the dependant variable "width" as a function of "day". We plot the residuals against "day", i.e. case iii) from residual analysis.



Residuals in regression model describing the dependent variable "width" as a linear function of the independent variable "day"

We see a pattern of residuals very akin to sketch 4 in figure 3.2. This indicates we need a quadratic term in our model. We add that and inspect the residuals again.



Residuals after fitting the same data, but now as a linear function of "day" and "day squared"

The residuals after adding a quadratic term are more akin to sketch 1 in figure 3.2, and it seems we have a well specified model.

3.2.2 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 3.2 above. Here we will briefly consider 2).

The deletion formula

Re-calculation of parameter estimates when discarding a single observation can be done using the formula

$$(\mathbf{A} - \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^T\mathbf{A}^{-1}}{1 - \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}},$$

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 $\mathbf{A} = \mathbf{x}^T\mathbf{x}$ and $\mathbf{u} = \mathbf{v} = \mathbf{x}_i^T$ we get

$$(\mathbf{x}^T\mathbf{x} - \mathbf{x}_i^T\mathbf{x}_i)^{-1} = (\mathbf{x}^T\mathbf{x})^{-1} + \frac{(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}_i^T\mathbf{x}_i(\mathbf{x}^T\mathbf{x})^{-1}}{1 - h_{ii}}$$

If we denote the \mathbf{x} -matrix where the i 'th row is removed $\mathbf{x}(i)$ we have that

$$\mathbf{x}(i)^T\mathbf{x}(i) = \mathbf{x}^T\mathbf{x} - \mathbf{x}_i^T\mathbf{x}_i.$$

The proof is omitted.

We can now state the relevant expressions.

Cook's D

A confidence region for the parameter $\boldsymbol{\theta}$ is all the vectors $\boldsymbol{\theta}^*$, which satisfy

$$\frac{1}{p\hat{\sigma}^2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^T\mathbf{x}^T\mathbf{x}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \leq F(p, n-p)_{1-\alpha}.$$

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

$$\mathbf{y}(i) = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)^T$$

and therefore have

$$\hat{\theta}(i) = [\mathbf{x}(i)^T \mathbf{x}(i)]^{-1} \mathbf{x}(i)^T \mathbf{y}(i).$$

Cook's D then equals

$$\frac{1}{p\hat{\sigma}^2} (\hat{\theta} - \hat{\theta}(i))^T \mathbf{x}^T \mathbf{x} (\hat{\theta} - \hat{\theta}(i)).$$

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 **STUDENT RESIDUAL**, 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

$$\text{COVRATIO}_i = \frac{\det[\hat{\sigma}(i)^2(\mathbf{x}(i)^T \mathbf{x}(i))^{-1}]}{\det[\hat{\sigma}^2(\mathbf{x}^T \mathbf{x})^{-1}]}$$

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

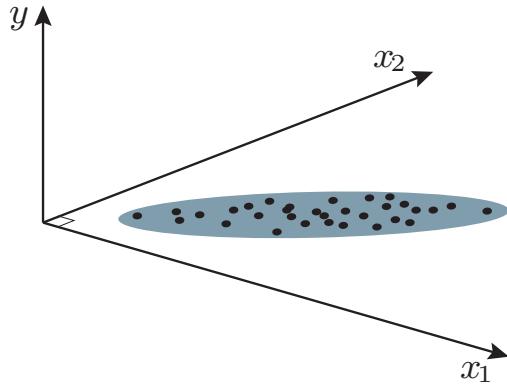
$$h_{ii} = \mathbf{x}_i(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}_i^T$$

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 $\pm 2\sqrt{p/(n-p)}$.

$$\begin{aligned} \text{DFFITS} &= \frac{\hat{y}_i - \hat{y}(i)}{\hat{\sigma}(i)\sqrt{h_{ii}}} \\ &= \frac{\mathbf{x}_i[\hat{\theta} - \hat{\theta}(i)]}{\hat{\sigma}(i)\sqrt{h_{ii}}}. \end{aligned}$$



If we have a model

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

and must estimate β_1 β_2 that can not be done in a reliable way since we cannot vary one x with the other fixed.

Figure 3.3 – All the (x_1, x_2) are in the shaded (blue) area

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 $\pm 2/\sqrt{n}$.

$$\text{DFBETAS}_j = \frac{\hat{\theta}_j - \hat{\theta}(i)_j}{\hat{\sigma}(i) \sqrt{(\mathbf{x}^T \mathbf{x})_{jj}^{-1}}}.$$

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 3.3. 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 $\mathbf{x}\mathbf{x}^T$.

||| Definition 3.15

We define the *tolerance (TOL)* and the *variance inflation (VIF)* as

$$\begin{aligned} \text{TOL}_i &= 1 - R^2(x_i | \text{all other } x\text{-variables}) \\ \text{VIF}_i &= \frac{1}{\text{TOL}_i} \end{aligned}$$

As a rule of thumb, $\text{TOL} < 0.1$ or equivalently $\text{VIF} > 10$ indicates a multi-collinearity problem.

||| Definition 3.16

We define the *condition number* as the square root of the largest eigenvalue of $\mathbf{x}\mathbf{x}^T$ 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 tol vif collin;
```

Model statements etc. are the same in REG as in GLM. The diagnostic tests come with the options / `r influence tol vif collin`.

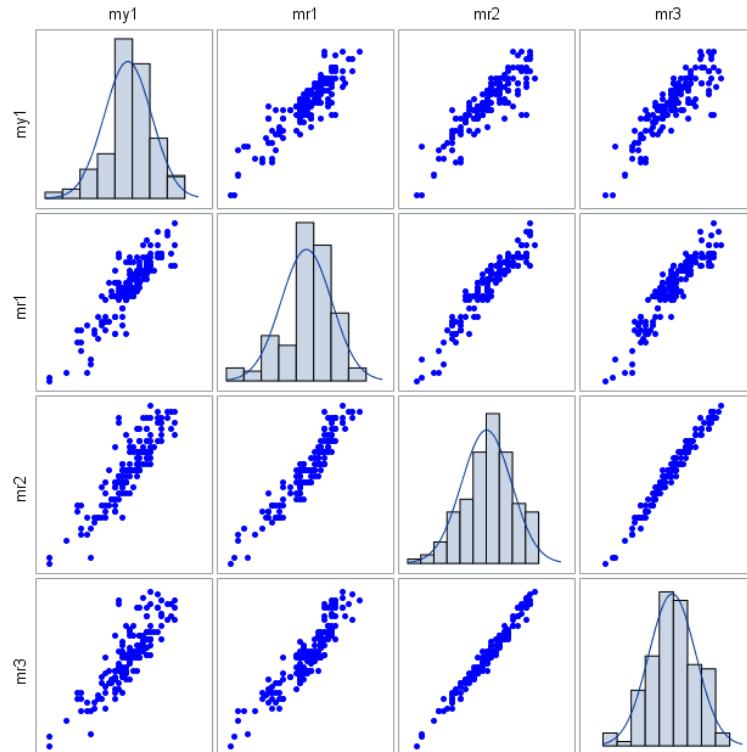
We will illustrate some of the diagnostics provided by SAS in the following

||| Example 3.17

We consider 148 randomly chosen pixel values from two Landsat images of the same area in India, the first collected in March, the second in May. We have shown the first 10 observations in the table below. Each row represents measured reflections in given wavelengths from a pixel in the image, the first six (mr1-mr6) from the March image, the following six (my1-my6) from the May image. The variables mr1-mr3 and my1-my3 are in the visible range. There is a more thorough description of the satellite images in Section 6.2.

Obs	mr1	mr2	mr3	mr4	mr5	mr6	my1	my2	my3	my4	my5	my6
1	162	89	135	105	210	144	137	71	105	83	172	120
2	161	88	136	104	210	143	134	69	104	79	172	119
3	165	92	142	109	212	148	143	78	120	93	189	133
4	166	93	144	109	216	150	142	79	119	93	187	133
5	163	94	145	111	215	151	146	79	120	93	190	133
6	165	93	146	111	216	152	143	78	121	94	188	136
7	164	93	144	109	215	152	139	75	115	90	189	131
8	164	90	138	106	216	147	133	69	102	79	172	118
9	166	93	145	111	218	153	140	71	107	84	177	122
10	158	82	123	96	201	135	132	65	94	74	163	108

We now want to investigate how well the first visible channel my1 from the May image may be described as an affine function of the corresponding three variables mr1-mr3 from the March image. Before any further analysis we plot the four histograms and the scatterplots in the figure below. We see that all variables are well correlated.



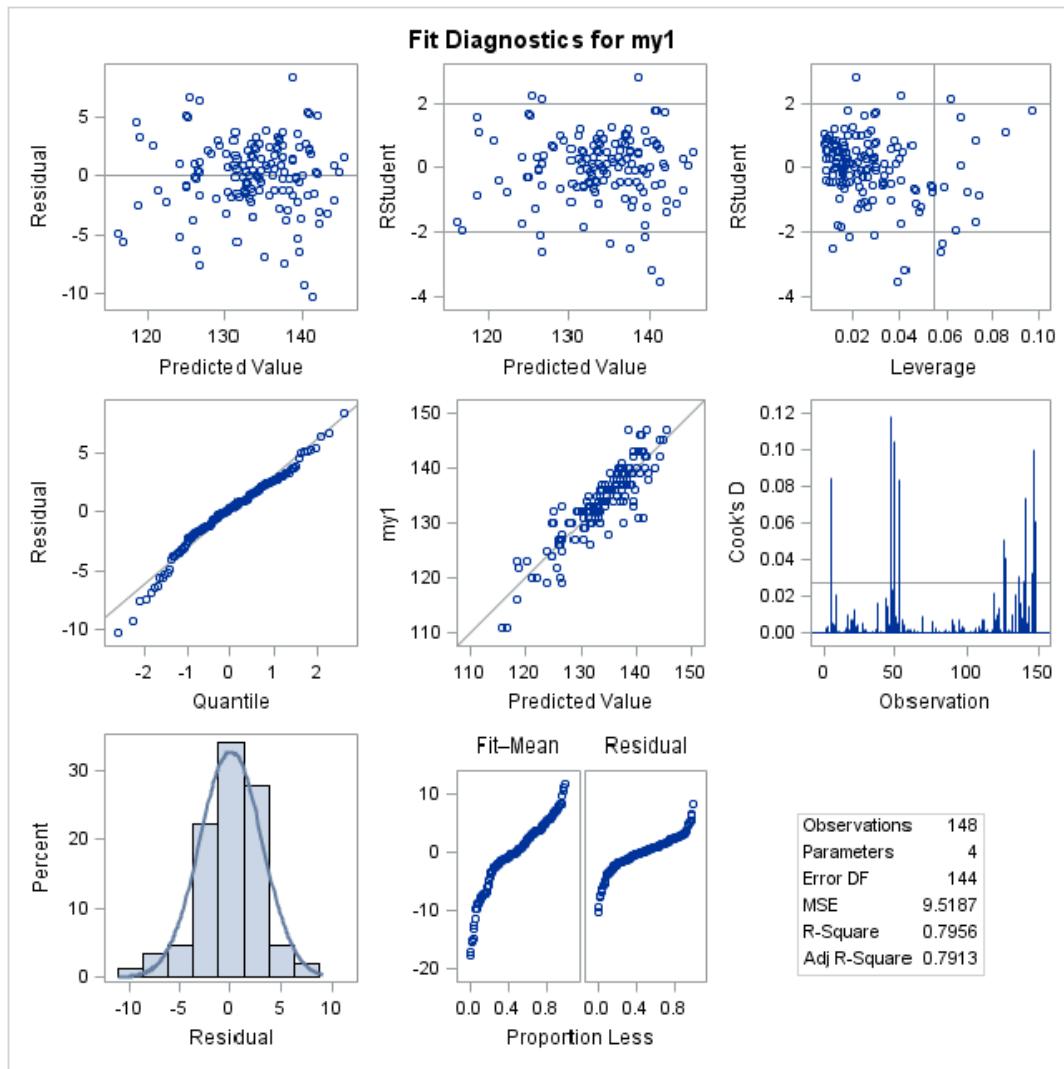
We have fitted a regression model of the form 'model my1=mr1 mr2 mr3' and obtain the parameter estimates given in the table below

Variable	DF	Parameter Estimate	Standard Error	t-Value	Pr > t
Intercept	1	16.35126	7.24638	2.26	0.0255
mr1	1	0.48318	0.11196	4.32	<.0001
mr2	1	0.74819	0.34939	2.14	0.0339
mr3	1	-0.16998	0.16439	-1.03	0.3029

The most significant predictor is - not surprisingly - mr1, i.e. the March measurement

with the same wavelength.

In order to validate the model further, one should investigate the residuals from the model. These analyses are summarized in the Fit-Diagnostic plot shown in the figure below.



We shall briefly discuss the different subplots.

If we take a look at the figure we see nothing remarkable in the first plot in the panel. The residuals seem to be randomly distributed around a horizontal line.

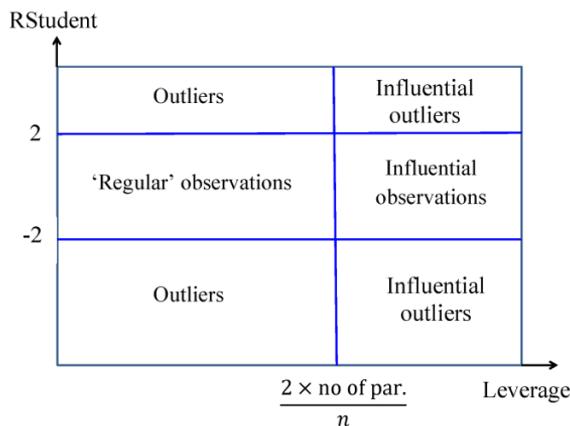
In the next plot however, we see that around 10 observations fall outside the range $[-2, 2]$ for the RStudent, and another 10 are very close to the limits. This indicates that the distribution of Rstudent has a heavier tail than a normal distribution.

From the last plot in the first row follows that three of the outliers are influential. A closer examination show that those are observations no 53, 141, and 146 (by speci-

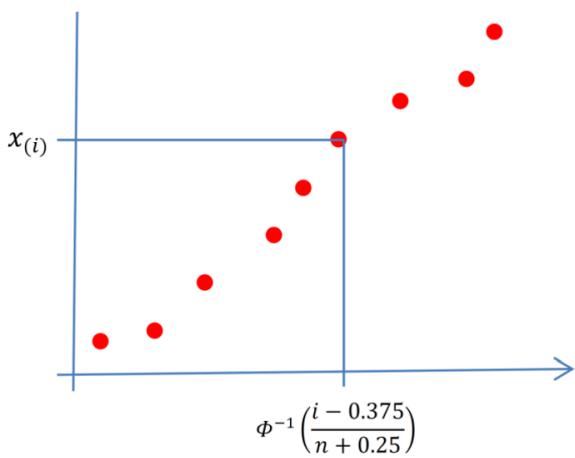
fying a label option in SAS, the observation numbers are added to the extreme observations in the plots), and those are also showing very large values of Cook's D, shown last in row two. The observed versus the predicted values does not show anything extraordinary.

The Q-Q plot show a systematic deviation from a straight line corresponding to having heavier tails than a normal distribution. This is also seen from the comparison of the histogram and the fitted normal curve. The fit-mean plot show that the range for the residuals is considerably smaller than the range of the centered fit, indicating that despite the deficiencies described earlier, the estimated model reduces the variation considerably. This is in good accordance with the R-square value 07913, i.e. almost 80% of the variation in the dependent variable is explained by the three explanatory variables.

A more detailed explanation of 3 of the subplots are given below



The RStudent by leverage plot show which observations are influential outliers. These observations will have a disproportional influence on the parameter estimates and they should be inspected for possible irregularities.



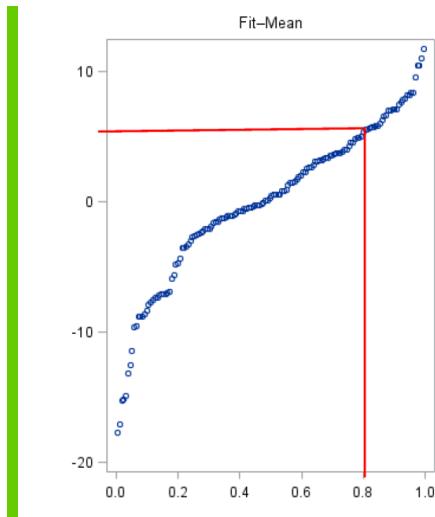
For the order statistics

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

the normal distribution Q-Q plot consists of the points

$$\left(\Phi^{-1}\left(\frac{i-0.375}{n+0.25}\right), x_{(i)}\right), \quad i = 1, \dots, n$$

If the points app. fall on a straight line, the distribution is well described by a normal distribution



The Fit – Mean plot (read Fit minus Mean) show the fraction of (centered) observations smaller than the abscissa. In the figure, 80% are smaller than 5. When comparing the plot for the fit with the corresponding plot for the residuals, a good fit corresponds to that the residuals show a much narrower range.

3.3 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.

3.3.1 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 $\xi_i, i = 0, 1, \dots, k$ are known polynomials of i 'th degree in t . We assume that

$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

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 3.18

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 3.19

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

$$\boldsymbol{\xi}^T \boldsymbol{\xi} = \begin{bmatrix} n\xi_0^2 & 0 & \cdots & 0 \\ 0 & \sum \xi_{1j}^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sum \xi_{kj}^2 \end{bmatrix},$$

i.e. $\boldsymbol{\xi}^T \boldsymbol{\xi}$ is a diagonal matrix. We therefore find

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{\xi}^T \boldsymbol{\xi})^{-1} \boldsymbol{\xi}^T \mathbf{Y} = \begin{bmatrix} \bar{Y}/\xi_0 \\ \sum \xi_{1j} Y_j / \sum \xi_{1j}^2 \\ \vdots \\ \sum \xi_{kj} Y_j / \sum \xi_{kj}^2 \end{bmatrix}$$

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

$$\begin{aligned} SS_{\text{res}} &= \|\mathbf{Y} - \xi \hat{\beta}\|^2 \\ &= (\mathbf{Y} - \xi \hat{\beta})^T (\mathbf{Y} - \xi \hat{\beta}) \\ &= \mathbf{Y}^T \mathbf{Y} - \hat{\beta}^T \xi^T \xi \hat{\beta} \\ &= \sum Y_j^2 - \{\hat{\alpha}^2 n \xi_0^2 + \hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2\} \\ &= \sum (Y_j - \bar{Y})^2 - \{\hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2\}. \end{aligned}$$

From this we immediately have

||| Theorem 3.20

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 3.21

The sums of squares which have been stated in the previous theorem are stochastically independent with expected values

$$\begin{aligned} E(\text{SS}_{i.\text{deg}}) &= E(\hat{\beta}_i^2 \sum_j \xi_i(t_j)^2) \\ &= \sigma^2 + \beta_i^2 \sum_j \xi_i(t_j)^2, \quad i = 1, \dots, k. \end{aligned}$$

and

$$E(\text{SS}_{\text{res}}) = E\left[\sum_j (\gamma_j - \bar{Y} - \dots - \hat{\beta}_k \xi_k(t_j))^2\right] = (n - k - 1)\sigma^2.$$

Finally

$$\frac{1}{\sigma^2} \text{SS}_{\text{res}} \sim \chi^2(n - k - 1),$$

and if $\beta_i = 0$

$$\frac{1}{\sigma^2} \text{SS}_{i.\text{deg}} \sim \chi^2(1).$$

||| 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	$\sigma^2 + \beta_3^2 \sum_j \xi_3(t_j)^2$
:	:	:	:
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 3.22**

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.

3.3.2 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. 219 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_1	ξ_2																			
1	-1	1	-3	1	-1	-2	2	-1	1	-5	5	-5	1	-1	-3	5	-1	3	-1	-7	7	-7	-7		
2	0	-2	-1	-1	3	-1	-1	2	-4	-3	-1	7	-3	5	-2	0	1	-7	4	-5	1	5	-13	23	
3	1	1	1	-1	-3	0	-2	0	6	-1	-4	4	3	-10	-1	-3	1	1	-5	-3	-3	7	-3	-17	
4		3	1	1	1	-1	-2	-4	1	-4	-4	2	10	0	-4	0	6	0	-1	-5	3	9	-15		
5			2	2	1	1	3	-1	-7	-3	-5	1	-3	-1	1	5	1	-5	-3	9	9	15			
6							5	5	5	1	1	2	0	-1	-7	-4	3	-3	-7	-3	17				
7												3	5	1	3	1	5	1	-5	-13	-23				
8																	7	7	7	7	7				
D	2	6	20	4	20	10	14	10	70	84	180	28	252	28	84	6	154	84	168	168	264	616	2184		
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 3.3.2: Values of orthogonal polynomials.

We now give an illustrative

||| Example 3.23

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

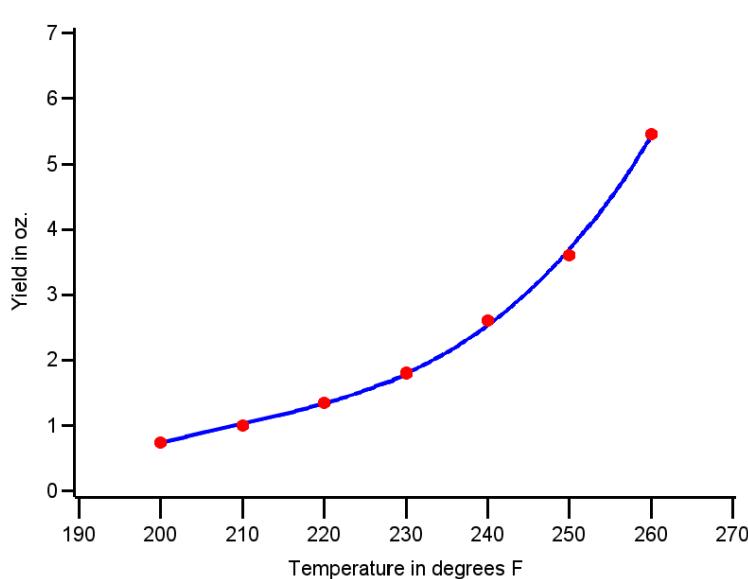
$$\begin{aligned}
 \hat{\mu}(t) &= 2.36 + 1 \cdot \hat{\beta}_1 \xi_1(t) + 1 \cdot \hat{\beta}_2 \xi_2(t) + \frac{1}{6} \hat{\beta}_3 \xi_3(t) \\
 &= 0.0236t^3 - 0.1409t^2 + 0.5631t + 0.2818.
 \end{aligned}$$

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 $\hat{\mu}(t)$.

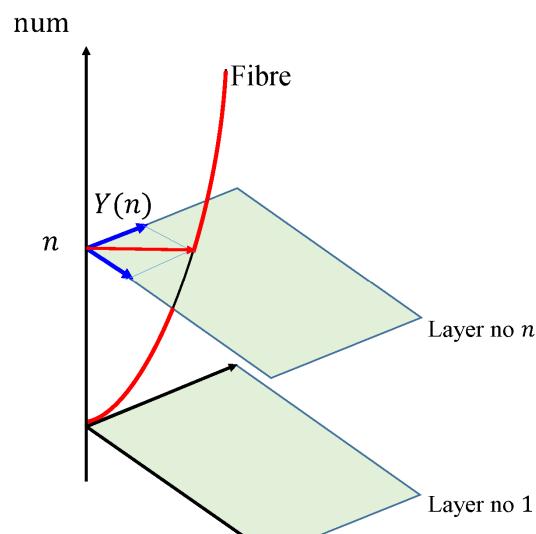
The estimated polynomial is shown together with the original data in the following figure.



Observed (red) and fitted (blue) yield as function of reaction temperature.

||| Example 3.24 Analysis of fibre directions

In Emerson et al. (2018) properties of fiber reinforced polymers are investigated. The fibres are uni-directional and could come from materials used in e.g. wind turbine blades. A sample is analyzed in a CT-scanner giving a 3 dimensional image of the sample consisting of e.g 127 layers. Each individual fibre has been identified, and we place each fibre in a coordinate system like the one shown in figure x. We describe the fibre geometry with variables $\{X(\text{num}), Y(\text{num}), \text{num}\}$ where 'num' is the layer number, and (X, Y) describe the deviation from vertical of the fibre.



The coordinate system used in describing the fibre geometry.

For some types of analysis it is of interest to have a polynomial approximation to the (X and) Y-variables. We shall show how orthogonal polynomials may be used for this. In our case, the independent variable, num, the layer number varies from 1 to 127. We use the formulas in 3.2.2 Determination of orthogonal polynomials to establish formulas for computing the 10 first orthogonal polynomials. This is quite straightforward, and the resulting formulas are given in the SAS code presented in the text box q0-q10 are the direct solution, and p0-p10 are scaled versions satisfying that the values of the polynomials vary between 1 and -1. The polynomials are shown in figure x..

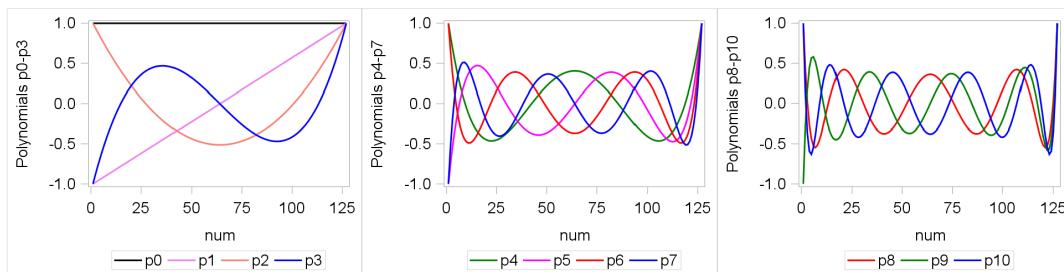
The data have been analyzed using [PROC REG](#). The ANOVA table is given below, and we see that there hardly is any variation left.

Source	DF	Sum of Squares	MeanSquare	F Value	Pr > F
Model	11	143397971	13036179	4.348E9	<.0001
Error	116	0.34780	0.00300		
Uncorrected Total	127	143397971			

The parameter estimates in the table below show that almost all parameters are significantly different from 0.

Variable	DF	ParameterEstimate	StandardError	t Value	Pr > t
p0	1	1062.59532	0.00486	218691	<.0001
p1	1	2.62400	0.00835	314.26	<.0001
p2	1	-5.60936	0.01061	-528.63	<.0001
p3	1	0.73066	0.01226	59.59	<.0001
p4	1	-0.44992	0.01347	-33.39	<.0001
p5	1	0.22815	0.01432	15.93	<.0001
p6	1	-0.41701	0.01485	-28.09	<.0001
p7	1	0.07652	0.01509	5.07	<.0001
p8	1	0.02539	0.01509	1.68	0.0951
p9	1	-0.08785	0.01486	-5.91	<.0001
p10	1	0.13080	0.01443	9.06	<.0001

The first 10 orthogonal polynomials for n=127 values of the independent variable is shown below

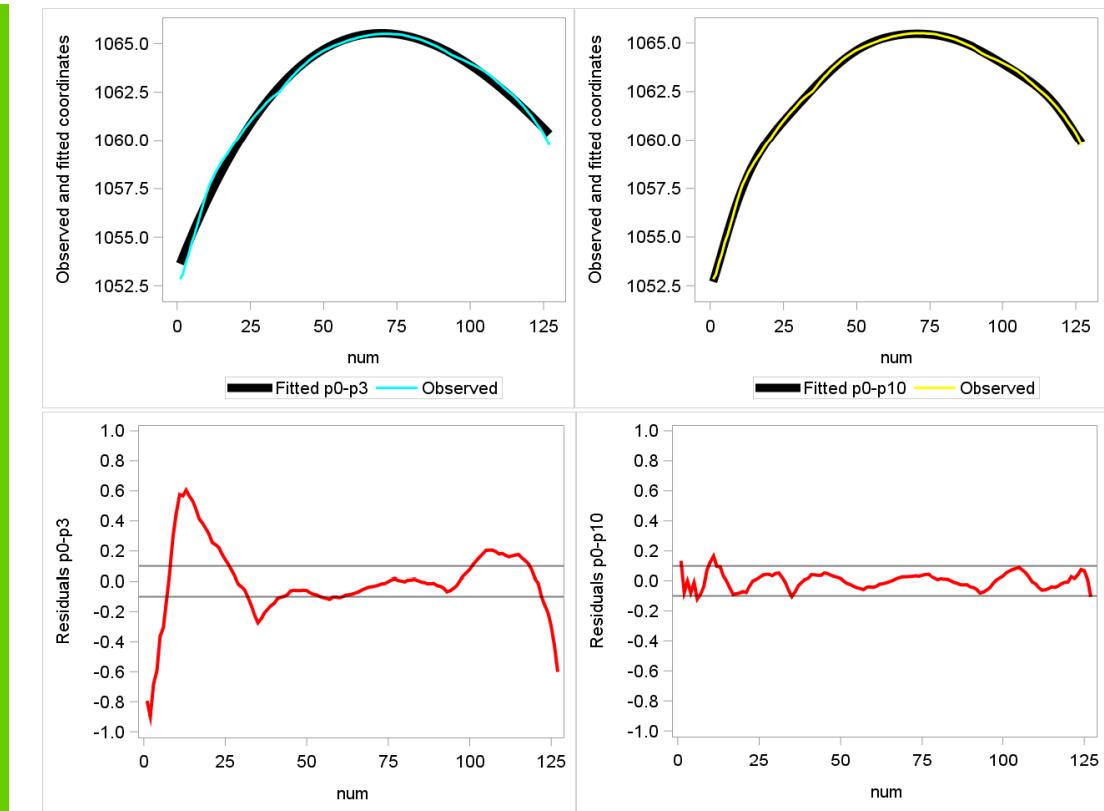


SAS code for generating orthogonal polynomials.

```
data fibrels;set fibrels;
q0=1;
q1=(num-64);
q2=(q1*q1-1*(16129-1)/12);
q3=(q1*q2-4*(16129-4)*q1/60);
q4=(q1*q3-9*(16129-9)*q2/140);
q5=(q1*q4-16*(16129-16)*q3/252);
q6=(q1*q5-25*(16129-25)*q4/396);
q7=(q1*q6-36*(16129-36)*q5/572);
q8=(q1*q7-49*(16129-49)*q6/780);
q9=(q1*q8-64*(16129-64)*q7/1020);
q10=(q1*q9-81*(16129-81)*q8/1292);
p0=q0;
p1=q1/63;
p2=q2/2625;
p3=q3/97650;
p4=q4/3431700;
p5=q5/116296500;
p6=q6/3837784500;
p7=q7/123989960769;
p8=q8/3934614800000;
p9=q9/122898850000000;
p10=q10/3783990900000000;
drop q0-q10;
run;
```

We have compared the fits obtained by using p0-p3 with the fits obtained by using p0-p10. Not surprisingly, the fit using up to 10th degree is better than the one using up to third degree. But the figure also shows that orthogonal polynomials, even of a high degree, are well suited for this type of approximation.

Below we see fitted polynomials and observed values corresponding to using the first 3 and the first 10 polynomials. In the bottom row we see the residuals.



3.4 Selection 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.

3.4.1 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 3.25

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

$$E(Y|x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

or in matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix},$$

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}).$$

In the numerical appendix (p. 234) 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 e.g.

$$M \supseteq H_{12} \supseteq H_2 \supseteq H_0.$$

The above mentioned sequence of sub-spaces corresponds to successive testing of the hypothesis

$$\beta_3 = 0, \quad \beta_1 = 0, \quad \beta_2 = 0.$$

This is only one of the possible sequences. 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. 34 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.

3.4.2 Examination of all regressions.

This method can of course only be used if there are reasonably few variables. We summarise the result from the numerical appendix (section 3.4.6) 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).

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 ($E(y|x) = \alpha + \beta_3 x_3$)

$$\frac{\|p_{H_0}(\mathbf{y}) - p_{H_3}(\mathbf{y})\|^2/1}{\|\mathbf{y} - p_{H_3}\|^2/3} = \frac{21.868 - 6.38}{6.38/3} \simeq 7.28.$$

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 :

$$E(Y|x) = \alpha + \beta_3 x_3 \simeq 0.4 + 2.2x_3.$$

Here \simeq means estimated at). The estimate of the error (the variance) on the measurements is (estimated with 3 degrees of freedom):

$$s^2 = 2.13.$$

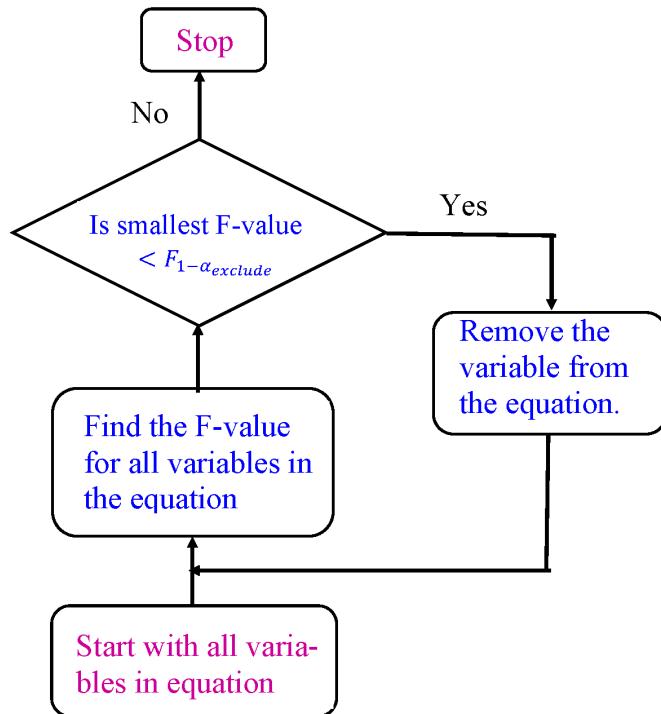


Figure 3.4 – Flow diagram for Backwards-elimination procedure in stepwise regression analysis.

3.4.3 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 : $E(y) = \alpha + \beta_3 x_3$ when we use a significance level α , greater than 8%.

Step	F-value for test of $\beta_i = 0$	/ Quantile in F-distribution
Model : $E(Y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$		
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 : $E(Y) = \alpha + \beta_2x_2 + \beta_3x_3$		
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 : $E(Y) = \alpha + \beta_3x_3$		
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.

3.4.4 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. 238) we see that x_3 has the greatest correlation coefficient with y , viz. 0.8416. We test if β_3 in the model $E(Y) = \alpha + \beta_3x_3$ can

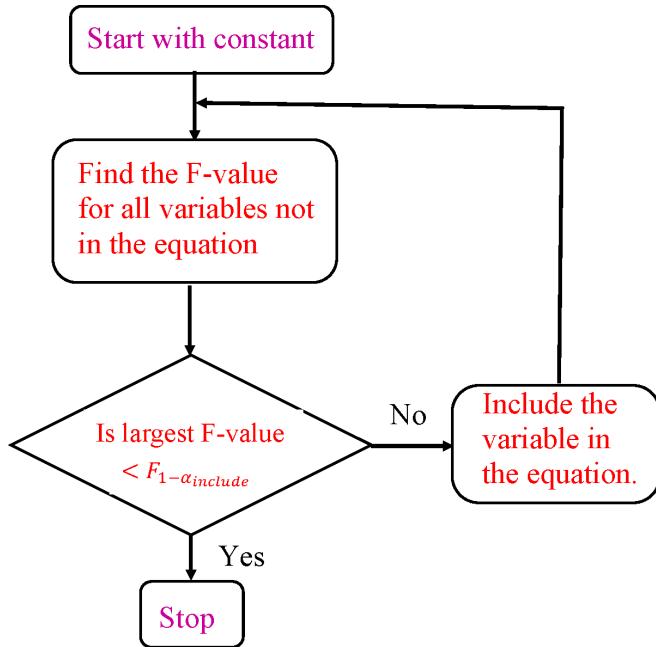


Figure 3.5 – Flow diagram for Forward-selection procedure in stepwise regression analysis

be assumed to be 0 we have the test statistic (see p. 238).

$$\frac{15.488/1}{6.38/3} = 7.28 \simeq F(1, 3)_{0.92}.$$

If we use a significance level $\alpha = 10\%$ we continue (since we then reject $\beta_3 = 0$).

2) From the partial correlation matrix given x_3 (p. 239) 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. 238)

$$\frac{2.095/1}{4.2855/2} = 0.98 \simeq F(1, 2)_{0.55}.$$

Since we were using significance level $\alpha = 10\%$, 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 3.26**

If we had used a significance level $\alpha = 50\%$ we would have continued the analysis and considered the partial correlations given x_2 and x_3 . According to the matrix p. 239 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 $E(y) = \alpha + \beta_1x_1 + \beta_2x_2 + \beta_3x_3$ is significantly different from 0. The test statistic is (p. 238)

$$\frac{3.652/1}{0.634/1} = 5.76 \simeq F(1, 1)_{0.71}.$$

In the case we have seen that the equation was extended considerably just by changing α . It is important to note that changes in the significance α can have drastic consequences for the resulting model.

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.

3.4.5 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. 238 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 3.27

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

$$0.98 < F(1, 2)_{0.55} > F_{1-\alpha_{ind}}.$$

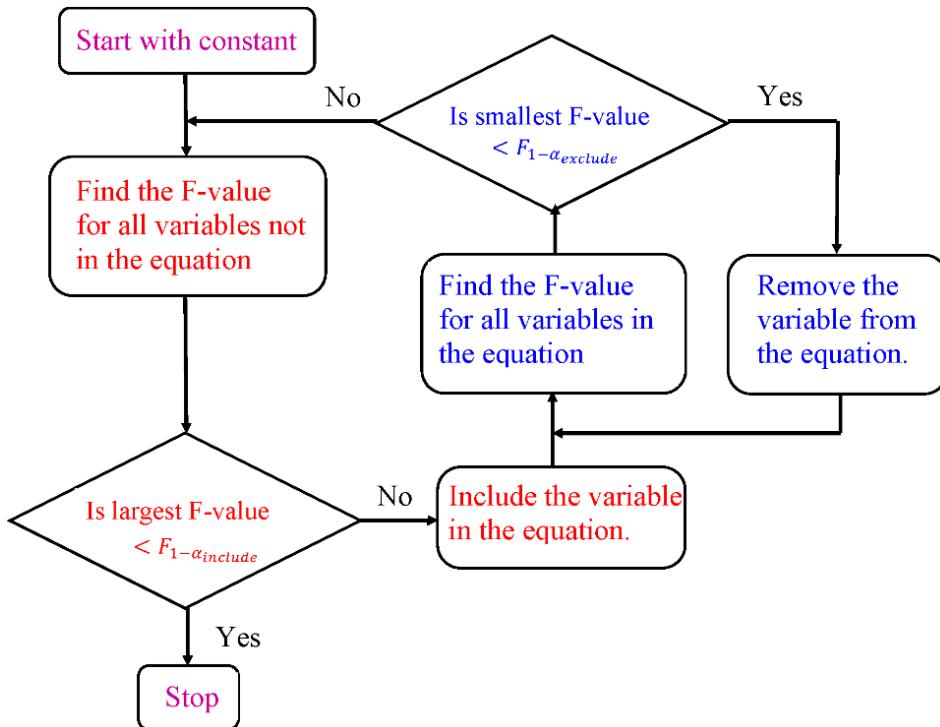


Figure 3.6 – Flow diagram for Stepwise-Regression procedure in stepwise regression analysis.

3.4.6 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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

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

$$\begin{pmatrix} 1 & 0.4199 \\ 0.4199 & 1 \end{pmatrix} - \begin{pmatrix} 0.4263 \\ 0.5728 \end{pmatrix} [1]^{-1} [0.4263 - 0.5728] = \begin{pmatrix} 0.8183 & 0.6641 \\ 0.6641 & 0.6718 \end{pmatrix},$$

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.

3.5 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.

3.5.1 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 -

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y}.$$

Furthermore we assume that the independent variables are scaled so that $\mathbf{x}^T \mathbf{x}$ 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 $\mathbf{x}^T \mathbf{x}$ 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 $\hat{\beta}$ that has not been dealt with earlier. We have

$$D(\hat{\beta}) = \sigma^2(\mathbf{x}^T \mathbf{x})^{-1}.$$

If we put L equal to the distance from $\hat{\beta}$ to β ,

$$L^2 = (\hat{\beta} - \beta)^T (\hat{\beta} - \beta),$$

we get

$$E(L^2) = \sum_{i=1}^p E[(\hat{\beta}_i - \beta_i)^2] = \sum_{i=1}^p V(\hat{\beta}_i) = \sigma^2 \text{tr}(\mathbf{x}^T \mathbf{x})^{-1}.$$

Since

$$L^2 = \hat{\beta}^T \hat{\beta} - 2\hat{\beta}^T \beta + \beta^T \beta,$$

we get that the expected value of the squared length of $\hat{\beta}$ is

$$E(\hat{\beta}^T \hat{\beta}) = \sigma^2 \text{tr}(\mathbf{x}^T \mathbf{x})^{-1} + \beta^T \beta.$$

If we denote the eigenvalues for $\mathbf{x}^T \mathbf{x}$

$$\lambda_1 \geq \dots \geq \lambda_p,$$

we obtain (accordingly to Theorem A.27 and the results on p. 472)

$$E(L^2) = \sigma^2 \left(\frac{1}{\lambda_1} + \dots + \frac{1}{\lambda_p} \right) > \frac{\sigma^2}{\lambda_p},$$

and

$$E(\hat{\beta}^T \hat{\beta}) = \sigma^2 \left(\frac{1}{\lambda_1} + \dots + \frac{1}{\lambda_p} \right) + \beta^T \beta > \frac{\sigma^2}{\lambda_p} + \beta^T \beta.$$

If the independent variables are strongly correlated the eigenvalues of $\mathbf{x}^T \mathbf{x}$ 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 3.7.

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

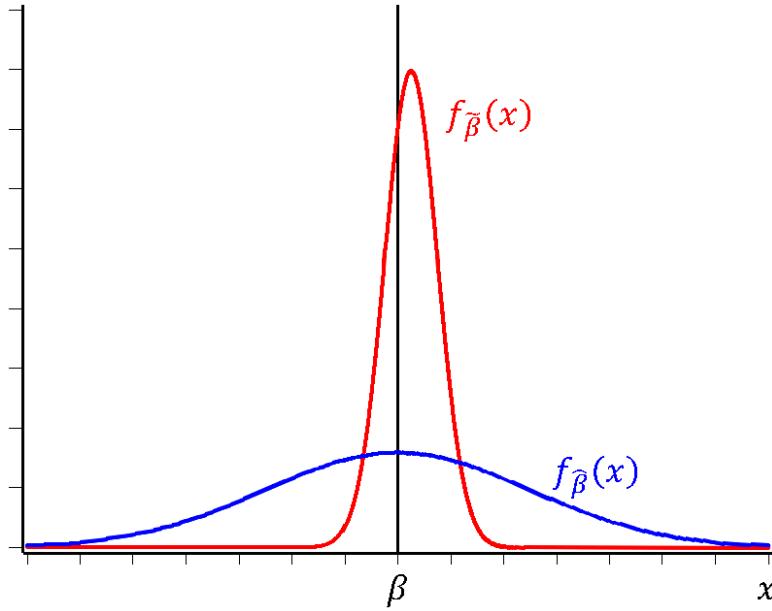


Figure 3.7 – Bias-variance trade-off for an estimator. The estimator $\hat{\beta}$ is unbiased, i.e. $E(\hat{\beta}) = \beta$, whereas $\tilde{\beta}$ is biased $E(\tilde{\beta}) \neq \beta$, but $V(\hat{\beta}) > V(\tilde{\beta})$ and we see that the outcomes of $\tilde{\beta}$ are distributed much narrower around the true value than the outcomes of $\hat{\beta}$ are.

variance, this would obviously be preferable. This is exactly what is obtained with the ridge estimator introduced in the definition below.

We continue this section by investigating some of the consequences of the fact that the expected squared length of an OLS estimator may be much larger than the squared length of the parameter. We consider an arbitrary estimator $t = \hat{\theta}$ in a general linear model. It has the residual sum of squares

$$\begin{aligned}\phi(t) &= \|y - xt\|_2^2 = \|y - xt\|^2 = (y - xt)^T(y - xt) \\ &= (y - xt)^T(y - xt) \\ &= (y - x\hat{\theta})^T(y - x\hat{\theta}) + (t - \hat{\theta})^T x^T x (t - \hat{\theta}) \\ &= \min_t \phi(t) + (t - \hat{\theta})^T x^T x (t - \hat{\theta})\end{aligned}$$

We now look for an estimator $t = \tilde{\theta}$ with minimum length and residual sum of square equal to $\min_t \phi(t) + c = \phi(\hat{\theta}) + c$ for some constant c , i.e. we seek a t that minimizes

$$t^T t \text{ subject to } (t - \hat{\theta})^T x^T x (t - \hat{\theta}) = c$$

We use a Lagrange multiplier λ and get the function

$$F(\mathbf{t}, \lambda) = \mathbf{t}^T \mathbf{t} + \lambda \{ (\mathbf{t} - \widehat{\boldsymbol{\theta}})^T \mathbf{x}^T \mathbf{x} (\mathbf{t} - \widehat{\boldsymbol{\theta}}) - c \}.$$

Furthermore

$$\frac{\partial F}{\partial \mathbf{t}} = 2\mathbf{t} + 2\lambda \{ \mathbf{x}^T \mathbf{x} \mathbf{t} - \mathbf{x}^T \mathbf{x} \widehat{\boldsymbol{\theta}} \} = 2\mathbf{t} + 2\lambda \{ \mathbf{x}^T \mathbf{x} \mathbf{t} - \mathbf{x}^T \mathbf{Y} \} = \mathbf{0}$$

or

$$\mathbf{t} + \lambda \mathbf{x}^T \mathbf{x} \mathbf{t} = \lambda \mathbf{x}^T \mathbf{Y} \Leftrightarrow \left\{ \mathbf{x}^T \mathbf{x} + \frac{1}{\lambda} \mathbf{I} \right\} \mathbf{t} = \mathbf{x}^T \mathbf{Y}$$

For $k = 1/\lambda$ we thus get the estimator

$$\widehat{\boldsymbol{\theta}}_{(k)} = \{ \mathbf{x}^T \mathbf{x} + k \mathbf{I} \}^{-1} \mathbf{x}^T \mathbf{Y}$$

In numerical mathematics this is known as the Tikonov regularized solution to the least squares problem. Later we shall call this a Ridge estimator corresponding to the Ridge parameter k .

||| Example 3.28

We illustrate the above results in an example using simulated data from a general linear model. We assume that each case consists of 25 observations drawn from the same model, and we simulate the outcomes of 20 cases. We thus have data

$$\begin{bmatrix} Y_{1,1} & \cdots & Y_{1,20} \\ \vdots & & \vdots \\ Y_{25,1} & \cdots & Y_{25,20} \end{bmatrix} = \begin{bmatrix} 5.87 & 5.81 & 5.73 \\ \vdots & \vdots & \vdots \\ 5.99 & 4.80 & 5.44 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} & \cdots & \varepsilon_{1,20} \\ \vdots & & \vdots \\ \varepsilon_{25,1} & \cdots & \varepsilon_{25,20} \end{bmatrix}$$

Part of the data are presented in the table below

Obs no.	Case no.								The design matrix		
	1	2	3	...	18	19	20	x			
1	47.53	46.27	47.39		44.51	48.97	46.3	5.87	5.81	5.73	
2	23.77	28.47	24.83		25.33	31.19	30.22	3.94	3.25	3.65	
3	40.38	35.01	36.10		33.07	37.39	33.47	3.54	5.26	4.33	
4	40.90	44.08	42.99		39.62	40.28	40.29	5.42	4.94	5.00	
5	40.87	41.33	42.75		39.62	40.08	39.96	4.93	5.45	5.13	
:								:			
20	38.51	37.75	38.70		40.41	37.16	39.31	4.47	4.77	4.77	
21	44.44	47.08	41.79		43.29	42.55	46.86	5.21	5.52	5.26	
22	38.89	42.02	44.04		40.90	37.70	40.27	4.64	5.69	5.13	
23	54.64	57.31	51.32		49.34	52.40	53.93	6.66	6.87	6.70	
24	48.99	47.01	50.86		53.11	50.08	48.65	4.52	7.57	5.98	
25	43.38	42.34	44.61		43.44	39.42	40.22	5.99	4.80	5.44	

Simulating outcomes from 20 cases of 25 observations from a GLM

The eigenvalues of $\mathbf{x}^T \mathbf{x}$ are 2113.8071, 15.954549, and 0.1552099 giving a condition number $\lambda_1/\lambda_3 = 13619.0$, showing a high degree of multicollinearity between the covariates. We have therefore computed the 20 different realizations of the Ridge, i.e. the Tikonov regularized estimators $\tilde{\theta}_{(k)}$ for $k = 0.01, 0.05, 0.1, 0.2, 0.3, 0.5$. In the table below, we show the outcomes of as well $\hat{\theta}$ as $\tilde{\theta}_{(k)}$ for $k = 0.5$. Knowing the true value $\theta = [2 \ 4 \ 2]^T$ with squared length 24, enables us to compute the squared deviation between the true and the estimated parameter. Furthermore we have shown the residual sum of squares. Firstly we see that the squared length of the Ridge estimates are smaller, in some cases much smaller than the squared length of the OLS estimates. We note that the Ridge estimates in all cases are closer to the true value than the OLS estimates.

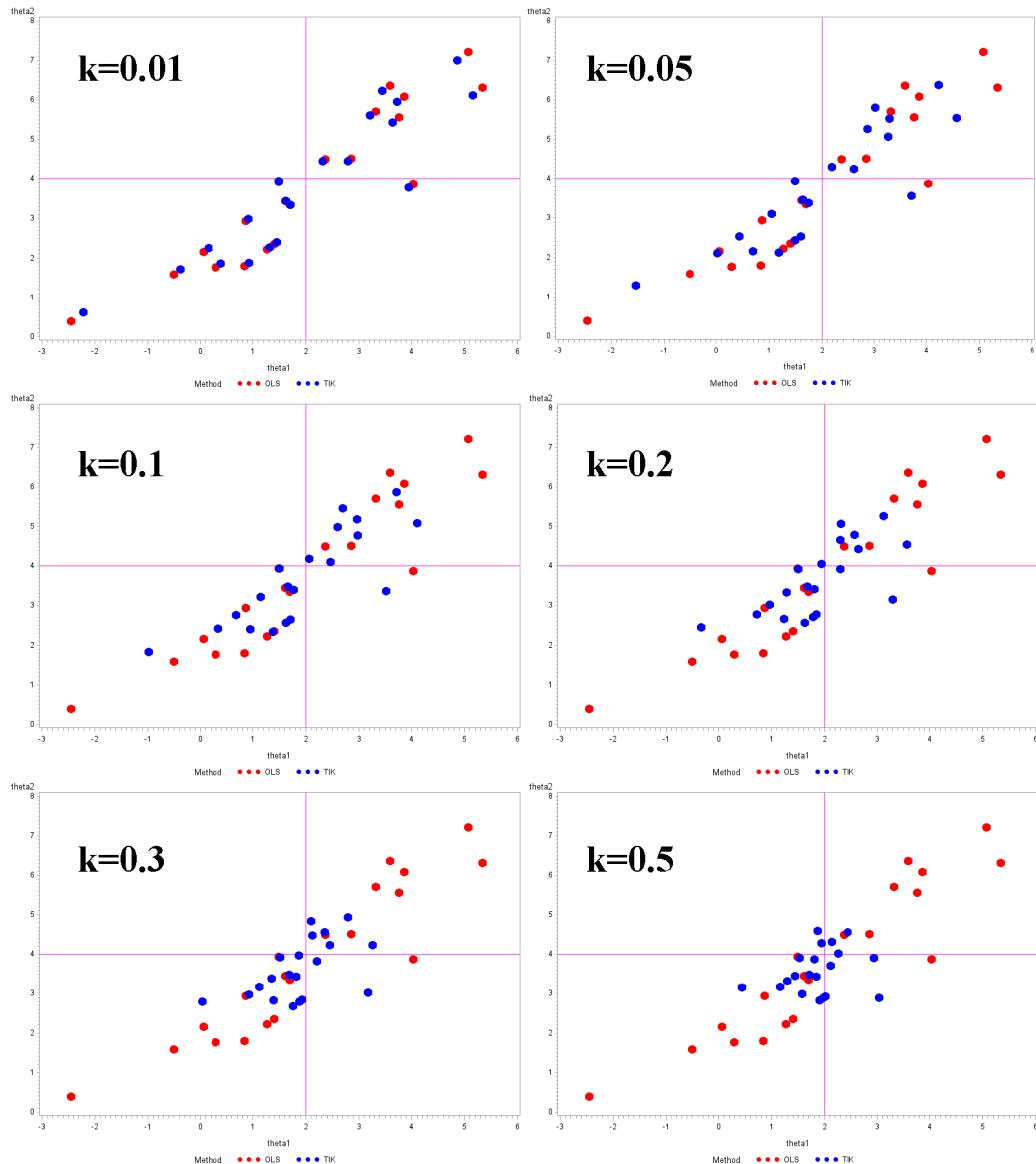
Case	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\ \hat{\theta}\ ^2$	$\ \theta - \hat{\theta}\ ^2$	$\ \mathbf{Y} - \hat{\mathbf{Y}}\ ^2$	$\tilde{\theta}_{(k)1}$	$\tilde{\theta}_{(k)2}$	$\tilde{\theta}_{(k)3}$	$\ \tilde{\theta}_{(k)}\ ^2$	$\ \theta - \tilde{\theta}_{(k)}\ ^2$	$\ \mathbf{Y} - \tilde{\mathbf{Y}}_{(k)}\ ^2$
1	2.37	4.50	1.09	27.02	1.21	98.35	1.80	3.87	2.28	23.44	0.13	98.72
2	0.84	1.80	5.41	33.25	17.82	102.33	1.90	2.82	3.33	22.65	3.16	103.35
3	3.76	5.56	-1.31	46.69	16.44	87.22	2.26	4.02	1.72	24.26	0.15	89.38
4	1.61	3.45	2.79	22.24	1.07	33.85	1.71	3.48	2.65	22.08	0.78	33.88
5	1.70	3.35	2.98	22.99	1.49	57.83	1.85	3.44	2.75	22.76	0.90	57.87
6	0.29	1.76	5.96	38.68	23.60	94.86	1.58	2.98	3.44	23.25	3.30	96.35
7	3.59	6.37	-1.90	57.03	23.34	90.68	1.87	4.59	1.59	27.10	0.53	93.58
8	1.49	3.94	2.71	25.08	0.77	71.65	1.52	3.90	2.71	24.88	0.73	71.70
9	3.32	5.71	-1.02	44.75	13.82	123.12	1.95	4.28	1.79	25.31	0.12	125.01
10	2.86	4.51	0.64	28.88	2.83	63.52	2.11	3.71	2.18	22.99	0.13	64.10
11	-2.45	0.39	10.09	107.86	98.21	68.03	0.44	3.16	4.42	29.71	9.02	75.55
12	5.07	7.22	-4.24	95.79	58.76	117.46	2.45	4.56	1.04	27.85	1.44	123.99
13	0.05	2.15	5.83	38.58	21.83	103.12	1.29	3.32	3.42	24.35	2.97	104.51
14	5.34	6.31	-3.62	81.50	48.14	88.74	2.94	3.91	1.18	25.31	1.57	94.13
15	-0.51	1.59	6.94	50.94	36.49	85.26	1.16	3.18	3.68	24.96	4.20	87.77
16	4.03	3.87	0.09	31.20	7.77	92.50	3.04	2.89	2.05	21.85	2.31	93.40
17	3.86	6.09	-1.91	55.63	23.14	96.51	2.13	4.31	1.59	25.68	0.29	99.41
18	1.41	2.35	4.18	24.96	7.81	129.19	2.01	2.92	3.00	21.58	2.18	129.52
19	0.86	2.94	4.04	25.68	6.55	89.21	1.44	3.45	2.95	22.65	1.51	89.52
20	1.26	2.22	4.47	26.46	9.81	72.69	1.98	2.89	3.08	21.72	2.39	73.15

Ordinary least squares estimators and Ridge, i.e. Tikonov regularized estimators for the 20 cases. The regularization parameter is $k = 0.5$.

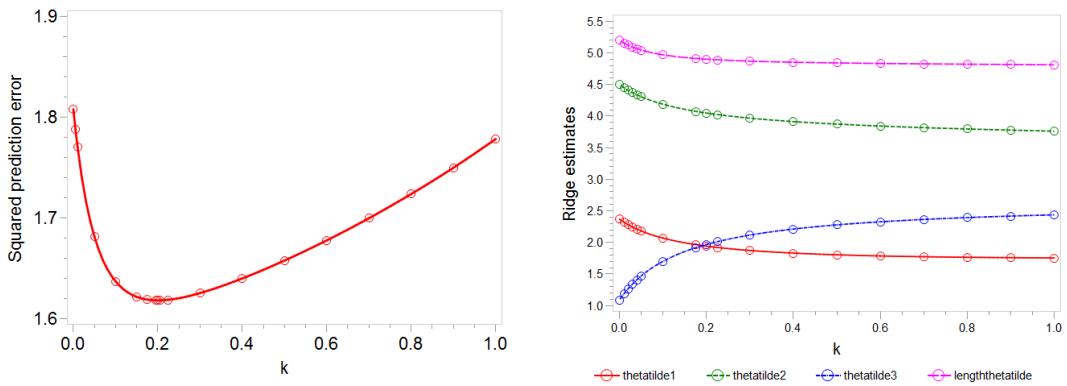
We have compared the two types of estimates in the figure below. We see how the Ridge estimates move in the direction of the true value $(2, 4)$ for increasing values of k .

On the other hand, the OLS estimates do give a smaller residual sum of squares (the definition of the least squares estimates), i.e.

$$\|\mathbf{Y} - \mathbf{x}\hat{\theta}\|^2 = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 \leq \|\mathbf{Y} - \widetilde{\mathbf{Y}}_{(k)}\|^2 = \|\mathbf{Y} - \mathbf{x}\widetilde{\theta}_{(k)}\|^2$$



Comparison between the OLS estimators and the Tikonov regularization (Ridge estimates) for different k -values. In red: OLS $\{\widehat{\theta}_1, \widehat{\theta}_2\}$, in blue: Tikonov $\{\widetilde{\theta}_{1(k)}, \widetilde{\theta}_{2(k)}\}$



The squared prediction error $\|x\theta - x\tilde{\theta}_{(k)}\|^2$ and the Ridge estimates as function of k . We see that the estimates more or less stabilize from 0.2 and up, the same value where we have the minimum squared prediction error.

||| Definition 3.29

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

$$(x^T x + k \cdot I) \hat{\beta}^* = x^T Y,$$

i.e.

$$\hat{\beta}^* = (x^T x + k \cdot I)^{-1} x^T Y.$$

Here k is a constant $\in [0, 1]$.

||| Remark 3.30

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

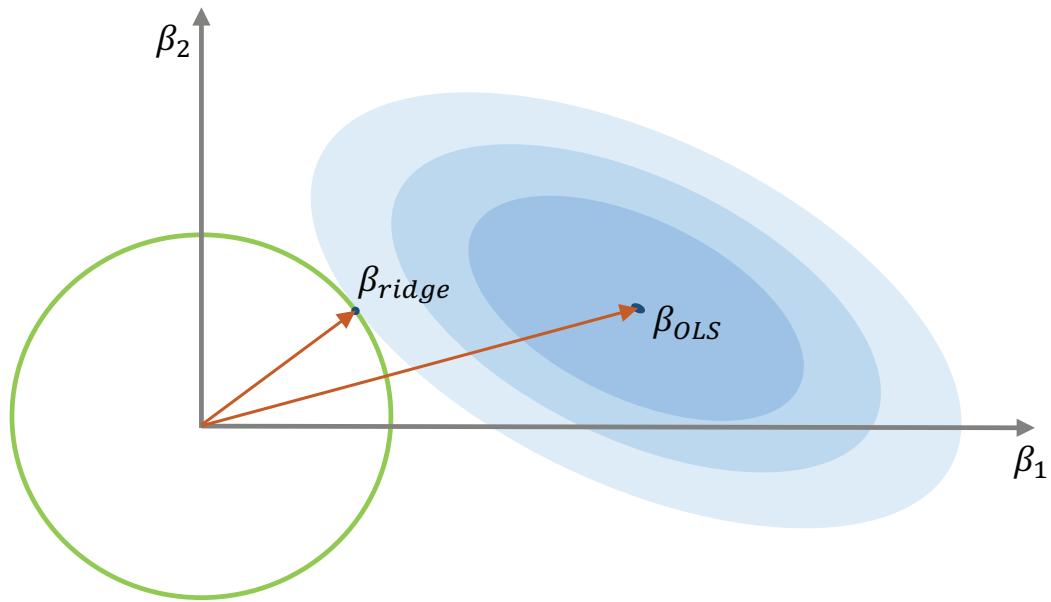


Figure 3.8 – Geometrical depiction of ridge regression

|||| Theorem 3.31

Let the situation be like in the definition. We put $\mathbf{x}^T \mathbf{y} = g$ and denote the observed residual sum of squares for an arbitrary estimator $\tilde{\beta}$ equal to

$$H(\tilde{\beta}) = (\mathbf{y} - \mathbf{x} \tilde{\beta})^T (\mathbf{y} - \mathbf{x} \tilde{\beta}).$$

Then the gradient of H in $\tilde{\beta} = \mathbf{0}$ 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 omitted

Not very complicated but omitted. The reader is referred to Hoerl and Kennard (1970) and Marquardt (1970).

Figure 3.8 depicts the situation geometrically in the case $p = 2$. The point β_{OLS}

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 large ellipse. We see that $\hat{\beta}_{ridge}$ is the shortest vector that has a residual sum of squares as small as the value of H on the large ellipse.

Other properties of the ridge estimator are given in the following theorem

||| Theorem 3.32

Let the situation be as described above. Then $\hat{\beta}_k^* = \hat{\beta}^*$ is a linear transformation of $\hat{\beta}$ since

$$\hat{\beta}^* = \mathbf{z}_k \hat{\beta} = (\mathbf{x}^T \mathbf{x} + k\mathbf{I})^{-1}(\mathbf{x}^T \mathbf{x}) \hat{\beta}.$$

$\hat{\beta}^*$ is biased since

$$E(\hat{\beta}^*) = \mathbf{z}_k \beta.$$

The dispersion matrix of $\hat{\beta}^*$ is

$$D(\hat{\beta}^*) = \sigma^2 [\mathbf{x}^T \mathbf{x} + k\mathbf{I}]^{-1} (\mathbf{x}^T \mathbf{x}) [\mathbf{x}^T \mathbf{x} + k\mathbf{I}]^{-1},$$

and the expected squared distance to β is

$$E[(\hat{\beta}^* - \beta)^T (\hat{\beta}^* - \beta)] = \text{tr}(D(\hat{\beta}^*)) + \beta^T (\mathbf{z}_k - \mathbf{I})^T (\mathbf{z}_k - \mathbf{I}) \beta.$$

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 3.33

If $\beta^T \beta$ 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 $\beta^T (\mathbf{z}_k - \mathbf{I})^T (\mathbf{z}_k - \mathbf{I}) \beta$ 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. [Wahba \(1990\)](#) for results on using *cross validation*), but the ridge trace is a straight forward method.

||| Definition 3.34

By the *ridge trace* we understand the mapping of the individual coefficients in the ridge estimator as a function of k .

||| Remark 3.35

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 3.36

The data are observations from the Landsat Thematic Mapper satellite taken over the same area in India at two time points, one in March and the other in May. Each observation consists of values of reflected light from six spectral bands shown in the table below.

Variable	Spectral band (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

The pixel size is $30 \text{ m} \times 30 \text{ m}$. False color composites of the satellite images are shown in the introduction to section 6.2 Canonical variables and correlations. 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 each of 600 pixel have measurements

$$\{mr_1, mr_2, mr_3, mr_4, mr_5, mr_6, my_1, my_2, my_3, my_4, my_5, my_6\}$$

To these we add sums, squares, and cross products like

$$sum_{ij} = mr_i + mr_j, \quad sqmr_i = mr_i^2, \quad cpmr_{ij} = mr_i \times mr_j$$

We now want to predict the outcome for the longest wavelength in May, i.e. my_6 from the other measurements in May and in March, also using some of the generated variables like sums and products. We split the 600 observations randomly into two datasets, Indiatest with 521 observations and Indiatrain with 79 observations. In order to get quantities that are easily compared, we standardize both datasets so that all independent variables (covariates) are centered and have standard deviation 1, and the dependent variable is centered but keeps its variability. A SAS-program doing this is given in the textbox below.

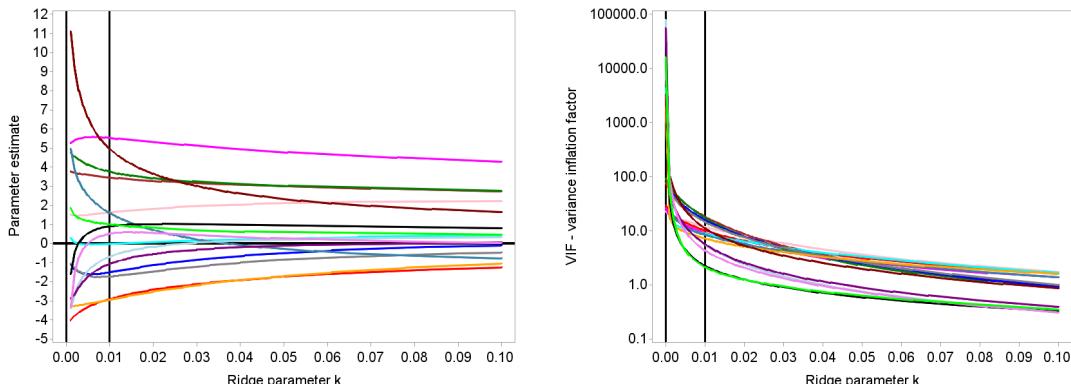
```
%let covars = my1 - my5 mr1 - mr6 sqmr1 - sqmr6 cpmr12 cpmr13 cpmr14 cpmr15
cpmr16 cpmr23 cpmr24 cpmr25 cpmr26 cpmr34 cpmr35 cpmr36 cpmr45 cpmr46
cpmr56 ;
%let depvar= my6;
proc standard data = indiatest out = indiatest_std mean = 0 std = 1; var &covars;
run;
proc standard data = indiatest_std out = indiatest_std mean = 0; var &depvar; run;
```

The covariates are highly collinear which is seen from the Variance Inflation Factors in the table at the end of this example. We recall the definition of the VIF for covariate no j as $1/(1 - R_j^2)$, where R_j^2 is the squared multiple correlation between covariate no j and all other covariates. As a rule of thumb, VIFs above 5-10 should be considered indicative of multicollinearity problems. In the present case the largest VIF is close to 80 000! We have therefore run a ridge regression analysis using the SAS program below.

The `outvif` option in the `PROC REG` statement adds the variance inflation factors (suitably modified to take the changed error structure for ridge estimates into account) to the output dataset beta.

```
Title 'Beta, Ridgeparm = 0 to 0.1 by 0.001';
proc reg data = indiatrain_std plots = ridge(unpack) outvif outset = beta
ridge = 0 to 0.1 by 0.001 ;
my6hat: model my6 = my1 - my5 mr1 - mr6 cpmr16 cpmr26 cpmr36 cpmr46 cpmr56;
run;
```

In order to determine the ridge constant k , we look at the ridge trace, the mapping of parameter estimates as function of k , and the corresponding plot of the variance inflation factors.



The ridge trace and the variance inflation plot. In the ridge trace we have not mapped the values for $k = 0$ since the OLS parameters deviate with up to (almost) two orders of magnitude from the ridge estimates. The first values mapped correspond to $k = 0.001$. The ordinate axis for the VIF plot is given in logarithmic scale. The major part of the variation in both plots take place for $k \leq 0.01$, the value chosen as the final ridge constant.

Based on these plots, $k = 0.01$ was chosen as the ridge constant. Judged from the graphs, this value is probably somewhat too small. Irrespective of this, we see that $\hat{\beta}_{Ridge}$ performs better than $\hat{\beta}_{OLS}$. Firstly, the parameter estimates are in general (much) smaller for the ridge case, and secondly the ridge estimators have much smaller VIFs. More important are the performance with respect to the values of the Average Squared Error. This is defined as

$$\text{ASE} = \frac{1}{n} \sum_{i=1}^n (Y_i - Y_i^{(pred)})^2$$

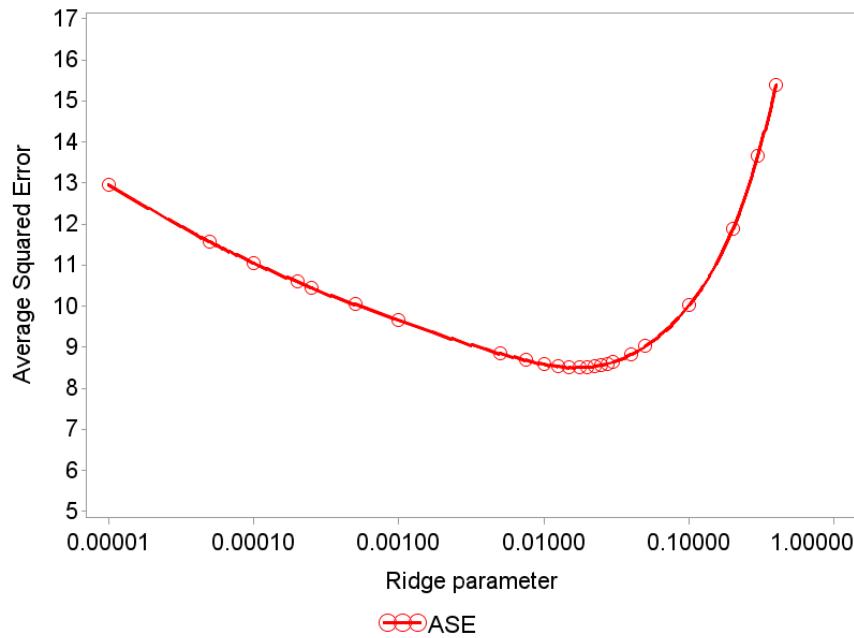
where $Y_i^{(pred)}$ is either $x_i \tilde{\beta}_{Ridge}$ or $x_i \hat{\beta}_{OLS}$ and x_i is the i 'th row in the design matrix (the covariates corresponding to observation no i). The ASE has the number of terms n as divisor instead of the degrees of freedom as in the Mean Squared Error (MSE). The latter concept is closely linked to the notion of unbiasedness, which we have skipped in using ridge estimators. From the table below it follows that the ASE_{Ridge} computed for the observations in the test dataset (8.29) is somewhat smaller than ASE_{OLS} computed for the test dataset (14.38), i.e. the ridge estimators give a better prediction on the test dataset, that has not been used in the estimation. That the opposite is the case for the training dataset is a trivial consequence of that the OLS estimators are chosen to minimize the sum of squared prediction errors.

The ASE may also be used in establishing an estimate of the ridge constant. A common strategy in cases like the present is dividing the data into three datasets: A training, an evaluation, and a test dataset. One might then estimate e.g. ridge parameters from the training data, find the k-value minimizing the ASE based on the evaluation data, and do a final testing of the model properties on the test data.

Below we have shown corresponding values of ASE and ridge parameters computed on the test dataset. Since the variation for small values of k is substantial, we have decided to plot the relation I a coordinate system where the abscissa is given in a logarithmic scale (This has the effect that we cannot map $k = 0$). We see that the global minimum of the ASE lies between 0.01 and 0.1. A closer inspection of the raw data shows that the global minimum is close to 0.0175. This value would presumably be a better choice for the final ridge constant. We shall compare the ridge estimates with other models in the sequel and shall use $k=0.01$ and continue using `indiatest_std` as our test dataset.

Variable	OLSEstimate	OLS-VIF	Ridge estimate	Ridge-VIF
Intercept	0	-	0	-
my1	1.46343	22.87	1.6243	11.92
my2	4.25394	75.88	3.4652	18.52
my3	5.86525	91.70	3.7737	17.55
my4	-5.05150	29.51	-2.8986	10.24
my5	4.42072	26.15	5.54774	9.16
mr1	16.50876	4364.05	-0.0091	8.86
mr2	12.01468	19742.77	-1.4939	16.03
mr3	-44.57964	15322.01	-1.7157	14.80
mr4	21.26081	1959.09	1.6050	8.50
mr5	2.37202	2024.28	-2.93727	7.27
mr6	35.68530	3263.03	4.9692	11.27
cpmr16	-42.61545	27460.28	0.9170	2.28
cpmr26	-30.41615	79756.68	-0.6830	5.64
cpmr36	82.31979	55115.19	-1.0561	5.72
cpmr46	-39.36497	9052.51	0.5195	4.38
cpmr56	-9.21564	16264.81	1.01908	2.16
ASE, train	2.46861	-	3.39018	-
ASE, test	14.37973	-	8.59058	-

Parameter estimates, their standard errors and Variation Inflation Factors (VIFs) from an ordinary least squares regression analysis and for a Ridge regression analysis with $k = 0.01$. Average Squared Errors (ASE) for the training and for the test datasets.



The Average Squared Errors based on the evaluation dataset (here = `indiatest_std`) for different ridge parameters. The value for $k = 0$ (14.38) is not shown since we use a logarithmic scale for the ridge constant.

3.5.2 The Lasso, the Elastic Net, and the LARS selection algorithm

The Ridge estimator is not the only shrinking estimator one can think of. An alternative to the Ridge estimator is the so-called **Lasso estimator** introduced by [Tibshirani \(1996\)](#), where we replace the Euclidian (two-norm) constraint $\lambda_2 \|\beta\|_2^2 \leq t$ for some t with a one-norm constraint. The acronym Lasso stands for “least absolute shrinkage and selection operator”. We formulate this in the definition

|||| Definition 3.37

The *Lasso estimator* $\tilde{\beta}_{Lasso}$ is the solution to the minimization problem

$$\min_{\beta} \|\mathbf{y} - \mathbf{x}\beta\|^2 \text{ subject to } \lambda_1 \|\beta\|_1 \leq t \text{ for some } t$$

|||| Remark 3.38

The interesting part here – beyond the shrinking is that some of the coordinates in $\tilde{\beta}_{Lasso}$ will be zero as shown in figure 3.9, where we compare the Lasso-solution with the Ridge solution. In this sense, the Lasso is also a variable selection method.

We shall not go into any details with this, but just observe that some of the shortcomings of the Lasso are solved by the Elastic Net of [Zou and Hastie \(2005\)](#) that combines virtues of the Ridge and the Lasso.

|||| Definition 3.39

The *Elastic Net* estimator $\tilde{\beta}_{Elastic\ Net}$ is the solution to the minimization problem

$$\min_{\beta} \{ \|\mathbf{y} - \mathbf{x}\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \}$$

The Least Angle Regression (and Shrinkage) algorithm LARS (or LAR) was introduced in [Efron et al. \(2004\)](#). This paper gives a very comprehensive exposition including an interesting discussion. LARS is an iterative algorithm similar to the forward selection method with one important exception, namely that it at each step don't. It We shall shortly indicate how the algorithm works and use some figures in facilitating this. One should have in mind that it is not possible to give an adequate illustration of such high-dimensional phenomena. We recall that forward selection finds the variable that maximizes the correlation between the current residual and the new variable. It includes that variable and updates the the ordinary least squares estimators. LARS chooses the same variable but is not letting it have full weight.

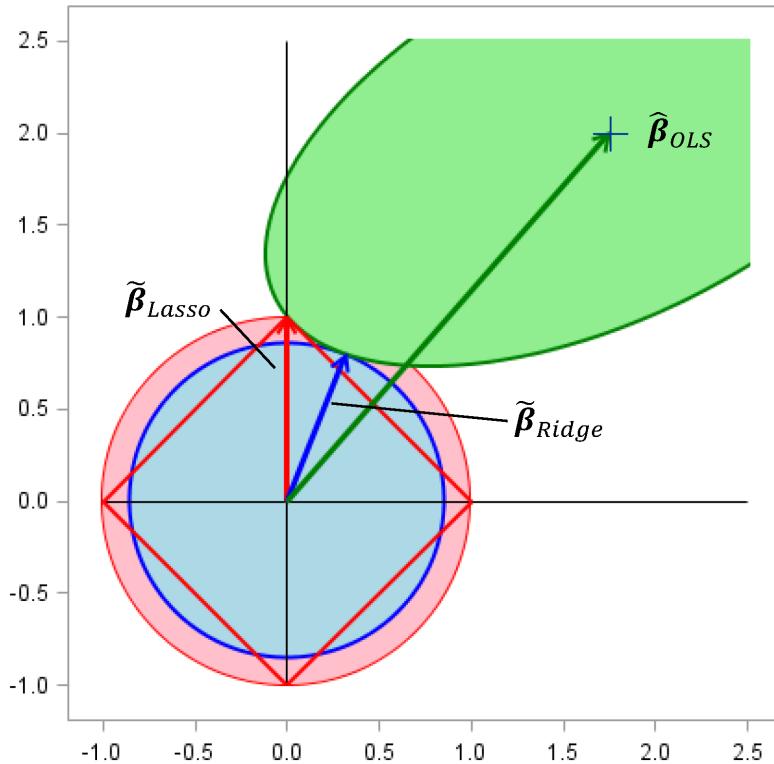


Figure 3.9 – Comparison between Ordinary Least Squares, Ridge, and Lasso estimates of a parameter β . The green ellipse $\{\beta | \|y - x\beta\|_2^2 = c\}$ is a contour curve of the criterion $\phi(\beta)$ we minimize in ordinary least squares. The criterion has minimum in the least squares estimate $\hat{\beta}_{OLS}$. The pink circle $\{\beta | \|\beta\|_2^2 = \beta_1^2 + \beta_2^2 = 1\}$ are the β values with euclidean norm 1 and the red square $\{\beta | \|\beta\|_1 = |\beta_1| + |\beta_2| = 1\}$ are the β values with 1-norm equal to 1. It is seen that $\tilde{\beta}_{Ridge}$ is the shortest (in Euclidian norm) vector with the criterion value $\phi(\tilde{\beta}_{Ridge}) = c$. $\tilde{\beta}_{Lasso}$ is the shortest (in 1-norm) vector with the criterion value $\phi(\tilde{\beta}_{Lasso}) = c$

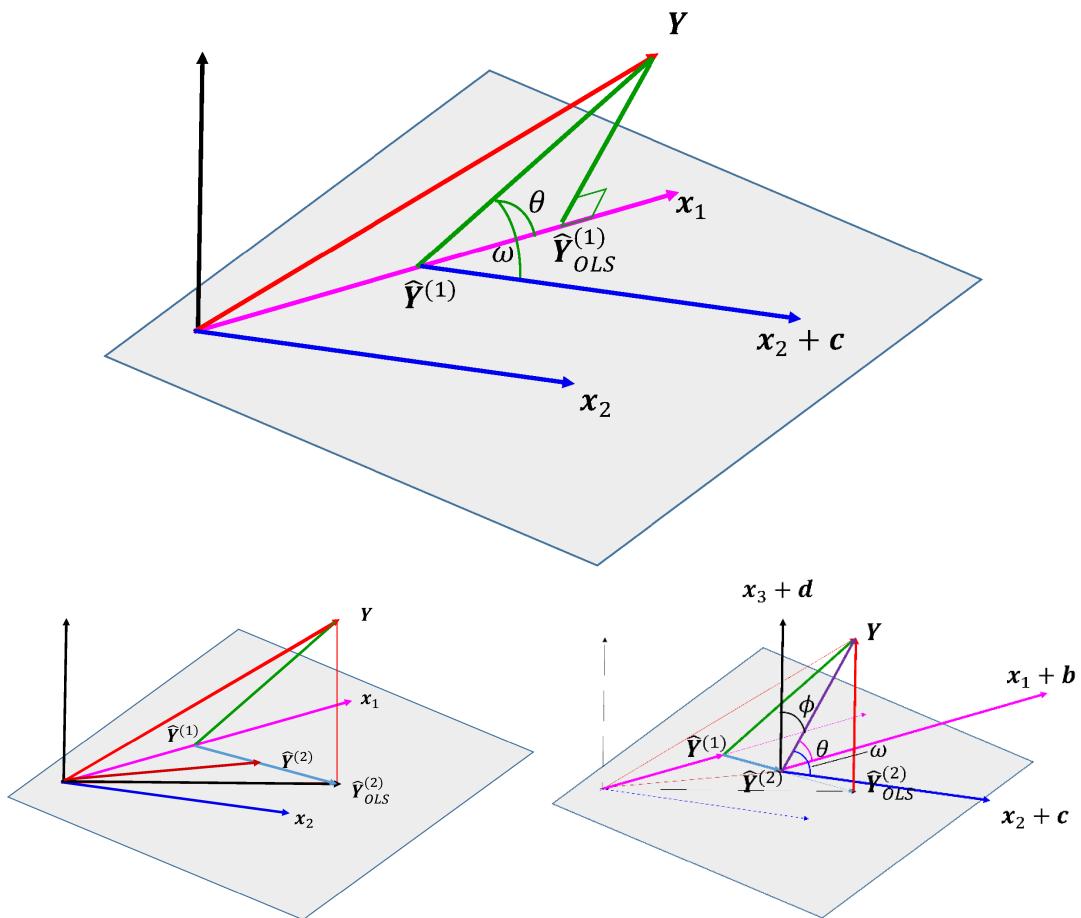


Figure 3.10 – The first two steps in the LARS algorithm, the first in the first row, the second in the second row.

But first some notation. We assume that all vectors – where relevant - are scaled to length 1. By doing this, the empirical correlation between vectors v and w is simply $v^T w$, which is the cosine of the angle between v and w . After step k we have a set of k independent variables. We call this set \mathcal{A}_k , the active set, and the corresponding design matrix $x_{\mathcal{A}_k}$. The OLS estimators based on the active set are – for the regression coefficients

$$\hat{\beta}_{OLS}^{(k)} = \{x_{\mathcal{A}_k}^T x_{\mathcal{A}_k}\}^{-1} x_{\mathcal{A}_k}^T Y$$

and for the estimated mean

$$\hat{Y}_{OLS}^{(k)} = x_{\mathcal{A}_k} \hat{\beta}_{OLS}^{(k)}$$

The corresponding values for the LARS method are called

$$\hat{\beta}_{LARS}^{(k)} = \hat{\beta}^{(k)}$$

$$\hat{Y}_{LARS}^{(k)} = \hat{Y}^{(k)}$$

I. In the first step we shall determine \hat{Y}^1 , positioned somewhere between 0 and $\hat{Y}_{OLS}^{(1)}$.

$$\hat{Y}^{(1)} = \gamma_1 \hat{Y}_{OLS}^{(1)}$$

In other words, we are looking for the LARS update in the direction of the ordinary least squares estimate, but without reaching it. The angle θ between the current residual vector $Y - \hat{Y}^1$ and the first variable chosen, say x_1 decreases from $\pi/2$ (corresponding to correlation 0) to the minimum value for any x (corresponding to the maximum correlation between Y and any x) as \hat{Y}^1 approaches 0. Therefore there is an intermediate position where θ is equal to ω , the angle between $Y - \hat{Y}^1$ and the next variable entered, say x_2 . This corresponds to

$$x_2^T \{Y - \gamma_1 \hat{Y}_{OLS}^{(1)}\} = x_1^T \{Y - \gamma_1 \hat{Y}_{OLS}^{(1)}\}$$

or

$$x_2^T \{Y - \hat{Y}^{(1)}\} = x_1^T \{Y - \hat{Y}^{(1)}\}$$

This equi-angle position is the LARS update. Using the above, we may obtain an expression for γ_1 and thus an expression for computing the first LARS update.

II. In the next step we move from $\hat{Y}^{(1)}$ in the direction of the OLS estimate based on $\{x_1, x_2\}$, i.e. $\hat{Y}_{OLS}^{(2)}$. The LARS update is determined by

$$\hat{Y}^{(2)} - \hat{Y}^{(1)} = \gamma_2(\hat{Y}_{OLS}^{(2)} - \hat{Y}^{(1)})$$

where γ is determined so that

$$x_3^T \left\{ Y - \hat{Y}^{(1)} - \gamma_2 (\hat{Y}_{OLS}^{(2)} - \hat{Y}^{(1)}) \right\} = x_1^T \left\{ Y - \hat{Y}^{(1)} - \gamma_2 (\hat{Y}_{OLS}^{(2)} - \hat{Y}^{(1)}) \right\}$$

or

$$x_3^T \left\{ Y - \hat{Y}^{(2)} \right\} = x_1^T \left\{ Y - \hat{Y}^{(2)} \right\}$$

and

$$x_3^T \left\{ Y - \hat{Y}^{(1)} - \gamma_2 (\hat{Y}_{OLS}^{(2)} - \hat{Y}^{(1)}) \right\} = x_2^T \left\{ Y - \hat{Y}^{(1)} - \gamma_2 (\hat{Y}_{OLS}^{(2)} - \hat{Y}^{(1)}) \right\}$$

or

$$x_3^T \left\{ Y - \hat{Y}^{(2)} \right\} = x_2^T \left\{ Y - \hat{Y}^{(2)} \right\}$$

i.e. so that the angles θ , ω , and ϕ are equal or, equivalently, that the corresponding correlations are equal. γ_2 and thus also the update may be determined using these expressions. We may continue in this way until we eventually reach the full least squares solution. A manageable and precise presentation including the necessary up-date formulas is given in [Sjöstrand et al. \(2018\)](#).

An interesting and very useful property of the sequence of LARS solutions is that we, with a minor modification, also will get a path of Lasso solutions. The modification needed is that whenever a path of parameter estimates crosses zero, we exclude that variable from the model! The elastic net solution is based on Lasso estimation of an augmented model where the design matrix is augmented with a scaled identity matrix and the observations are augmented with a corresponding number of zeros. We shall not go into further details on these computational matters but refer to the numerous research papers (and software manuals) on the topic.

We may continue in this

||| Example 3.40

We consider the data from Example xx and shall compare different selection methods. When using variable selection the stopping criterion is obviously very important. As indicated in example 3.36, one may divide the dataset into training, evaluation and test datasets, and based on the predictive performance on the evaluation set choose the model. If the amount of data is limited, a useful alternative to this is performing a k-fold cross validation, i.e. split the dataset in k parts, keep one of those out of the analysis and fit the model the remaining k-1 parts, and compute the Predicted Residual Sum of Squares (PRESS) statistic

$$PRESS = \sum_{i=1}^m \frac{r_i^2}{(1 - h_{ii})^2}$$

on the omitted part. Here r_i equals the residual at observation i and h_{ii} is it's leverage. Repeat this for all parts. Finally the k PRESS statistics are added to give the CVPRESS statistic – the 'final' estimate of the prediction error for that combination of effects. CVPRESS may thus be used as stopping criterion for the selection process.

We shall illustrate the concepts by means of the SAS procedure [PROC GLMSELECT](#) with the selection criteria Forward, Backward, Stepwise, Lars, Lasso, and Elasticnet. The SAS code used is given below.

```
proc glmselect data=indiatrian_std plots=all ;
model my6 = my1 - my5 mr1 - mr6 cpmr16 cpmr26 cpmr36 cpmr46 cpmr56
sum12 sum23 sum34 sum45 sum56/
selection = lasso (choose = cv steps=15)
cvMethod = split(10) cvDetails = all details=all;
run;
```

Most terms are more or less self-explanatory. The statement `cvMethod=split(k)` asks for a k-fold cross validation of the following type. If $n = k \times m$, then the dataset is split into k parts each of size m having observation numbers

Part 1: $\{1, 1 + k, 1 + 2k, \dots, 1 + (m - 1)k\}$

Part 2: $\{2, 2 + k, 2 + 2k, \dots, 2 + (m - 1)k\}$

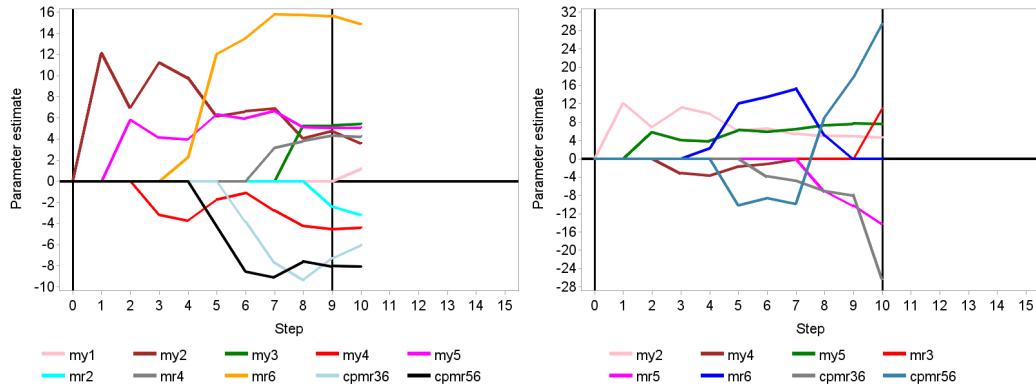
\vdots

Part k: $\{k, k + k, k + 2k, \dots, k + (m - 1)k\}$

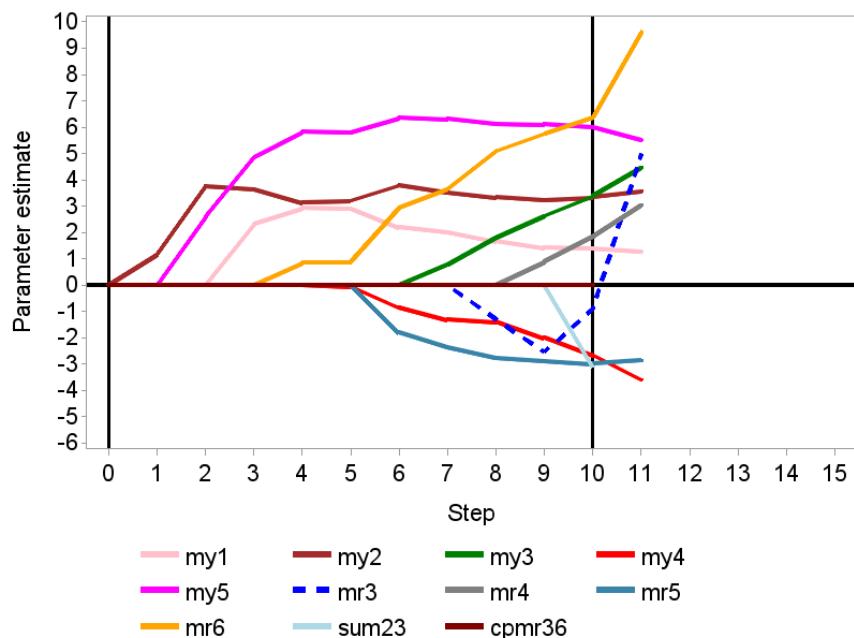
Two other, commonly used methods are `cvMethod=block` and `cvMethod=random`, where the dataset is divided into blocks of m consecutive observations or divided

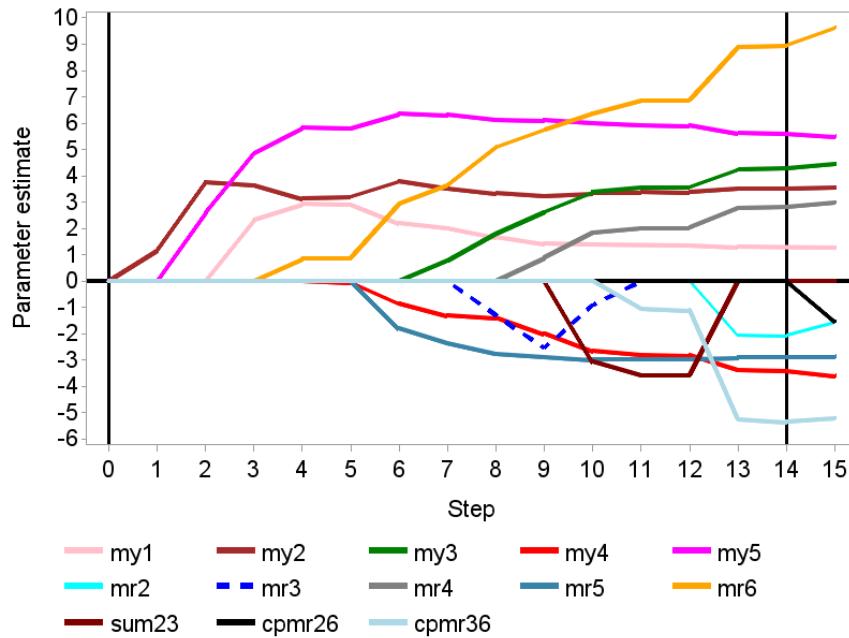
randomly into parts with m observations. There are obvious modifications for n/k not an integer.

We have shown the parameter traces (mapping the estimated parameter as function of step number) in the figures below

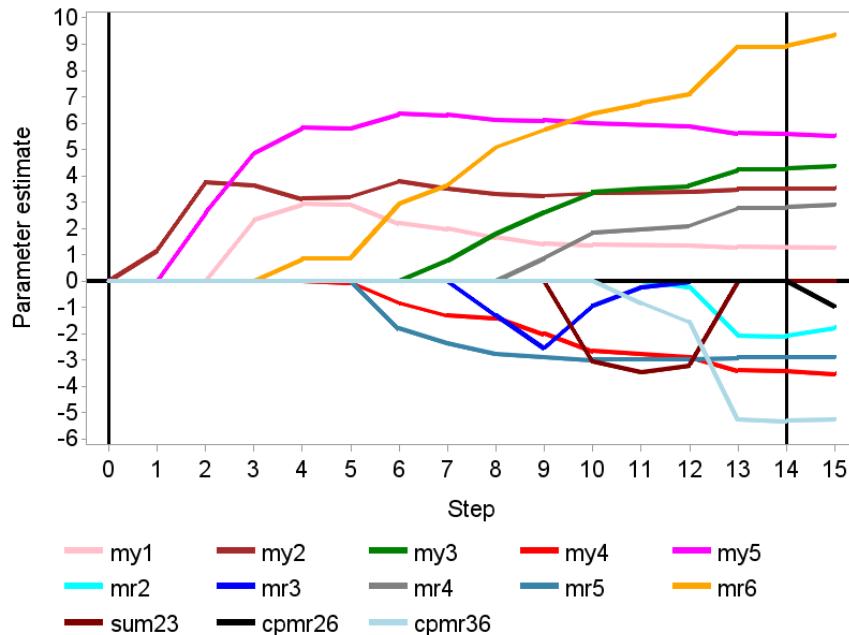


Parameter trace (parameter values mapped against step number) in Forward selection and Stepwise selection. The vertical lines show the optimal values according to the CVPRESS criterion.





Parameter trace in Lars selection and Lasso selection. The vertical lines show the optimal values according to the CVPRESS criterion. The dotted blue line for mr3 shows a sign shift of the coefficient in the Lars selection and therefore mr3 is excluded using the Lasso criterion.



Parameter trace the Elastic Net selection method (bottom). The vertical line at 14 represents the optimal value according to the CVPRESS criterion.

A summary of the effects entered and removed is presented in first table below and the average squared errors based on the training and on the test datasets are given in the second. In the first table we have also presented the variables retained with the backward elimination method. In [PROC GLMSELECT](#), variables of the form $sum_{ij} = mr_i + mr_j$ are automatically discarded from the Backward procedure due to the linear dependency they cause in the set of covariates.

Step	Forward	Stepwise		LARS	LASSO		Elastic net		Backward
	Effect Entered	Effect entered	Effect Re-removed	Effect Entered	Effect Entered	Effect Re-removed	Effect Entered	Effect Re-removed	Effects retained
0	Intercept	Intercept		Intercept	Intercept		Intercept		Intercept
1	my2	my2		my2	my2		my2		my1
2	my5	my5		my5	my5		my5		my2
3	my4	my4		my1	my1		my1		my3
4	mr6	mr6		mr6	mr6		mr6		my4
5	cpmr56	cpmr56		my4	my4		my4		my5
6	cpmr36	cpmr36		mr5	mr5		mr5		mr1
7	mr4		my4	my3	my3		my3		mr3
8	my3	mr5		mr3	mr3		mr3		mr4
9	mr2		mr6	mr4	mr4		mr4		mr6
10	my1	mr3		sum23	sum23		sum23		cpmr16
11				cpmr36	cpmr36		cpmr36		cpmr26
12						mr3	mr2		cpmr36
13					mr2			mr3	cpmr46
14						sum23		sum23	
15					cpmr26		cpmr26		

	OLS	Ridge (0.01)	Forward	Backward	Stepwise	Lars	Lasso	Elastic net
Train	2.46861	3.39018	2.98185	2.51312	3.69349	3.33582	3.02352	3.02490
Test	14.37973	8.59058	9.90599	14.83447	9.65059	8.45545	8.80589	8.80306

We see that there are some consistency in the selection patterns when looking at the first variables entered. When it comes to prediction in the test dataset, OLS and Backward elimination give the poorest results. Forward selection and Stepwise regression present a clear improvement, but the smallest prediction errors occur when using a shrinking method – in the present case with Lars selection as (marginally) the best.

3.5.3 Logistic regression

We start with an illustrative example

||| Example 3.41

We consider some data which concern conserving of iron items from the Iron Age. The data are from [Salomonsen \(1977\)](#). 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

$$X_i \sim B(n_i, p_i)$$

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

$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i},$$

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

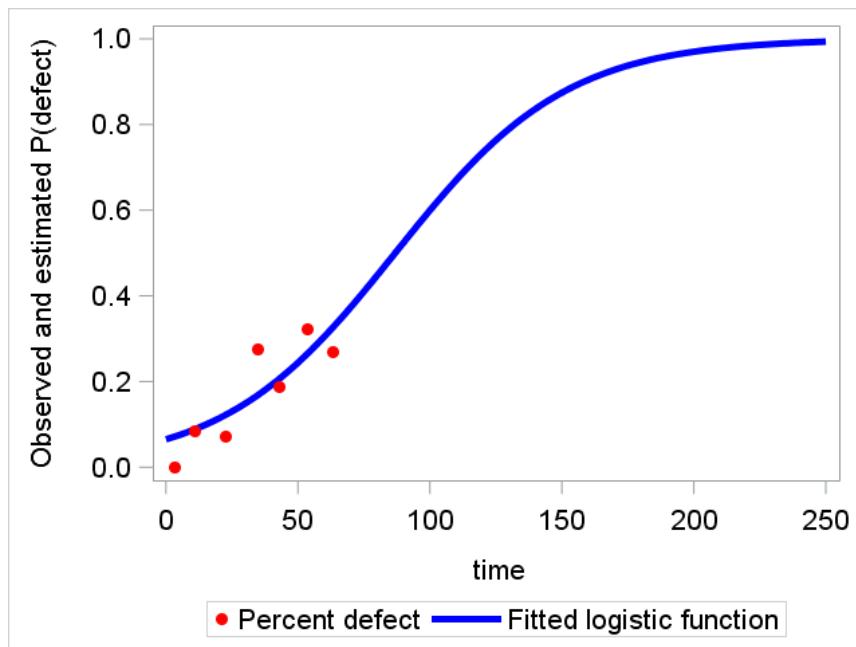
$$\begin{aligned} \log L(\alpha, \beta) &= \\ &= c - \sum_i x_i \log(1 + \exp(-\alpha - \beta t_i)) - \sum_i (n_i - x_i)(\alpha + \beta t_i) \\ &\quad - \sum_i (n_i - x_i) \log(1 + \exp(-\alpha - \beta t_i)) \\ &= c - \sum_i n_i \log(1 + \exp(-\alpha - \beta t_i)) - \sum_i (n_i - x_i)(\alpha + \beta t_i). \end{aligned}$$

These equations are solved – or the likelihood function is maximized by alternative methods - using the following SAS code

```
proc logistic data=conservation plots=all;
model def/artefacts = time;
run;
```

The maximum likelihood estimates for the parameters in the logistic model for the probability of observing an artefact with further degradation (defect) are given below.

Parameter	DF	Estimate	Standard	WaldChi-Square	Pr > ChiSq
Intercept	1	-2.6661	0.4046	43.4116	<.0001
time	1	0.0307	0.00883	12.1149	0.0005



The observed and estimated defect proportions for the annealed iron artefacts.

In the example above, we see that we have a linear relation between the logit and the time, the explanatory variable. This may be generalized to the case with a multidimensional explanatory variable. We formulate this in

||| Definition 3.42

We thus again consider a binary response Y ($= 0$ or 1), this time accompanied by a set of explanatory variables x so that the probability of getting a 0 depends on x as follows

$$p(x) = P\{Y = 0 \mid x\} = \frac{1}{1 + \exp(-\alpha - \beta^T x)},$$

corresponding to that the logit

$$\text{logit}(p(x)) = \log \frac{p(x)}{1 - p(x)} = \alpha + \beta^T x$$

is a linear (affine) function of the explanatory variables. This model is called the *logistic regression model*.

A parameter often used in interpretation of the parameters in the logistic regression model is the odds ratio given in

||| Definition 3.43

The parameter

$$o(\mathbf{x}) = \frac{p(\mathbf{x})}{1 - p(\mathbf{x})} = e^{\alpha + \beta^T \mathbf{x}}$$

is called the *odds*, and the ratio between two odds evaluated at two different \mathbf{x} -values

$$\frac{o(\mathbf{x})}{o(\mathbf{z})} = e^{\beta^T (\mathbf{x} - \mathbf{z})}$$

is the odds ratio. If \mathbf{x} and \mathbf{z} only deviate with 1 in the j th coordinate, we get the odds ratio

$$e^{\beta_j}$$

i.e. independent of \mathbf{x} .

In everyday language, describing probabilities of events by using odds is fairly ambiguous. In some cases the term odds is used synonymously with probability. In gambling a statement like the odds of a given outcome is say $s : t$ ("s" to "t") will in general mean that the probability p of the outcome is the solution to

$$\frac{p}{1 - p} = \frac{s}{t} \text{ or } p = \frac{s}{s + t}$$

so – to make things clear - odds 3 to 1, i.e. 3, corresponds to a probability 0.75, and a probability 0.4 corresponds to odds 2 to 3, i.e. 0.66667.

The odds ratio (and confidence limits) are part of the standard output from **PROC LOGISTIC**. From the artefact example we get

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
time	1.031	1.014	1.049

and it is easily verified that this value is $\exp(0.0307)$. the interpretation of this value is then that the odds of having a defect increase with 3.1% per year (irrespective of the time that has elapsed).

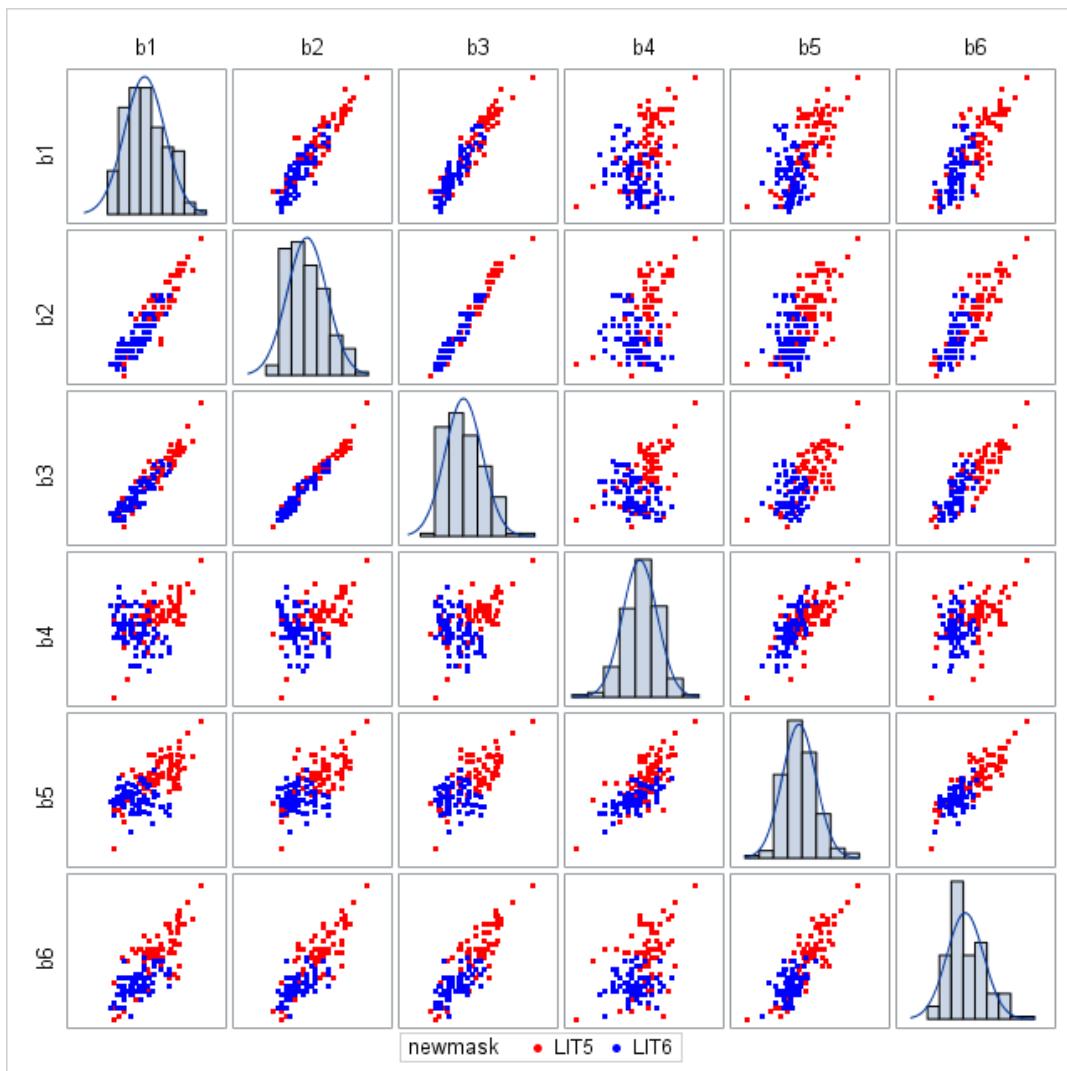
It is of course arbitrary that we have modelled the probability of getting a zero. We could just as well have modelled the complementary value, the probability of getting a one. One should notice however, that the Logistic procedure in SAS models “the probability of the lower response levels”.

The logistic regression model may be analyzed using maximum likelihood methods using vocabulary and yielding results that may be interpreted very similarly to results from linear regression analysis. We shall not go into details with this, but refer to the substantial literature and the extensive descriptions in e.g. the SAS manuals that also contains useful references.

We shall illustrate some of the uses of these models in

||| Example 3.44

We consider 161 randomly selected pixels from a Landsat satellite image from Ymer Ø in Central East Greenland. 78 of those pixels come from a lithological unit that we call unit 5, and the remaining 83 from another unit, no 6. We have measured the electromagnetic reflection in 6 bands, b1 – b6. The scatterplots and the histograms are shown in the figure below.



Histograms and scatterplots for the variables b1 – b6. In the scatterplots, we have depicted observations from lithology 5 as red, and pixels from lithology 6 as blue.

We can now model the probability of coming from lithology 5 (mask=5 in the dataset train56) using the SAS code presented in the textbox below

```
Title 'Logistic Regression on Ymer Data';
proc logistic data=train56 outest=betas;
model mask(event='5') = b1-b6;
output out=pred p=phat predprob = ( individual crossvalidate);
run;
```

The estimated probabilities of getting lithologies 5 and 6 are given in the dataset pred, and crossvalidate in the prediction probability statement give an estimate of class belonging for each case based on parameter estimates excluding that case. By fixing a threshold between 0 and 1 we may then 'classify' the pixels into one of the lithologies depending on whether the (estimated) probability of class belonging is under or above the chosen threshold.

We have shown the estimated parameters in the table below

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	StandardError	WaldChi-Square	Pr > ChiSq
Intercept	1	-17.9135	4.5008	15.8409	<.0001
b1	1	0.0459	0.0853	0.2897	0.5904
b2	1	-0.0416	0.2339	0.0316	0.8589
b3	1	0.0486	0.1876	0.0671	0.7956
b4	1	0.0662	0.0854	0.6005	0.4384
b5	1	0.0783	0.0678	1.3343	0.2480
b6	1	0.1475	0.0970	2.3102	0.1285

The classification results are shown in the table below. This should only be thought of as an illustration! It is by no means a given thing that 50% should be used as threshold. The evaluation of such items is a main topic in Chapter 5, Discriminant analysis and classification. For later reference we have in the final table shown the result of using two normal distribution based methods, linear and quadratic discriminant analysis.

		Logistic				
		Individual		Cross Validation		
From		5	6	5	6	Total
5		66	12	63	15	78
6		9	74	11	72	83
Total		75	86	74	87	161

	Classified into								
	Linear				Quadratic				
	Resubstitution		Cross Validation		Resubstitution		Cross Validation		
From	5	6	5	6	5	6	5	6	Total
5	62	16	61	17	69	9	68	10	78
6	6	77	9	74	7	76	11	72	83
Total	68	93	70	91	76	85	79	82	161

3.5.4 Non-linear regression

In ordinary regression analysis, the mean of the observed random variables are linear functions of the unknown parameters. However, one will often be in a situation where this is not the case. Under such circumstances, one may still estimate the unknown parameters using the maximum likelihood method com-

bined with suitable routines for numerical maximization of the likelihood function. Furthermore, one may under mild regularity conditions use the results on asymptotic distributions of the ML estimators to get uncertainties and observed levels of significance (observed p-values)

The SAS procedure **PROC NLIN** for non-linear regression collects the relevant approximative results and put them in a framework very similar to linear regression. We shall illustrate this in the next

||| Example 3.45

The data in this example regards fermentation of salamis, and the possibilities of pausing the fermentation process by freezing the samples. This is of interest in order to enable a sensory panel to evaluate salami samples from the same batch but of different ages at the same time. The situation is described in [Trinderup et al. \(2018a\)](#). To investigate the possibility, we are considering the shrinking of salamis over time and with the measured width in mm, we are obtaining an objective measure of time change. We have 132 samples of salamis and have measured the width for non-frozen and for frozen and subsequently thawed samples at timepoints between 2 and 42 days. Mean and standard deviation are given below.

Variable	N	Mean	Std Dev	Minimum	Maximum
day	132	12.73	11.50	2	42
f	132	0.45	0.50	0	1
mmWidth	132	56.60	2.78	31.35	60.98

A possible consequence of the freezing and thawing process is accelerated loss of water. Therefore we may see a more rapid time development for the diameter for the frozen samples, and it would be very interesting to see whether this may be described by a time-shift of the non-frozen curve. If the amount of water evaporating is proportional to the amount of water present, the development of the diameter over time will be exponential. We therefore end up with a model like

$$E(\text{Width at time } t) = \mu(t) = \begin{cases} \lambda + \exp\left(-\frac{t-\alpha}{\beta}\right), & f = 0 \\ \lambda + \exp\left(-\frac{t-\alpha-\delta}{\beta}\right), & f = 1 \end{cases}$$

where δ is the time-shift parameter. We can estimate the parameters by means of **PROC NLIN**, using the SAS code given below.

```

proc nlin data = salami42;
parameters alfa = 25 lambda = 50 beta = 10 delta = -1 ;
if (f = 0) then do;
mean = lambda+exp( - (day - alfa)/beta);
end;
else do;
mean = lambda+exp( - (day - alfa - delta)/beta);
end;
model mmwidth = mean;
output out = predwidth predicted = p;
run;

```

The results are presented in tables below. And we see that we – except for the iteration summary – are obtaining the same type of information as we do in ordinary linear regression with the – important – difference, that uncertainties and probabilities are approximate.

Iterative Phase					
Iter	alfa	lambda	beta	delta	Sum of Squares
0	25.0000	50.0000	10.0000	-1.0000	486.0
1	25.2488	52.3734	10.9216	-0.9802	63.2494
2	25.0186	52.4114	11.0039	-0.9768	60.1584
3	24.9723	52.4161	10.9889	-0.9772	60.1556
4	24.9801	52.4152	10.9922	-0.9771	60.1556
5	24.9783	52.4154	10.9915	-0.9771	60.1556
6	24.9787	52.4154	10.9916	-0.9771	60.1556

Iteration summary using the Gauss-Newton method.

Source	DF	Sum of Squares	Mean Square	F Value	ApproxPr > F
Model	3	950.0	316.7	673.84	<.0001
Error	128	60.1556	0.4700		
Corrected Total	131	1010.2			

The resulting ANOVA table.

Parameter	Estimate	ApproxStd Error	Approximate ConfidenceLimits	95% ConfidenceLimits
alfa	24.9787	1.8507	21.3168	28.6406
lambda	52.4154	0.2260	51.9681	52.8627
beta	10.9916	0.7784	9.4514	12.5319
delta	-0.9771	0.2684	-1.5083	-0.4460

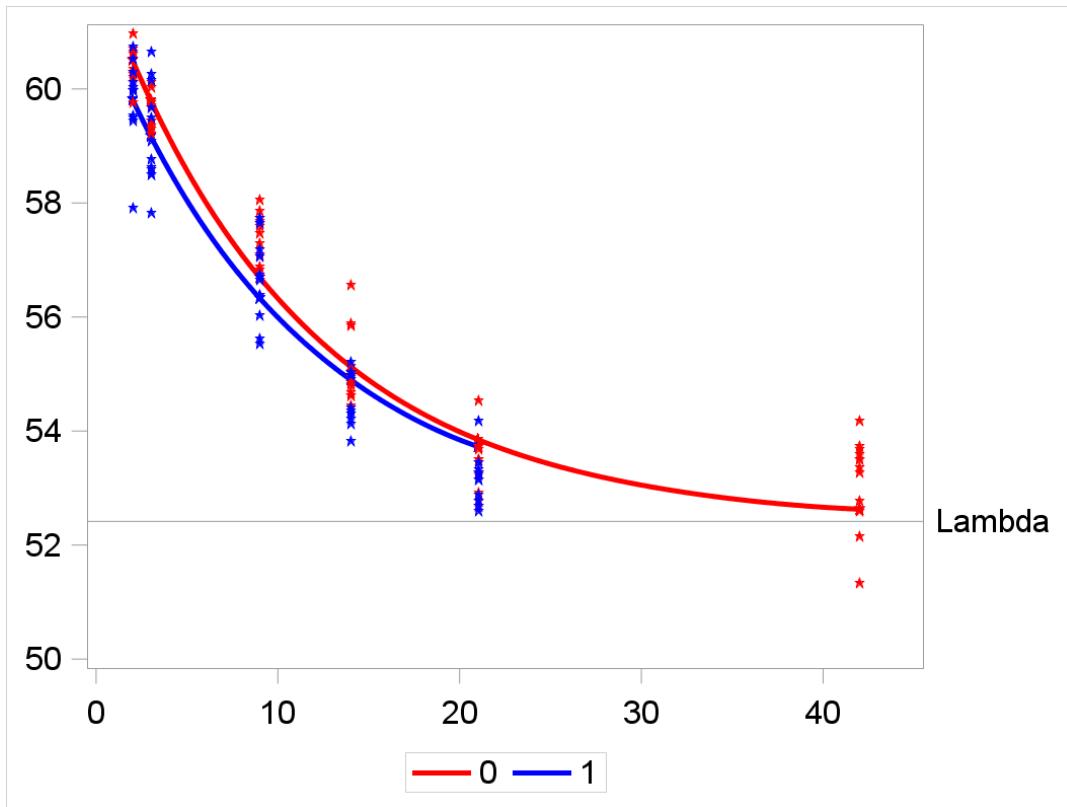
The resulting parameter estimates and their approximate uncertainties.

Approximate Correlation Matrix				
	alfa	lambda	beta	delta
alfa	1.0000000	-0.9152783	0.9895138	-0.2008268
lambda	-0.9152783	1.0000000	-0.8718694	0.0957032
beta	0.9895138	-0.8718694	1.0000000	-0.1445356
delta	-0.2008268	0.0957032	-0.1445356	1.0000000

The approximate correlation matrix for the estimated parameters.

The time development is shown in the figure below, and we see that the model with a time-shift of one day gives a reasonable explanation of the data. We notice that the estimated correlations between the time shift δ and the other parameters are small compared to the other correlations. This justifies a larger confidence in this specific parameter estimate.

The time deveopment of the salami width for non-frozen (red) and frozen (blue) samples corresponding to a timeshift -0.98, very close to one day.



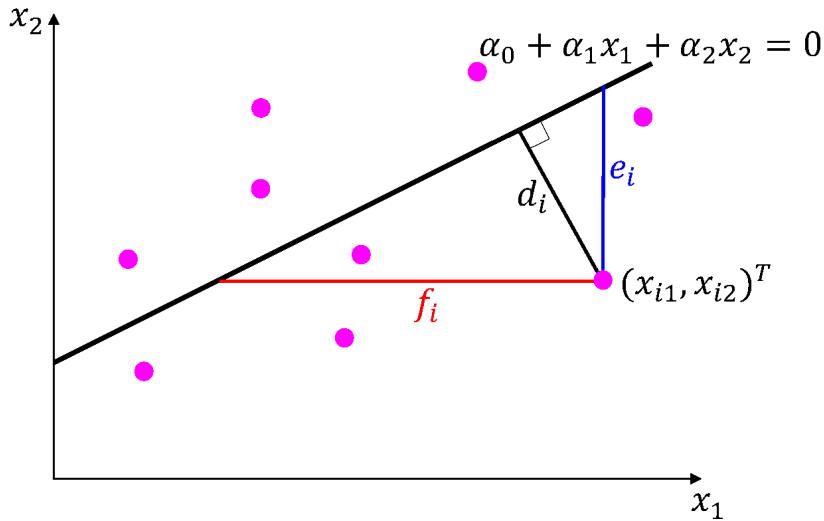


Figure 3.11 – In regression of x_2 on x_1 the line parameters are determined by minimizing the sum of the squares of the vertical distances, i.e. $\sum e_i^2$, in regression of x_1 on x_2 by minimizing the sum of squares of the horizontal distances, i.e. $\sum f_i^2$. In orthogonal regression the parameters are determined by minimizing the sum of squares of the orthogonal distances, i.e. $\sum d_i^2$.

3.5.5 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, see figure 3.11.

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. [Kendall and Stuart \(1967\)](#) p. 392) that this plane gives the maximum likelihood estimator of the α 's. We formulate the solution to the problem in

|||| Theorem 3.46

We consider n points $x_1, \dots, x_n \in \mathbb{R}^p$ and the hyperplane

$$\alpha_0 + \alpha_1 x_1 + \dots + \alpha_p x_p = 0$$

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 smalles eigenvalue of the empirical 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

$$\begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

The distance from a point with the coordinates $x = (x_1, \dots, x_p)^T$ to the hyperplane ia

$$\frac{|\alpha_0 + \alpha_1 x_1 + \dots + \alpha_p x_p|}{\sqrt{\alpha_1^2 + \dots + \alpha_p^2}}.$$

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

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \alpha,$$

then the equations system can be written as

$$\hat{\Gamma} \alpha = \frac{\lambda}{n-1} \alpha,$$

i.e. α 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 3.47

The result which has been stated in the theorem has a close connection to the results which will be shown in chapter 6.1 on principal components.

|||| 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

$$X_i \sim N_p(\mu, \Sigma),$$

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 1.57 p. 56 we have the following results on the empirical mean vector \bar{X} and the empirical

variance-covariance matrix \mathbf{S}

$$\begin{aligned}\bar{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i & \sim \mathcal{N}_p(\boldsymbol{\mu}, \frac{1}{n} \boldsymbol{\Sigma}) \\ \mathbf{S} &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T & \sim W(n-1, \frac{1}{n-1} \boldsymbol{\Sigma})\end{aligned}$$

$\bar{\mathbf{X}}$ 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

$$\begin{aligned}\mathbf{Y} &\sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ m\mathbf{U} &\sim W(m, \boldsymbol{\Sigma}),\end{aligned}$$

furthermore let \mathbf{Y} and \mathbf{U} be stochastically independent. We now let

$$T^2 = \mathbf{Y}^T \mathbf{U}^{-1} \mathbf{Y}.$$

Then the following holds

$$\frac{m-p+1}{mp} T^2 \sim F(p, m-p+1; \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}),$$

i.e. the left hand side is non-centrally F-distributed with non-centrality parameter $\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$ 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

$$\frac{m-p+1}{mp} T^2 \sim F(p, m-p+1).$$

||| Proof omitted

See e.g. [Anderson \(1958\)](#), p. 106.

We now have the following main result

|||| **Theorem 4.2**

We will use the notation

$$T^2 = n(\bar{X} - \mu_0)^T \mathbf{S}^{-1} (\bar{X} - \mu_0),$$

where \bar{X} , μ_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 4.1 we find that

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

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

$$T^2 = \frac{\det[\mathbf{S} + n(\bar{X} - \mu_0)(\bar{X} - \mu_0)^T]}{\det[\mathbf{S}]} - 1$$

||| Proof omitted

Purely technical and follows by using theorem A.7 p. 445 on the matrix

$$\begin{bmatrix} -1 & \sqrt{n}(\bar{X} - \mu_0)^T \\ \sqrt{n}(\bar{X} - \mu_0) & \mathbf{S} \end{bmatrix}$$

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

$$n(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T = \begin{pmatrix} 41.68 & 35.70 \\ 35.70 & 30.58 \end{pmatrix},$$

and

$$\mathbf{s} + n(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T = \begin{pmatrix} 42.86 & 35.39 \\ 35.39 & 31.45 \end{pmatrix}.$$

Then using theorem 4.5

$$t^2 = \frac{95.49}{0.9293} - 1 = 101.75.$$

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

$$\{\mu | n(\bar{x} - \mu)^T \mathbf{s}^{-1}(\bar{x} - \mu) \leq \frac{(n-1)p}{n-p} F(p, n-p)_{1-\alpha}\},$$

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

$$7(19.66 - \mu_1, 5.31 - \mu_2) \mathbf{s}^{-1} \begin{pmatrix} 19.66 - \mu_1 \\ 5.31 - \mu_2 \end{pmatrix} = \frac{12}{5} F(2, 5)_{0.95}$$

or

$$(19.66 - \mu_1, 5.31 - \mu_2) \mathbf{s}^{-1} \begin{pmatrix} 19.66 - \mu_1 \\ 5.31 - \mu_2 \end{pmatrix} = 1.9851.$$

We find

$$\mathbf{s}^{-1} = \begin{pmatrix} 0.9341 & 0.3310 \\ 0.3310 & 1.2692 \end{pmatrix}$$

with the eigenvalues 1.4727 and 0.7307 and the corresponding (normed) eigenvectors

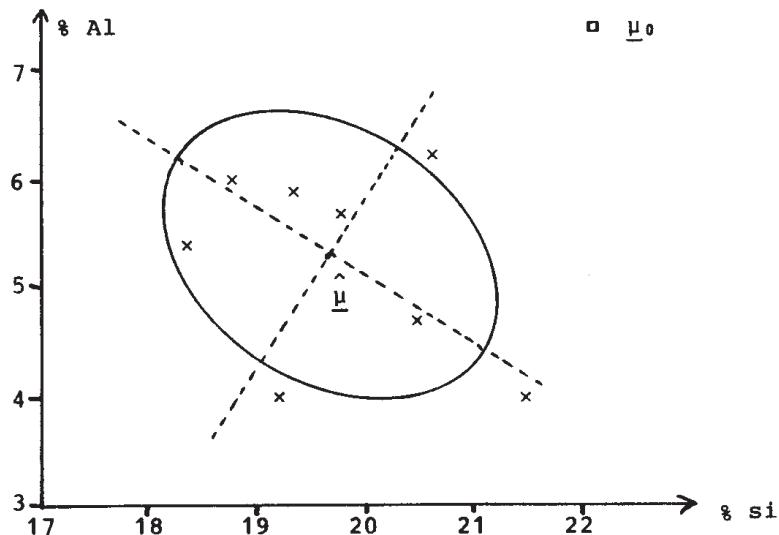
$$\begin{pmatrix} 0.5236 \\ 0.8520 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -0.8520 \\ 0.5236 \end{pmatrix}.$$

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

$$\frac{y_1^2}{1.1610^2} + \frac{y_2^2}{1.6482^2} = 1$$



Observations and confidence region.

In the figure above the confidence region and the observations are shown. Furthermore $\mu_0 = (22.10, 7.40)^T$ 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

$$\begin{aligned} X_i &\sim N_p(\mu, \Sigma) \\ Y_i &\sim N_p(\nu, \Sigma), \end{aligned}$$

and we wish to test

$$H_0 : \mu = \nu \quad \text{against} \quad H_1 : \mu \neq \nu.$$

We use the notation

$$\begin{aligned}\bar{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \\ \bar{\mathbf{Y}} &= \frac{1}{m} \sum_{i=1}^m \mathbf{Y}_i \\ \mathbf{S}_1 &= \frac{1}{n-1} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T \\ \mathbf{S}_2 &= \frac{1}{m-1} \sum_{i=1}^m (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})^T \\ \mathbf{S} &= \frac{(n-1)\mathbf{S}_1 + (m-1)\mathbf{S}_2}{n+m-2}\end{aligned}$$

From theorem 1.57 and theorem 1.56 we have

$$\begin{aligned}\bar{\mathbf{X}} &\sim N_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}) \\ \bar{\mathbf{Y}} &\sim N_p(\boldsymbol{\nu}, \frac{1}{m}\boldsymbol{\Sigma}) \\ \mathbf{S} &\sim W(n+m-2, \frac{1}{n+m-2}\boldsymbol{\Sigma}).\end{aligned}$$

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

$$T^2 = \frac{nm}{n+m} (\bar{\mathbf{X}} - \bar{\mathbf{Y}})^T \mathbf{S}^{-1} (\bar{\mathbf{X}} - \bar{\mathbf{Y}}).$$

Then the critical region for a test of H_0 against H_1 at level α is equal to

$$C = \{ \mathbf{x}_1, \dots, \mathbf{x}_n, \mathbf{y}_1, \dots, \mathbf{y}_m \mid \frac{n+m-p-1}{(n+m-2)p} t^2 > F(p, n+m-p-1)_{1-\alpha} \}$$

Here t^2 is the observed value of T^2 .

||| Proof

From lemma 4.5 and from the above mentioned relationships we find that

$$\frac{n+m-p-1}{(n+m-2)p} T^2 \sim F(p, n+m-p-1; (\mu - \nu)^T \Sigma^{-1} (\mu - \nu)),$$

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

$$\left\{ \delta \mid \frac{nm}{n+m} (\bar{x} - \bar{y} - \delta)^T \mathbf{s}^{-1} (\bar{x} - \bar{y} - \delta) \leq \frac{(n+m-2)p}{n+m-p-1} F(p, n+m-p-1)_{1-\alpha} \right\}.$$

||| 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 \mathbf{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. [Anderson \(1958\)](#), 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 below

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

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

$$X_i \sim N(\mu, \Sigma)$$

and

$$Y_i \sim N(\nu, \Sigma),$$

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

$$t^2 = \frac{16 \cdot 16}{16 + 16} (\bar{x} - \bar{y})^T \mathbf{s}^{-1} (\bar{x} - \bar{y}) = 52.4.$$

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 $\mathbf{Y}_1, \dots, \mathbf{Y}_n$,

$$\mathbf{Y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}).$$

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

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

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

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1^T \\ \vdots \\ \mathbf{Y}_n^T \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{|1} & \cdots & \mathbf{Y}_{|p} \end{bmatrix}$$

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

$$\mathbb{E}(\mathbf{Y}_i^T) = (x_{i1} \cdots x_{ik}) \begin{pmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{pmatrix} = \mathbf{x}_i^T \boldsymbol{\theta},$$

where $\mathbf{x}_i^T = \mathbf{x}_{-i}$ is the i 'th row in the \mathbf{x} -matrix. This readily gives

$$\mathbb{E}(\mathbf{Y}_i) = \boldsymbol{\theta}^T \mathbf{x}_i.$$

If the observations are rearranged into a column vector

$$\tilde{\mathbf{Y}} = \text{vc}(\mathbf{Y}) = \begin{bmatrix} \mathbf{Y}_{1|} \\ \vdots \\ \mathbf{Y}_{p|} \end{bmatrix},$$

we find from theorem 1.9, p. 10, that

$$\mathbb{D}(\mathbf{Y}) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n = \begin{bmatrix} \sigma_1^2 \mathbf{I}_n & \cdots & \sigma_{1p} \mathbf{I}_n \\ \vdots & & \vdots \\ \sigma_{p1} \mathbf{I}_n & \cdots & \sigma_p^2 \mathbf{I}_n \end{bmatrix},$$

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

$$\hat{\boldsymbol{\theta}} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y}.$$

||| Proof omitted

See e.g. [Anderson \(1958\)](#).

||| Remark 4.15

We see that

$$\hat{\theta}_{j|} = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y}_{j|},$$

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

$$M = (\mathbf{Y} - \mathbf{x} \theta)^T (\mathbf{Y} - \mathbf{x} \theta) - (\mathbf{Y} - \mathbf{x} \hat{\theta})^T (\mathbf{Y} - \mathbf{x} \hat{\theta})$$

is positive semidefinite. From this follows that

$$\text{eig}_i(\mathbf{Y} - \mathbf{x} \theta)^T (\mathbf{Y} - \mathbf{x} \theta) \geq \text{eig}_i(\mathbf{Y} - \mathbf{x} \hat{\theta})^T (\mathbf{Y} - \mathbf{x} \hat{\theta}),$$

where eig_i corresponds to the i 'th largest eigenvalue. From this follows again that $\hat{\theta}$ minimises

$$\det(\mathbf{Y} - \mathbf{x} \theta)^T (\mathbf{Y} - \mathbf{x} \theta)$$

and

$$\text{tr}(\mathbf{Y} - \mathbf{x} \theta)^T (\mathbf{Y} - \mathbf{x} \theta).$$

|||| **Remark 4.17**

Above we have silently assumed that $\mathbf{x}^T \mathbf{x}$ 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

$$\begin{aligned}\hat{\Sigma}^* &= \frac{1}{n} \sum_{i=1}^n (\mathbf{Y}_i - \hat{\theta}^T \mathbf{x}_i)(\mathbf{Y}_i - \hat{\theta}^T \mathbf{x}_i)^T \\ &= \frac{1}{n} (\mathbf{Y} - \mathbf{x}\hat{\theta})^T (\mathbf{Y} - \mathbf{x}\hat{\theta}) \\ &= \frac{1}{n} [\mathbf{Y}^T \mathbf{Y} - (\mathbf{x}\hat{\theta})^T (\mathbf{x}\hat{\theta})].\end{aligned}$$

The (i, j) 'th element can also be written

$$\hat{\sigma}_{ij}^* = \frac{1}{n} (\mathbf{Y}_{i|} - \mathbf{x}\hat{\theta}_{i|})^T (\mathbf{Y}_{j|} - \mathbf{x}\hat{\theta}_{j|}).$$

|||| **Proof**

The many identities between $\hat{\Sigma}$'s elements are found by simple matrix manipulations. For the results we refer to [Anderson \(1958\)](#).

■

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

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \text{vc}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\theta}_{|1} \\ \vdots \\ \boldsymbol{\theta}_{|p} \end{bmatrix} \\ \hat{\boldsymbol{\theta}} &= \text{vc}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} \hat{\boldsymbol{\theta}}_{|1} \\ \vdots \\ \hat{\boldsymbol{\theta}}_{|p} \end{bmatrix}.\end{aligned}$$

Then we have that $\hat{\boldsymbol{\theta}}$ is normally distributed

$$\hat{\boldsymbol{\theta}} = \text{vc}(\hat{\boldsymbol{\theta}}) \sim N_{pk}(\tilde{\boldsymbol{\theta}}, \boldsymbol{\Sigma} \otimes (\mathbf{x}^T \mathbf{x})^{-1}),$$

and $n\hat{\boldsymbol{\Sigma}}^*$ is Wishart distributed

$$n\hat{\boldsymbol{\Sigma}}^* \sim W(n - k, \boldsymbol{\Sigma}).$$

Finally $\hat{\boldsymbol{\Sigma}}^*$ and $\hat{\boldsymbol{\theta}}$ and therefore also $\hat{\boldsymbol{\Sigma}}^*$ and $\hat{\boldsymbol{\theta}}$ are stochastically independent.

|||| Proof

It is trivial that

$$E(\hat{\boldsymbol{\theta}}) = E[(\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{Y}] = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T \mathbf{x} \boldsymbol{\theta} = \boldsymbol{\theta}$$

and from this it follows that $E(\hat{\boldsymbol{\theta}}) = \tilde{\boldsymbol{\theta}}$. Furthermore $\hat{\boldsymbol{\theta}}$ is of course normally distributed.

Finally we have that

$$D(\hat{\boldsymbol{\theta}}_{|i}) = \sigma_{ii}(\mathbf{x}^T \mathbf{x})^{-1}$$

and

$$C(\hat{\boldsymbol{\theta}}_{|i}, \hat{\boldsymbol{\theta}}_{|j}) = (\mathbf{x}^T \mathbf{x})^{-1} \mathbf{x}^T C(\mathbf{Y}_{|i}, \mathbf{Y}_{|j}) \mathbf{x} (\mathbf{x}^T \mathbf{x})^{-1} = \sigma_{ij}(\mathbf{x}^T \mathbf{x})^{-1}.$$

From this the result concerning the variance covariance matrix for $\hat{\boldsymbol{\theta}}$ is readily seen.

The result concerning the distribution of $\hat{\boldsymbol{\Sigma}}^*$ and concerning the independence of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}^*$ 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. [Anderson \(1958\)](#).

■

From the theorem we readily find

||| Corollary 4.20

The unbiased estimate for Σ is equal to

$$\hat{\Sigma} = \frac{n}{n-k} \hat{\Sigma}^* = \frac{1}{n-k} (\mathbf{Y} - \mathbf{x} \hat{\theta})^T (\mathbf{Y} - \mathbf{x} \hat{\theta}).$$

||| 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

$$H_0 : \mathbf{A}\boldsymbol{\theta}\mathbf{B}^T = \mathbf{C} \quad \text{against} \quad H_1 : \mathbf{A}\boldsymbol{\theta}\mathbf{B}^T \neq \mathbf{C},$$

where $\mathbf{A}(r \times k)$, $\mathbf{B}(s \times p)$ and $\mathbf{C}(r \times s)$ are given (known) matrices. We introduce

$$\begin{aligned}\Delta &= \mathbf{A}\hat{\boldsymbol{\theta}}\mathbf{B}^T - \mathbf{C} \\ \mathbf{R} &= n\hat{\Sigma}^* = (\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}})^T(\mathbf{Y} - \mathbf{x}\hat{\boldsymbol{\theta}}) = \mathbf{Y}^T\mathbf{Y} - \hat{\boldsymbol{\theta}}^T(\mathbf{x}^T\mathbf{x})\hat{\boldsymbol{\theta}}\end{aligned}$$

and

$$\begin{aligned}\mathbf{E} &= \mathbf{B}\mathbf{R}\mathbf{B}^T \\ \mathbf{H} &= \Delta^T[\mathbf{A}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{A}^T]^{-1}\Delta.\end{aligned}$$

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' Λ* or *Anderson's U* . Since the distribution contains three parameters it is somewhat dif-

ficult to use in practise and we therefore give an approximation to an F-distribution in the following

■

||| Theorem 4.22

Let U be $U(s,r,n-k)$ -distributed and let

$$\begin{aligned} t &= \begin{cases} \frac{1}{\sqrt{\frac{s^2r^2-4}{s^2+r^2-5}}} & s^2 + r^2 = 5 \\ \sqrt{\frac{s^2r^2-4}{s^2+r^2-5}} & s^2 + r^2 \neq 5 \end{cases} \\ v &= \frac{1}{2}(2(n-k) + r - s - 1). \end{aligned}$$

Then

$$F = \frac{1 - U^{\frac{1}{t}}}{U^{\frac{1}{t}}} \cdot \frac{vt + 1 - \frac{1}{2}sr}{sr}$$

is approximately distributed as

$$F(sr, vt + 1 - \frac{1}{2}sr).$$

If either s or r 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	Grade 1 – 9	1 = worst, 9 = best
3: Ability to creep	Grade 1 – 9	1 = no runners, 9 = most runners
4: Activity	Grade 1 – 9	1 = weakest, 9 = strongest
5: Time of blooming	Grade 1 – 9	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

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \sim N(\mu_i, \Sigma),$$

then the unbiased estimate of Σ is

$$\hat{\Sigma} = \begin{bmatrix} 0.85897 & 0.07870 \\ 0.07870 & 0.29444 \end{bmatrix}.$$

The matrix $(\mathbf{x}^T \mathbf{x})^{-1}$ 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

$$D(\hat{\theta}) = \Sigma \otimes (\mathbf{x}^T \mathbf{x})^{-1} = \begin{pmatrix} \sigma_{11}(\mathbf{x}^T \mathbf{x})^{-1} & \sigma_{12}(\mathbf{x}^T \mathbf{x})^{-1} \\ \sigma_{21}(\mathbf{x}^T \mathbf{x})^{-1} & \sigma_{22}(\mathbf{x}^T \mathbf{x})^{-1} \end{pmatrix},$$

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

$$\mathbf{A} \boldsymbol{\theta} \mathbf{B}^T = (\theta_{41} \ \theta_{42}).$$

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

$$\mathbf{A} \boldsymbol{\theta} \mathbf{B}^T = \boldsymbol{\theta}_1.$$

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-way analysis of variance. While it is possible to perform 3-, 4-way or even higher orders, we will not treat them here. They are, however, easily accessible with the SAS procedure [PROC GLM](#). First

4.3.1 One-way multi-dimensional analysis of variance

We consider observations

$$\begin{aligned} Y_{11}, \dots, Y_{1n_1} \\ \vdots \qquad \vdots \\ Y_{k1}, \dots, Y_{kn_k} \end{aligned}.$$

These are assumed to be stochastically independent with

$$\mathbf{Y}_{ij} \sim \mathbf{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \quad i = 1, \dots, k; \quad j = 1, \dots, n_i,$$

i.e. p -dimensional normal distributed with the same variance-covariance matrix. We wish to test hypothesis

$$H_0 : \boldsymbol{\mu}_1 = \dots = \boldsymbol{\mu}_k$$

against

$$H_1 : \exists i, j (\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j).$$

Analogously to the univariate one-way analysis of variance we define sums of squares deviation matrices

$$\begin{aligned} \mathbf{T} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{Y}})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}})^T \\ \mathbf{W} &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)^T \\ \mathbf{B} &= \sum_{i=1}^k n_i (\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})^T \end{aligned}$$

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 $\boldsymbol{\Sigma}$ can use

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-k} \mathbf{W}.$$

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	$\mathbf{B} = \sum_i n_i (\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})^T$	$k - 1$
Error = variation within groups	$\mathbf{W} = \sum_i \sum_j (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)^T$	$n - k$
Total	$\mathbf{T} = \sum_i \sum_j (\mathbf{Y}_{ij} - \bar{\mathbf{Y}})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}})^T$	$n - 1$

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-way multidimensional analysis of variance

In this case we will only look at a two-way analysis of variance with 1 observation per cell. We will therefore assume that we have observations

$$\begin{aligned} & \mathbf{Y}_{11}, \dots, \mathbf{Y}_{1m} \\ & \vdots \quad \vdots , \\ & \mathbf{Y}_{k1}, \dots, \mathbf{Y}_{km} \end{aligned}$$

which are assumed to be p -dimensional normal distributed with the same variance-covariance matrix Σ and with mean values

$$E(\mathbf{Y}_{ij}) = \boldsymbol{\mu}_{ij} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_j,$$

where the parameters $\boldsymbol{\alpha}_i$ $\boldsymbol{\beta}_j$ satisfy

$$\sum_i \boldsymbol{\alpha}_i = \sum_j \boldsymbol{\beta}_j = \mathbf{0}.$$

We now want to test the hypothesis

$$H_0 : \boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_k = \mathbf{0}$$

against

$$H_1 : \exists i (\boldsymbol{\alpha}_i \neq \mathbf{0})$$

and

$$K_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_m = \mathbf{0}$$

against

$$K_1 : \exists j (\boldsymbol{\beta}_j \neq \mathbf{0}).$$

Analogous to the sums of squares of the one-dimensional (univariate) analysis of variance we define the matrices

$$\begin{aligned} \mathbf{T} &= \sum_{i=1}^k \sum_{j=1}^m (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})^T \\ \mathbf{Q}_1 &= \sum_{i=1}^k \sum_{j=1}^m (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{.j} + \bar{\mathbf{Y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{.j} + \bar{\mathbf{Y}}_{..})^T \\ \mathbf{Q}_2 &= m \sum_{i=1}^k (\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})^T \\ \mathbf{Q}_3 &= k \sum_{j=1}^m (\bar{\mathbf{Y}}_{.j} - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_{.j} - \bar{\mathbf{Y}}_{..})^T. \end{aligned}$$

Here we have used the usual notation

$$\begin{aligned} \bar{\mathbf{Y}}_{..} &= \frac{1}{km} \sum_{i=1}^k \sum_{j=1}^m \mathbf{Y}_{ij} \\ \bar{\mathbf{Y}}_{i.} &= \frac{1}{m} \sum_{j=1}^m \mathbf{Y}_{ij}, \quad i = 1, \dots, k \\ \bar{\mathbf{Y}}_{.j} &= \frac{1}{k} \sum_{i=1}^k \mathbf{Y}_{ij}, \quad j = 1, \dots, m. \end{aligned}$$

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

$$\{\mathbf{y}_{11}, \dots, \mathbf{y}_{km} \mid \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_2)} \leq U(p, k-1, (k-1)(m-1))_\alpha\}.$$

The ratio test at level α for test of K_0 against K_1 is given by the critical region

$$\{\mathbf{y}_{11}, \dots, \mathbf{y}_{km} \mid \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1 + \mathbf{q}_3)} \leq U(p, m-1, (k-1)(m-1))_\alpha\}.$$

|||| Proof omitted

Follows readily from theorem 4.21. See e.g. [Anderson \(1958\)](#).

We collect the results in a usual analysis of variance table

Source of variation	SS-matrix	Degrees of freedom	Test statistic
Differences between columns	$\mathbf{Q}_3 = k \sum_j (\bar{\mathbf{Y}}_{\cdot j} - \bar{\mathbf{Y}}_{\cdot \cdot})(\bar{\mathbf{Y}}_{\cdot j} - \bar{\mathbf{Y}}_{\cdot \cdot})^T$	$m - 1$	$\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1 + \mathbf{Q}_3)}$
	$\mathbf{Q}_2 = m \sum_i (\bar{\mathbf{Y}}_{i \cdot} - \bar{\mathbf{Y}}_{\cdot \cdot})(\bar{\mathbf{Y}}_{i \cdot} - \bar{\mathbf{Y}}_{\cdot \cdot})^T$	$k - 1$	$\frac{\det(\mathbf{Q}_1)}{\det(\mathbf{Q}_1 + \mathbf{Q}_2)}$
	$\mathbf{Q}_1 = \sum_i \sum_j (\bar{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i \cdot} - \bar{\mathbf{Y}}_{\cdot j} + \bar{\mathbf{Y}}_{\cdot \cdot}) \times (\bar{\mathbf{Y}}_{ij} - \bar{\mathbf{Y}}_{i \cdot} - \bar{\mathbf{Y}}_{\cdot j} + \bar{\mathbf{Y}}_{\cdot \cdot})^T$	$(k - 1)(m - 1)$	
Total	$\mathbf{T} = \sum_i \sum_j (\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{\cdot \cdot})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{\cdot \cdot})^T$	$km - 1$	

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.

||| Case 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.					
		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

Yield 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 [PROC GLM](#))

Source	DF	Sums of Squares	Mean Square	F-value	Pr > F
Model	7	22.16385172	3.16626453	3.18	0.0482
Error	10	9.97011456	0.99701146		
Corrected Total	17	32.13396628			

SAS provides the above table, with a test of both α and β being equal to zero. We want to investigate the plan and block effect separately and use the Type III SS table

Source	DF	Type III SS	Mean Square	F-value	Pr > F
plant	2	10.94560411	5.47280206	5.49	0.0246
block	5	11.21824761	2.24364952	2.25	0.1288

The test statistic for the hypothesis $\beta_1 = \dots = \beta_6 = 0$ is

$$F = \frac{MS_{block}}{MS_{Error}} = 2.25 < 3.33 = F_{95\%}(5, 10)$$

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

$$F = \frac{MS_{plant}}{MS_{Error}} = 5.49 > 4.10 = F_{95\%}(2, 10).$$

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:

Source	DF	Sums of Squares	Mean Square	F-value	Pr > F
Model	7	0.01883306	0.00269044	3.24	0.0456
Error	10	0.00831056	0.00083106		
Corrected Total	17	0.02714361			

Source	DF	Type III SS	Mean Square	F-value	Pr > F
plant	2	0.00803078	0.00401539	4.83	0.0340
block	5	0.01080228	0.00216046	2.60	0.0932

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

Source	DF	Sums of Squares	Mean Square	F-value	Pr > F
Model	7	521.8764857	74.5537837	3.91	0.0258
Error	10	190.6005383	19.0600538		
Corrected Total	17	712.4770240			

Source	DF	Type III SS	Mean Square	F-value	Pr > F
plant	2	260.1739723	130.0869862	6.83	0.0135
block	5	261.7025133	52.3405027	2.75	0.0817

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

$$\mathbf{Y}_{ij} = \begin{pmatrix} \text{content of green matter} & \text{in plant } i \text{ in blok } j \\ \text{content of nitrogen} & \text{---} \\ \text{content of dry matter} & \text{---} \end{pmatrix}.$$

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. 305 the matrices \mathbf{Q}_1 , \mathbf{Q}_2 and \mathbf{Q}_3 are found to be

$$\begin{aligned} \text{plant: } \mathbf{q}_2 &= \begin{bmatrix} 10.945604111 & 0.2626132778 & 52.369429167 \\ 0.2626132778 & 0.0080307778 & 1.3854275 \\ 52.369429167 & 1.3854275 & 260.17397233 \end{bmatrix} \\ \text{block: } \mathbf{q}_3 &= \begin{bmatrix} 11.218247611 & 0.3480119444 & 53.974648667 \\ 0.3480119444 & 0.0108022778 & 1.6712966667 \\ 53.974648667 & 1.6712966667 & 261.70251333 \end{bmatrix} \\ \text{error: } \mathbf{q}_1 &= \begin{bmatrix} 9.9701145556 & 0.2866667222 & 43.4540855 \\ 0.2866667222 & 0.0083105556 & 1.2551111667 \\ 43.4540855 & 1.2551111667 & 190.60053833 \end{bmatrix} \end{aligned}$$

The matrices have been found by means of [PROC GLM](#), with the statement

```
manova h=_all_ / printe printh;
```

Further, SAS provides the following tables

MANOVA					
Test Criteria and F Approximations for the Hypothesis of No Overall plant Effect					
H = Type III SSCP Matrix for plant					
E = Error SSCP Matrix					
S=2 M=0 N=3					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00332590	43.57	6	16	<.0001
Pillai's Trace	1.85881808	39.5	6	18	<.0001
Hotelling-Lawley Trace	40.44925895	51.86	6	9.0909	<.0001
Roy's Greatest Root	32.46449279	97.39	3	9	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

The (in this case exact) F-test statistic for Wilks' Lambda for a test of the hypothesis

$\alpha_1 = \alpha_2 = \alpha_3 = 0$, (i.e.. the hypothesis that all plants are equal) is 43.57. 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, which we also see directly from the p-value.

MANOVA					
Test Criteria and F Approximations for the Hypothesis of No Overall block Effect					
$H = \text{Type III SSCP Matrix for block}$					
$E = \text{Error SSCP Matrix}$					
$S=3 M=0.5 N=3$					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.06301303	2.58	15	22.486	0.0205
Pillai's Trace	1.45992638	1.90	15	30	0.0663
Hotelling-Lawley Trace	6.63821190	3.27	15	10.704	0.0283
Roy's Greatest Root	5.01196572	10.02	5	10	0.0012
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

We see from Wilks' Lambda, that now also the hypothesis of the blocks being equal is rejected at a level smaller than $\alpha = 2.5\%$.

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.

||| Remark 4.28

In example 4.27 above, we first performed individual ANOVA tests before performing a MANOVA. This was done for illustrative purposes. In real situations, one would first perform a MANOVA. If we find a significant overall effect, we can *then* perform individual ANOVA's to see which variables contribute the most. If there is no significant overall effect - there is no reason to perform tests on the individual variables.

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

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \boldsymbol{\Sigma}_{1k} \\ \vdots & & \vdots \\ \boldsymbol{\Sigma}_{k1} & \cdots & \boldsymbol{\Sigma}_{kk} \end{bmatrix}.$$

Our hypothesis is now that $\mathbf{X}_1, \dots, \mathbf{X}_k$ are independent i.e. that the variance-covariance matrix has the form

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{kk} \end{bmatrix}.$$

If we define $\hat{\boldsymbol{\Sigma}}$ computed on the basis of n realisations of \mathbf{X} in the usual way and if we partition $\hat{\boldsymbol{\Sigma}}$ analogously to the partitioning of $\boldsymbol{\Sigma}$, we have

|||| **Theorem 4.29**

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. [Anderson \(1958\)](#).

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 4.30**

We consider independent observations X_1, \dots, X_n with $X_i \sim N_p(\mu, \Sigma)$, and we let

$$\mathbf{A} = \sum (X_i - \bar{X})(X_i - \bar{X})^T.$$

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. [Anderson \(1958\)](#).

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 4.31**

We consider independent observations X_1, \dots, X_n with $X_i \sim N_p(\mu, \Sigma)$, and we let

$$\mathbf{A} = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T.$$

The quotient test statistic for a test of $H_0 : \Sigma = \Sigma_0$, where Σ_0 is known against all alternatives is

$$\lambda_1 = \left(\frac{e}{n}\right)^{pn/2} [\det(\mathbf{A} \Sigma_0^{-1})]^{\frac{n}{2}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{A} \Sigma_0^{-1})\right),$$

where e is Eulers number. When determining the critical region we can use that

$$P\{-2 \ln \lambda_1 \leq v\} \simeq P\{\chi^2\left(\frac{1}{2}p(p+1)\right) \leq v\}.$$

|||| **Proof omitted**

See e.g. [Anderson \(1958\)](#).

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

$$\begin{aligned} X_{11}, \dots, X_{1n_1}, \quad & X_{1j} \sim N_p(\mu_1, \Sigma_1) \\ \vdots & \\ X_{k1}, \dots, X_{kn_k}, \quad & X_{kj} \sim N_p(\mu_k, \Sigma_k) \end{aligned},$$

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

$$\begin{aligned} n &= \sum n_i, \\ \mathbf{W}_i &= \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)^T, \end{aligned}$$

and

$$\mathbf{W} = \sum_{i=1}^k \mathbf{W}_i,$$

cf. section 4.3.1.

We then have

|||| Theorem 4.32

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. [Anderson \(1958\)](#).

|||| 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

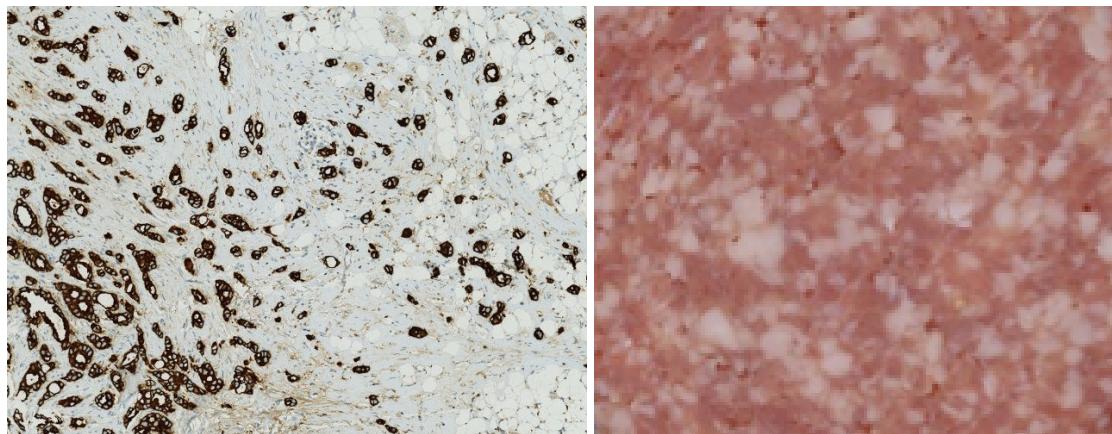


Figure 5.1 – Left. Tissue Micro Array core slice from breast carcinomas stained with PCK marking tumor tissue. The core were digitalized using a high resolution optical scanner [Wessel Lindberg et al. \(2017\)](#). Right. Salami section two days after fermentation start showing fat and meat fractions [Trinderup et al. \(2018b\)](#).

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.

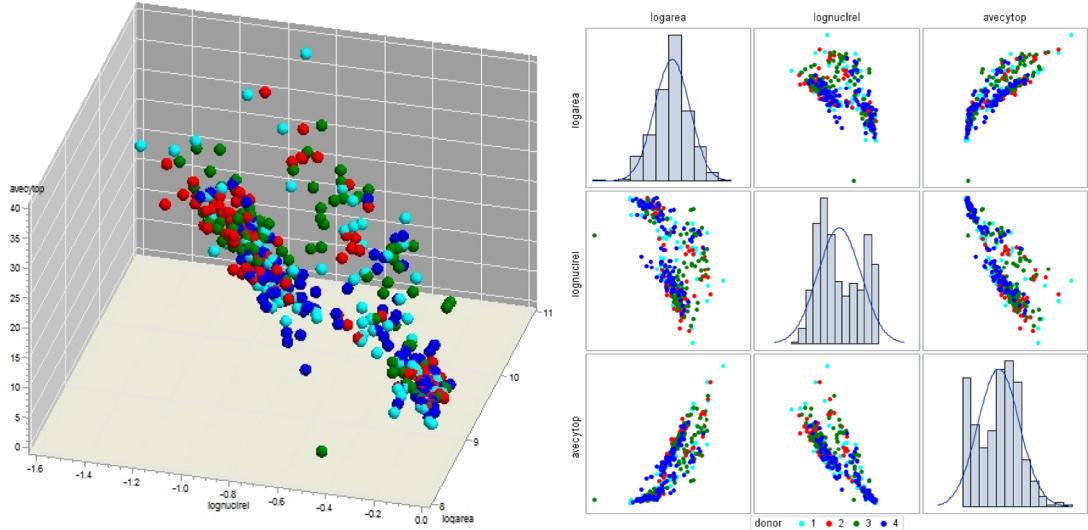


Figure 5.2 – Scatter plots and histograms of cell measurements (log area, log (area nucleus/area), average cytoplasma intensity. The color code corresponds to different individuals [Ottosen \(2015\)](#).

5.1 Discrimination between two populations

5.1.1 Logistic regression

Logistic regression is a probabilistic regression model, that maps the input to probabilities for belonging to a given class. See chapter [3.5.3](#) for more details.

5.1.2 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 :

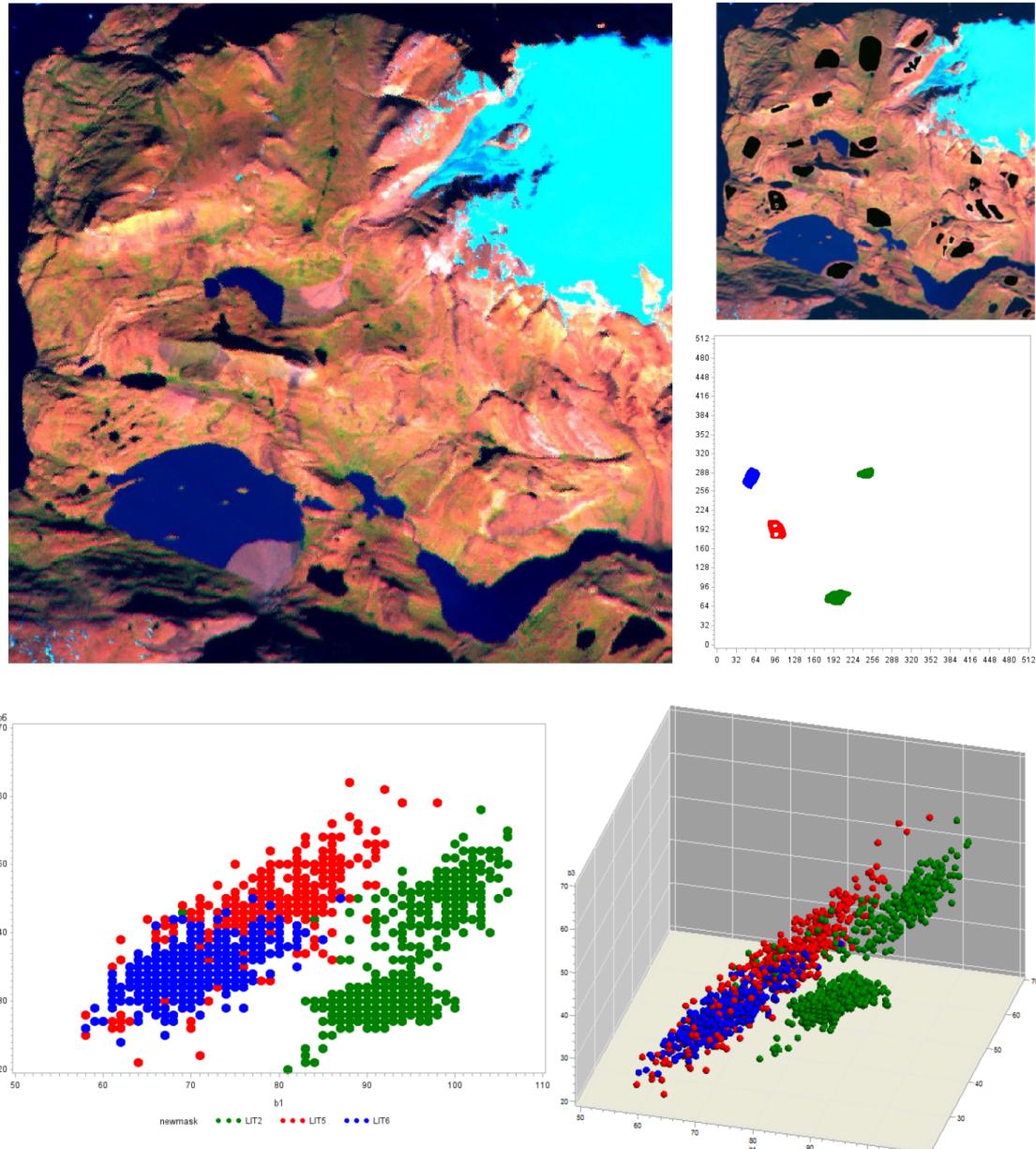


Figure 5.3 – Top. Landsat false color composite (RGB \sim bands 4, 6, 1) of area on Ymer Ø, Central East Greenland. Training areas from different geological units; locality of training areas for unit 2 (deltas and young alluvial fans, green), unit 5 (quartzites, red), and unit 6 (black shales, blue).

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 [Conradsen \(1987\)](#).

From		Classify as	
		π_1	π_2
Nature	π_1	0	$L(\pi_1, \pi_2) = L_{12}$
	π_2	$L(\pi_2, \pi_1) = L_{21}$	0

Table 5.1 – The loss function connected with the classification problem

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 $P\{X \in R_1 | \pi_i\}$ and $P\{X \in R_2 | \pi_i\}$.

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 $X = \mathbf{x}$. 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

$$\mathbf{x} \in R_1 \iff L(\pi_2, \pi_1) k(\pi_2 | \mathbf{x}) \leq L(\pi_1, \pi_2) k(\pi_1 | \mathbf{x})$$

$$\begin{aligned} &\iff \frac{L_{12}p_1f_1(\mathbf{x})}{L_{21}p_2f_2(\mathbf{x})} \geq 1 \\ &\iff \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{L_{21}p_2}{L_{12}p_1} \end{aligned}$$

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.

$$\begin{aligned} R(\pi_1, d_{R_1}) &= E_{\pi_1} L(\pi_1, d_{R_1}(\mathbf{X})) \\ &= L_{11} \times P\{\mathbf{X} \in R_1 \mid \pi_1\} + L_{12} \times P\{\mathbf{X} \in R_2 \mid \pi_1\} = L_{12} \times P\{\mathbf{X} \in R_2 \mid \pi_1\} \end{aligned}$$

$$R(\pi_2, d_{R_1}) = E_{\pi_2} L(\pi_2, d_{R_1}(\mathbf{X}))$$

$$= L_{21} \times P\{\mathbf{X} \in R_1 \mid \pi_2\} + L_{22} \times P\{\mathbf{X} \in R_2 \mid \pi_2\} = L_{21} \times P\{\mathbf{X} \in R_1 \mid \pi_2\}$$

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. See example 5.10 for how to find c in practise.

We will now consider the important special case where f_1 and f_2 are normal distributions.

5.1.3 Discrimination between two normal populations

If f_1 and f_2 are normal with the same dispersion matrix we have

|||| **Theorem 5.4**

Let $\pi_1 \sim N(\mu_1, \Sigma)$ and $\pi_2 \sim N(\mu_2, \Sigma)$. Then we have

$$\begin{aligned} \frac{f_1(x)}{f_2(x)} \geq c &\Leftrightarrow x^T \Sigma^{-1} (\mu_1 - \mu_2) - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 \geq \log c \\ &\Leftrightarrow \left[x^T \Sigma^{-1} \mu_1 - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 \right] - \left[x^T \Sigma^{-1} \mu_2 - \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 \right] \geq \log c. \end{aligned}$$

|||| **Proof**

We introduce the inner product ($|$) and the norm $\|\cdot\|$ by

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

and

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

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

$$\begin{aligned} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq c &\Leftrightarrow \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \log c \\ &\Leftrightarrow -\|\mathbf{x} - \boldsymbol{\mu}_1\|^2 + \|\mathbf{x} - \boldsymbol{\mu}_2\|^2 \geq 2\log c \\ &\Leftrightarrow -(\mathbf{x} - \boldsymbol{\mu}_1 | \mathbf{x} - \boldsymbol{\mu}_1) + (\mathbf{x} - \boldsymbol{\mu}_2 | \mathbf{x} - \boldsymbol{\mu}_2) \geq 2\log c \\ &\Leftrightarrow 2(\mathbf{x} | \boldsymbol{\mu}_1) - 2(\mathbf{x} | \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 | \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_2 | \boldsymbol{\mu}_2) \geq 2\log c \\ &\Leftrightarrow 2(\mathbf{x} | \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - (\boldsymbol{\mu}_1 | \boldsymbol{\mu}_1) + (\boldsymbol{\mu}_2 | \boldsymbol{\mu}_2) \geq 2\log c. \end{aligned}$$

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 $(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2)$, see figure 5.4. 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

$$\mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \log p_i, \quad i = 1, 2$$

are called *linear discriminant functions* for π_i . If we do not have a prior distribution, the term $\log p_i$ is omitted. The function

$$\mathbf{x}^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 - \log c$$

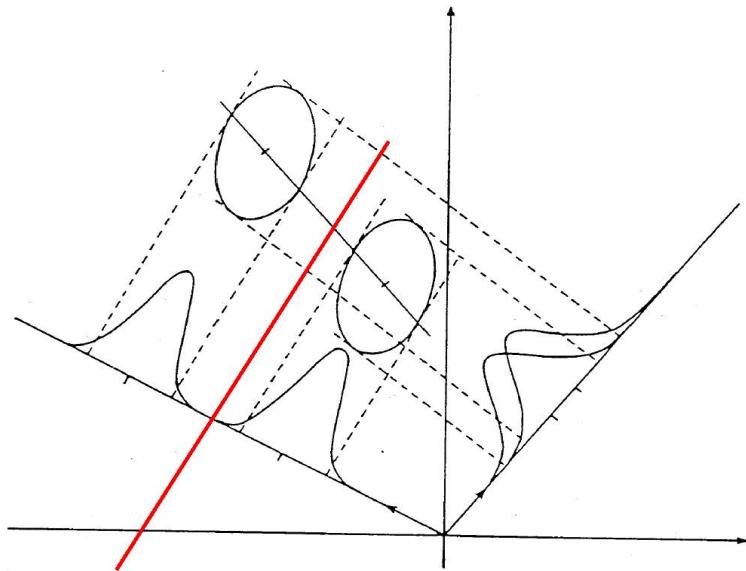


Figure 5.4 – Classification example

is called the *linear discriminator* between π_1 and π_2 .

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

$$g(d) = \frac{[E_1(X^T d) - E_2(X^T d)]^2}{V(X^T d)} = \frac{[(\mu_1 - \mu_2)^T d]^2}{d^T \Sigma d}$$

|||| **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

$$\mathbf{d}^T \boldsymbol{\Sigma} \mathbf{d} = 1$$

We introduce a Lagrange multiplier λ and seek the maximum of

$$\psi(\mathbf{d}) = [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{d}]^2 - \lambda (\mathbf{d}^T \boldsymbol{\Sigma} \mathbf{d} - 1) .$$

Now we have that

$$\frac{\partial \psi}{\partial \mathbf{d}} = 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{d} - 2\lambda \boldsymbol{\Sigma} \mathbf{d}$$

If we let this equal 0, we have

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{d} = \lambda \boldsymbol{\Sigma} \mathbf{d}$$

i.e.

$$\mathbf{d} = \left\{ \frac{1}{\lambda} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{d} \right\} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = k \boldsymbol{\delta}$$

where k is a scalar.

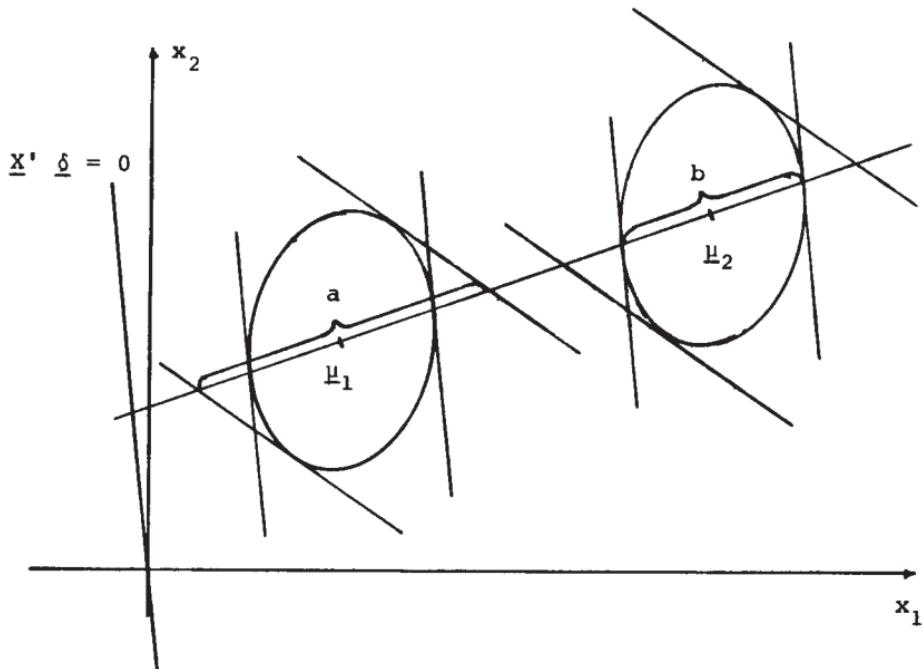
■

|||| **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 $-\log c$), 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 (theorem 5.4)

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} - \frac{1}{2}(2 \cdot 16 + 1 \cdot 4 - 2 \cdot 8) + \frac{1}{2}(2 \cdot 1 + 1 \cdot 1 - 2 \cdot 1) = 0$$

or

$$5x_1 - 2x_2 - 9.5 = 0$$

If we enter an arbitrary point, e.g. $\begin{bmatrix} 5 \\ 6 \end{bmatrix}$ we get

$$5 \cdot 5 - 2 \cdot 6 - 9.5 = 3.5 > 0 .$$

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:

		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

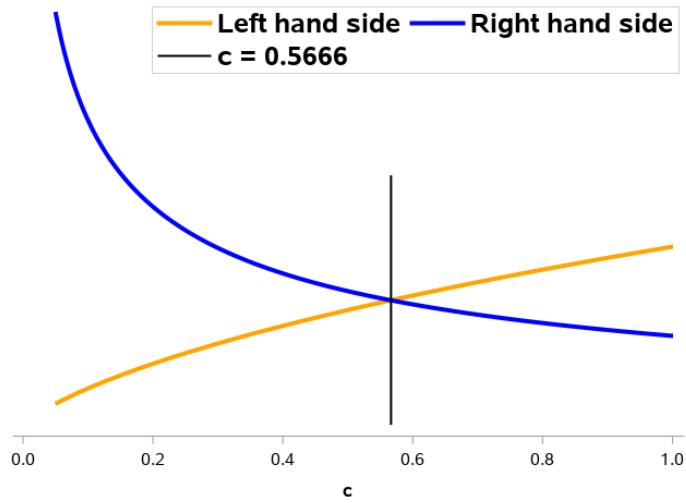
$$2P\left\{\frac{f_1(x)}{f_2(x)} < c \mid \pi_1\right\} = P\left\{\frac{f_1(x)}{f_2(x)} \geq c \mid \pi_2\right\}$$

Using theorem 5.8

$$\Leftrightarrow 2P\{Z < \log c \mid \pi_1\} = P\{Z \geq \log c \mid \pi_2\}$$

$$\begin{aligned} &\Leftrightarrow 2P\left\{N\left(\frac{1}{2} \cdot 13, 13\right) < \log c\right\} = P\left\{N\left(-\frac{1}{2} \cdot 13, 13\right) \geq \log c\right\} \\ &\Leftrightarrow 2P\left\{N(0, 1) < \frac{\log c - 6.5}{\sqrt{13}}\right\} = P\left\{N(0, 1) \geq \frac{\log c + 6.5}{\sqrt{13}}\right\} \end{aligned}$$

For determining c we first note that the left hand side of the equality sign is increasing as a function of c , and that the right hand side is decreasing. Inserting $c = 1$ gives the values 0.072 and 0.036 for the left and right hand side of the equation. Thus, the correct value of c must be smaller than 1. If we try $c = 0.5$ we get the values 0.046 and 0.054, and we conclude that the correct c is larger than 0.5. After a few iterations along those lines, we will arrive at a correct estimate. The value for c can of course also be found directly, e.g. by a graphical approach,



where we find the following value

$$c \simeq 0.5666 .$$

Using this value, the misclassification probabilities are

$$\text{If } \pi_1 \text{ is true : } P\left\{N(0, 1) < \frac{\log 0.5666 - 6.5}{\sqrt{13}}\right\} \simeq 0.025$$

$$\text{If } \pi_2 \text{ is true : } P\left\{N(0, 1) \geq \frac{\log 0.5666 + 6.5}{\sqrt{13}}\right\} \simeq 0.050$$

The discriminating line is now determined by

$$5x_1 - 2x_2 - 9.5 = \log 0.5666$$

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

$$\frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow -(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) + (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) \geq 2\log c$$

Since

$$\begin{aligned} (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) &= 2(x_1 - 4)^2 + (x_2 - 2)^2 - 2(x_1 - 4)(x_2 - 2) \\ &= 2x_1^2 + x_2^2 - 2x_1x_2 - 12x_1 + 4x_2 + 20, \end{aligned}$$

and

$$\begin{aligned} (x - \mu_2)^T \Sigma_2^{-1} (x - \mu_2) &= (x_1 - 1)^2 + (x_2 - 1)^2 \\ &= x_1^2 + x_2^2 - 2x_1 - 2x_2 + 2, \end{aligned}$$

then

$$\frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow -x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 \geq 2\log c$$

If we choose $c = 1$, we note that the curve which separates R_1 and R_2 is the hyperbola

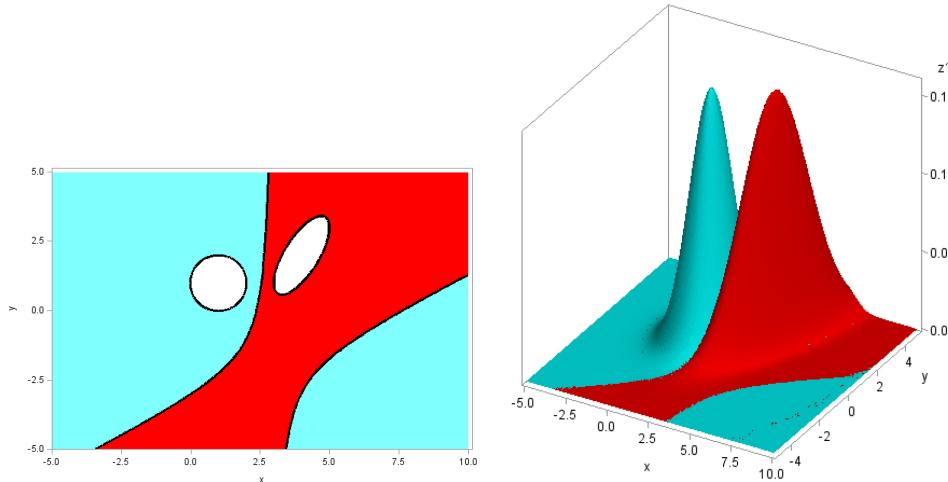
$$\{x \mid -x_1^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 = 0\}$$

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 $(1, 1)^T$. Furthermore 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.

5.1.4 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. Then we may construct discriminators from the estimated distributions the same way we did for the exact distributions.

Let us consider the normal case

$$\pi_1 \leftrightarrow N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$$

$$\pi_2 \leftrightarrow N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$$

where the parameters are unknown. If we have observations $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1}$ which we know come from π_1 and observations $\mathbf{X}_{21}, \dots, \mathbf{X}_{2n_2}$ which we know come from π_2 we can estimate the parameters as usual:

$$\hat{\boldsymbol{\mu}}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} \mathbf{X}_{1j} = \bar{\mathbf{X}}_1 \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\mathbf{X}_{1j} - \bar{\mathbf{X}}_1)(\mathbf{X}_{1j} - \bar{\mathbf{X}}_1)^T$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \mathbf{X}_{2j} = \bar{\mathbf{X}}_2 \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}_2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (\mathbf{X}_{2j} - \bar{\mathbf{X}}_2)(\mathbf{X}_{2j} - \bar{\mathbf{X}}_2)^T$$

and with $N = n_1 + n_2$ the *pooled estimate* of the dispersion matrix

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-2} \left[(n_1 - 1)\hat{\boldsymbol{\Sigma}}_1 + (n_2 - 1)\hat{\boldsymbol{\Sigma}}_2 \right]$$

We now estimate the appropriate decision rule by plugging these estimators into the formula given by theorem 5.4, i.e.

$$\mathbf{x}^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2) - \frac{1}{2} \hat{\boldsymbol{\mu}}_1^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_1 + \frac{1}{2} \hat{\boldsymbol{\mu}}_2^T \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_2 .$$

The exact distribution of this quantity if we substitute \mathbf{x} with a random variable $\mathbf{X} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ 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

$$D^2 = \| \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2 \|_{\hat{\boldsymbol{\Sigma}}^{-1}}^2 = (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$$

This is called the *empirical Mahalanobis' distance* as opposed to the (*theoretical*) *Mahalanobis' distance*

$$\Delta^2 = \| \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \|_{\boldsymbol{\Sigma}^{-1}}^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

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 $\mu_1 = \mu_2$ against all alternatives becomes

$$C = \left\{ x_{11}, \dots, x_{1n_1}, x_{21}, \dots, x_{2n_2} \mid \frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \cdot \frac{n_1 n_2}{n_1+n_2} d^2 > F(p, n_1 + n_2 - p - 1)_{1-\alpha} \right\}$$

Here d^2 is the observed value of D^2 .

|||| **Proof**

An immediate consequence of the relation to Hotellings T^2 statistic.

■

5.1.5 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 omitted**

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 X is $f_i(x)$.

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:

$$\text{we select } \pi_i \text{ rather than } \pi_j \text{ if } S_i^* > S_j^*$$

$$\Leftrightarrow -(\sum_v p_v f_v(\mathbf{x}) - p_i f_i(\mathbf{x})) > -(\sum_v p_v f_v(\mathbf{x}) - p_j f_j(\mathbf{x}))$$

$$\Leftrightarrow p_i f_i(\mathbf{x}) > p_j f_j(\mathbf{x}).$$

In this case we can therefore choose the discriminant score

$$S_i = S_i(\mathbf{x}) = p_i f_i(\mathbf{x})$$

In this case the Bayes' rule is that we select the population which has the largest posterior distribution, i.e.

$$\text{select } \operatorname{argmax}_i k(\pi_i | \mathbf{x})$$

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

$$S'_i = S'_i(\mathbf{x}) = f_i(\mathbf{x})$$

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

$$f_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \boldsymbol{\Sigma}_i}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right),$$

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)

$$S_i^Q(\mathbf{x}) = -\frac{1}{2} \log \det \boldsymbol{\Sigma}_i - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \log p_i$$

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

$$k^Q(\pi_v \mid \mathbf{x}) = \frac{\exp(S_v^Q(\mathbf{x}))}{\sum_{i=1}^k \exp(S_i^Q(\mathbf{x}))}.$$

If all the Σ_i are equal then the terms

$$-\frac{1}{2} \log \det \Sigma - \frac{1}{2} \mathbf{x}^T \Sigma^{-1} \mathbf{x}$$

are common for all $S_i^Q(\mathbf{x})$ and can therefore be omitted. We then get

$$S_i^L(\mathbf{x}) = \mathbf{x}^T \Sigma^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \boldsymbol{\mu}_i^T \Sigma^{-1} \boldsymbol{\mu}_i + \log p_i$$

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 (LDF)*. Like in the quadratic case we have the posterior probability

$$k^L(\pi_v \mid \mathbf{x}) = \frac{\exp(S_v^L(\mathbf{x}))}{\sum_{i=1}^k \exp(S_i^L(\mathbf{x}))}.$$

If there are only two groups we note that we choose group 1 if

$$S_1^L(\mathbf{x}) - S_2^L(\mathbf{x}) \geq 0 \Leftrightarrow \mathbf{x}^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2} \boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 > \log \frac{p_2}{p_1}$$

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

||| Example 5.14

We consider populations π_1, π_2 and π_3 given by normal distributions with expected values

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu}_3 = \begin{bmatrix} 2 \\ 6 \end{bmatrix},$$

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

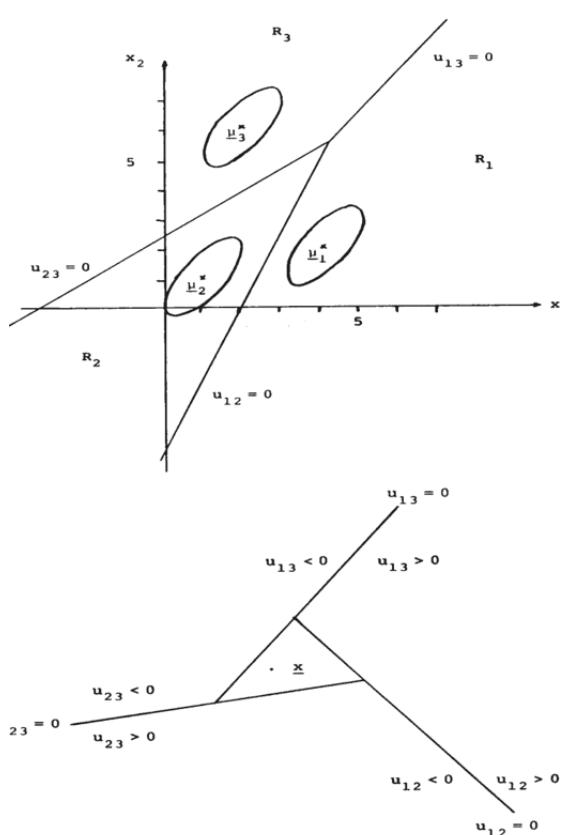
$$R_1 = \{\mathbf{x} \mid u_{12}(\mathbf{x}) > 0 \wedge u_{13}(\mathbf{x}) > 0\}$$

$$R_2 = \{\mathbf{x} \mid u_{12}(\mathbf{x}) < 0 \wedge u_{23}(\mathbf{x}) > 0\}$$

$$R_3 = \{\mathbf{x} \mid u_{13}(\mathbf{x}) < 0 \wedge u_{23}(\mathbf{x}) < 0\}$$

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

$$\left. \begin{array}{l} u_{23}(x) < 0 \quad \text{i.e. } f_2(x) < f_3(x) \\ u_{13}(x) > 0 \quad \text{i.e. } f_1(x) > f_3(x) \end{array} \right\} \Rightarrow f_1(x) > f_2(x)$$

$$u_{12}(x) < 0 \quad \text{i.e. } f_1(x) < f_2(x)$$

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 $N = \sum n_i$.

|||| **Definition 5.15**

Assuming that the hypothesis $H_0 : \Sigma_1 = \dots = \Sigma_k$ is true, we define *the squared generalized distance from $\hat{\mu}_j$ to population π_i* as

$$D_i^2(\hat{\mu}_j) = \begin{cases} (\hat{\mu}_i - \hat{\mu}_j)^T \hat{\Sigma}^{-1} (\hat{\mu}_i - \hat{\mu}_j) & \text{if the priors are equal} \\ (\hat{\mu}_i - \hat{\mu}_j)^T \hat{\Sigma}^{-1} (\hat{\mu}_i - \hat{\mu}_j) - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

If the hypothesis is *not* true, we define *the squared generalized distance from $\hat{\mu}_j$ to population π_i* as

$$D_i^2(\hat{\mu}_j) = \begin{cases} (\hat{\mu}_i - \hat{\mu}_j)^T \hat{\Sigma}_i^{-1} (\hat{\mu}_i - \hat{\mu}_j) & \text{if the priors are equal} \\ (\hat{\mu}_i - \hat{\mu}_j)^T \hat{\Sigma}_i^{-1} (\hat{\mu}_i - \hat{\mu}_j) + \log \det \hat{\Sigma}_i - 2\log p_i & \text{if the priors are not all equal} \end{cases}$$

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 4.1, 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 5.1.4.

|||| **Remark 5.16**

If we look at the most general expression of the squared distance from an observation \mathbf{x} to population π_i we get (replacing $\hat{\boldsymbol{\mu}}_j$ with \mathbf{x})

$$D_i^2(\mathbf{x}) = (\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^T \hat{\boldsymbol{\Sigma}}_i^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}_i) + \log \det \hat{\boldsymbol{\Sigma}}_i - 2\log p_i$$

which is the estimate of minus two times the formula for $S_i^Q(\mathbf{x})$ given earlier. Thus the discriminant score becomes

$$-\frac{1}{2} D_i^2(\mathbf{x})$$

and the estimated posterior probability for group i becomes

$$\hat{k}^Q(\pi_i | \mathbf{x}) = \frac{\exp(-\frac{1}{2} D_i^2(\mathbf{x}))}{\sum_{j=1}^k \exp(-\frac{1}{2} D_j^2(\mathbf{x}))}$$

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 $r^2 \hat{\boldsymbol{\Sigma}}_i$, i.e.

$$K_i(\mathbf{z}) = \frac{1}{r^p (2\pi)^{p/2} \det \hat{\boldsymbol{\Sigma}}_i^{1/2}} \exp \left(-\frac{1}{2r^p} \mathbf{z}^T \hat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{z} \right)$$

We then estimate the density function for group π_i as

$$\hat{f}_i(\mathbf{x}) = \frac{1}{n_i} \sum_{j=1}^{n_i} K_i(\mathbf{x} - \mathbf{x}_{ij})$$

Un 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 [PROC 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 $v(r_k(x))$ is the volume of the of ellipsoid $E_{r_k(x)}(x)$.

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 $E_{r_k(x)}(x)$ around 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 5.17

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 then classify using a simple majority vote principle.

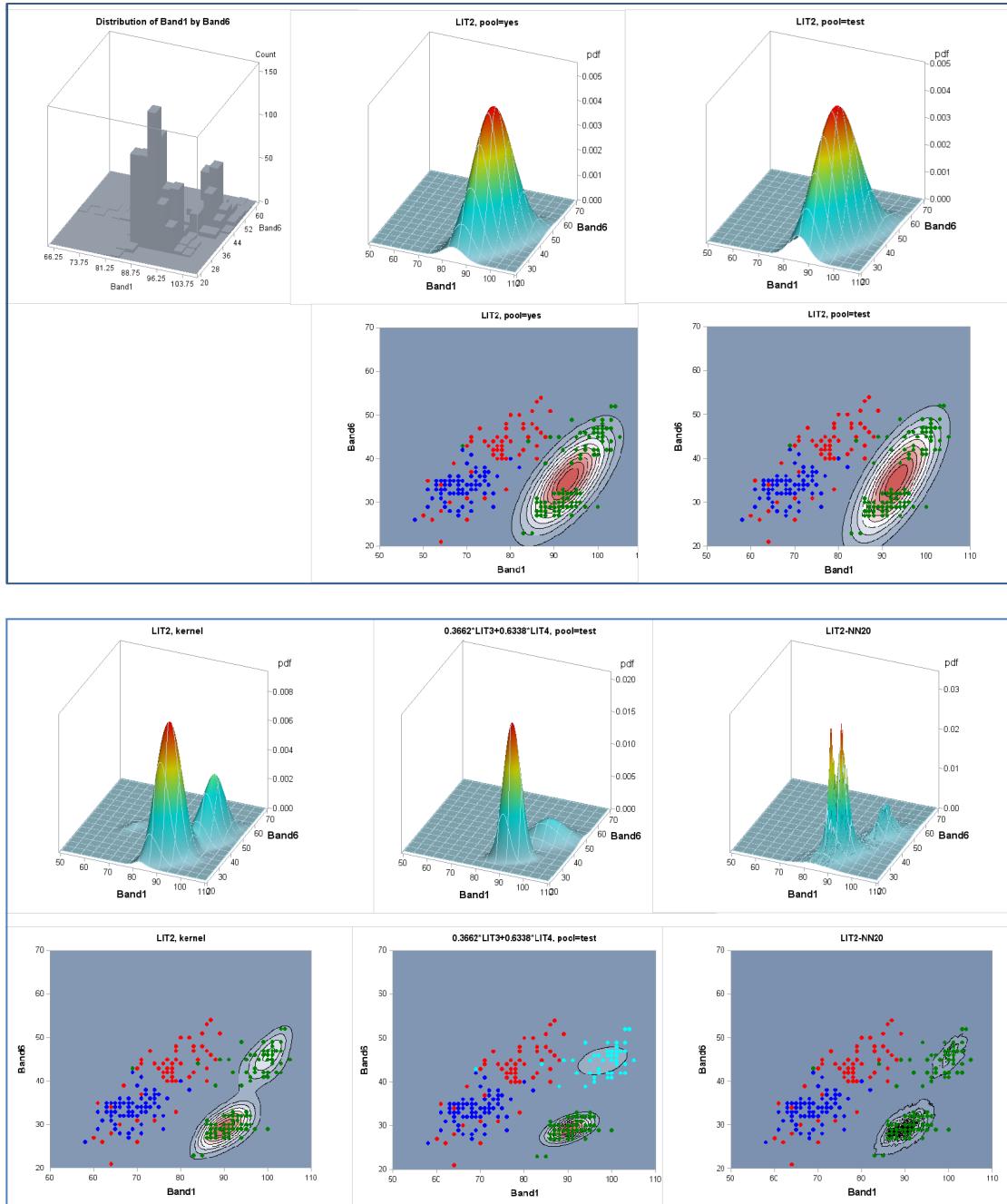


Figure 5.5 – Histogram and estimated distributions for unit 2, deltas and young alluvial fans using Landsat Band1 and band6. In the top frame the distribution is estimated as a single normal distribution, in the first case with the pooled dispersion matrix, in the second with the within class 2 estimated 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.

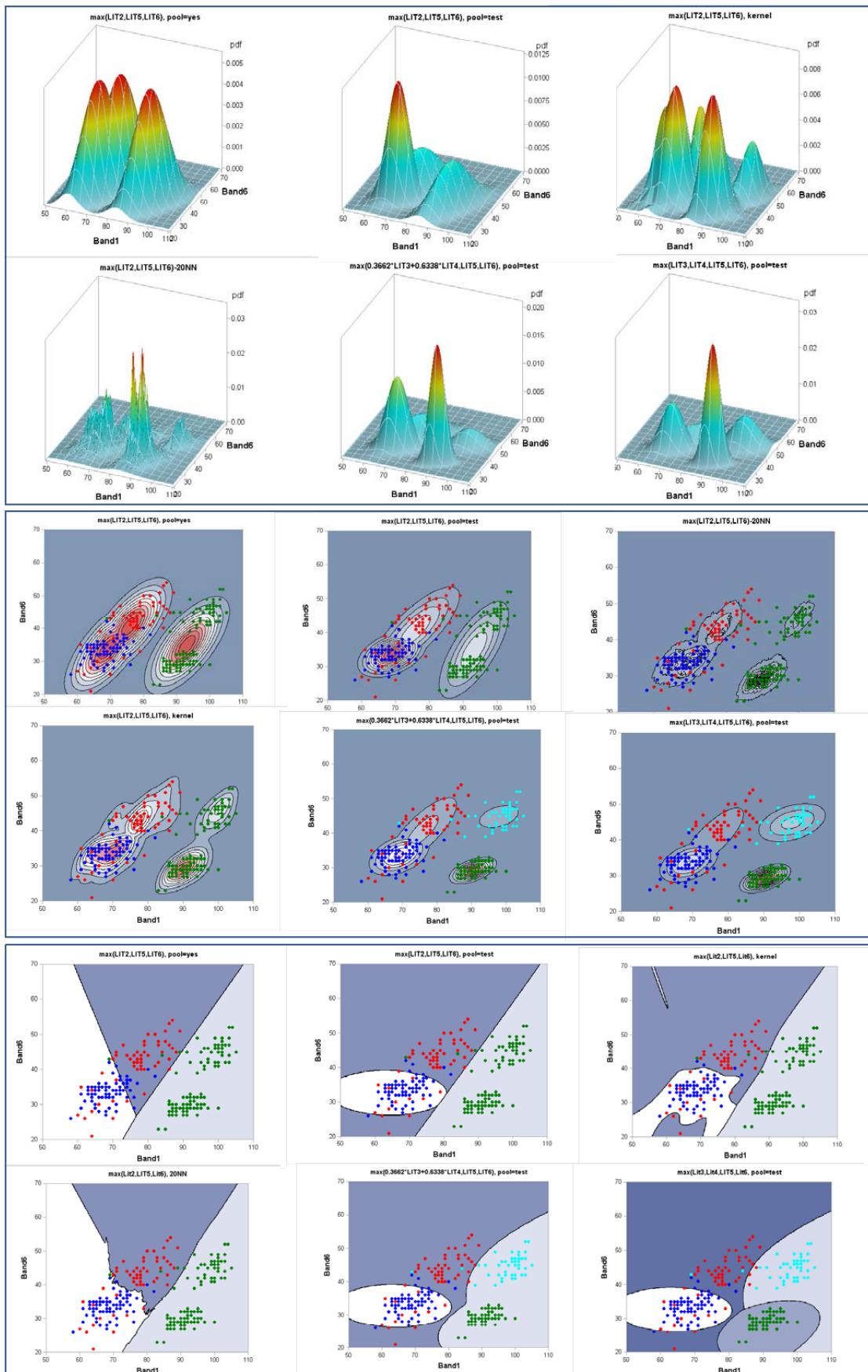


Figure 5.6 – The classification regions corresponding to the different types of estimate of the pdf's.

	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 5.2. 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.

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

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 [PROC 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.

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.

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 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 [PROC 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 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 $H_1 : \text{the class is positive}(\pi_1)$.

Number of observations from	Classified as		Sum
	<i>pos</i>	<i>neg</i>	
Nature	<i>pos</i>	$tp = N_{11}$ = # true positive	$fn = N_{12}$ = # false negative
	<i>neg</i>	$fp = N_{21}$ = # false positive	$tn = N_{22}$ = # true negative
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.

Selecting the positive class will therefore correspond to rejecting the hypothesis. 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	$tp_i = N_{ii}$ = # 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	$\frac{1}{k} \sum_{i=1}^k \frac{tp_i + tn_i}{TN} = \frac{2}{kN_{..}} \sum_{i=1}^k N_{ii} + \frac{(k-2)}{k} = 1 - \frac{2}{k} \frac{1}{N_{..}} \sum_{i \neq j} N_{ij}$
Average class error rate, misclassification rate	$\frac{1}{k} \sum_{i=1}^k \frac{fp_i + fn_i}{TN} = \frac{2}{k} - \frac{2}{k} \frac{1}{N_{..}} \sum_{i=1}^k N_{ii} = \frac{2}{k} \frac{1}{N_{..}} \sum_{i \neq j} N_{ij}$
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

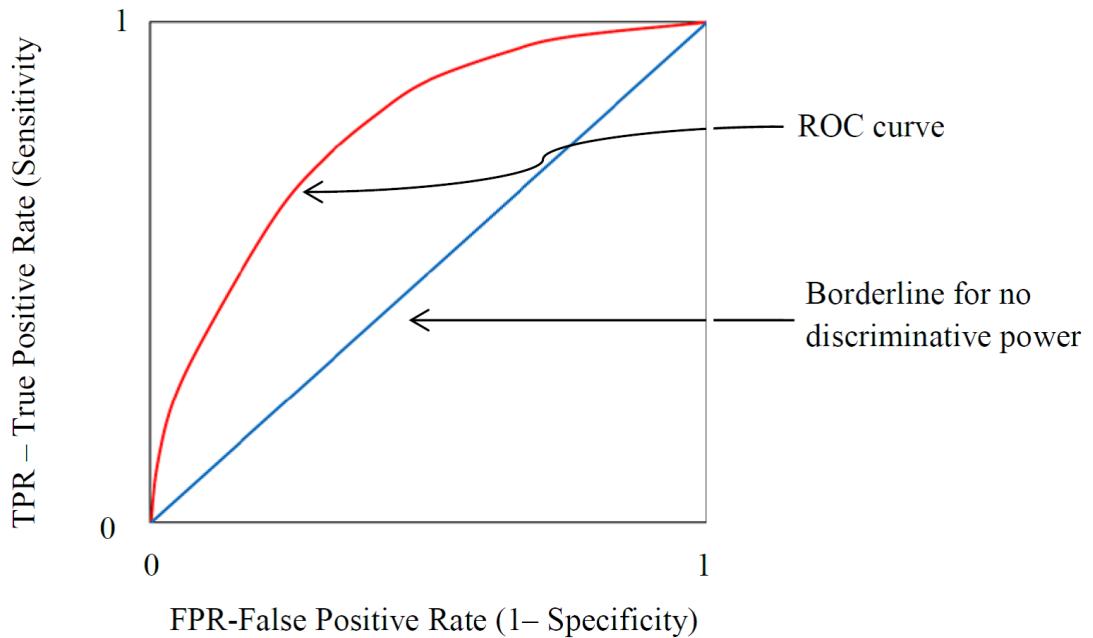


Figure 5.7 – Reciever Operator Curve (ROC)

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 and Mohri \(2005\)](#). 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 5.9 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.

|||| **Theorem 5.18**

We may now test whether the two classifiers perform equally well by using *McNemar's test statistic*

$$M = \frac{(|N_{fs} - N_{fs}| - 1)^2}{N_{fs} + N_{fs}}.$$

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 .

|||| **Proof omitted**

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 \mathbf{W} , \mathbf{B} , and \mathbf{T} as in section 5.2.3.

With $N = \sum n_i$ we have the following estimates of means and the dispersion matrix:

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

$$\widehat{\Sigma} = \frac{1}{N-k} W$$

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:

$$X_{ij} = \begin{bmatrix} X_{ij,1} \\ \vdots \\ X_{ij,q} \\ X_{ij,q+1} \\ \vdots \\ X_{ij,p} \end{bmatrix} = \begin{bmatrix} X_{ij}^{(1)} \\ X_{ij}^{(2)} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_i^{(1)} \\ \mu_i^{(2)} \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right).$$

Thus $X_{ij}^{(1)}$ corresponds to the first q coordinates in X_{ij} and $X_{ij}^{(2)}$ 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 $X_{ij}^{(2)}$ given $X_{ij}^{(1)}$ are

$$\begin{aligned} E\left(X_{ij}^{(2)} \mid X_{ij}^{(1)} = \mathbf{x}^{(1)}\right) &= \boldsymbol{\mu}_i^{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\left(\mathbf{x}^{(1)} - \boldsymbol{\mu}_i^{(1)}\right) \\ &= \boldsymbol{\mu}_i^{(2)} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_i^{(1)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}^{(1)} \\ &= \boldsymbol{\mu}_i^{(2|1)} \end{aligned}$$

$$D\left(X_{ij}^{(2)} \mid X_{ij}^{(1)} = \mathbf{x}^{(1)}\right) = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{2|1}$$

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 $\boldsymbol{\Sigma}_{2|1}$ is the Schur complement $\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{11}$ of $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Sigma}_{11}$ and therefore

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}| \left| \boldsymbol{\Sigma}_{2|1} \right|$$

We furthermore introduce the same partitioning of W , B , and T as we have for $\boldsymbol{\Sigma}$:

$$\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}$$

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 5.19

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 et al. \(1973\)](#) or [McLachlan \(2004\)](#). 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 5.20

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

$$\begin{aligned} \Delta^2 &= \begin{bmatrix} (\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)})^T & (\boldsymbol{\mu}_1^{(2)} - \boldsymbol{\mu}_2^{(2)})^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)} \\ \boldsymbol{\mu}_1^{(2)} - \boldsymbol{\mu}_2^{(2)} \end{bmatrix} \\ \Delta_1^2 &= (\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)})^T \boldsymbol{\Sigma}_{11}^{-1} (\boldsymbol{\mu}_1^{(1)} - \boldsymbol{\mu}_2^{(1)}). \end{aligned}$$

Since

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} + \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} & -\boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \\ -\boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{2|1}^{-1} & \boldsymbol{\Sigma}_{2|1}^{-1} \end{bmatrix}$$

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 5.1.4, we get the corresponding empirical measures

$$\begin{aligned} D_{(2|1)}^2 &= (\widehat{\boldsymbol{\mu}}_1^{(2|1)} - \widehat{\boldsymbol{\mu}}_2^{(2|1)})^T \widehat{\boldsymbol{\Sigma}}_{2|1}^{-1} (\widehat{\boldsymbol{\mu}}_1^{(2|1)} - \widehat{\boldsymbol{\mu}}_2^{(2|1)}) \\ D^2 &= (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)^T \widehat{\boldsymbol{\Sigma}}^{-1} (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) \\ D_1^2 &= (\widehat{\boldsymbol{\mu}}_1^{(1)} - \widehat{\boldsymbol{\mu}}_2^{(1)})^T \widehat{\boldsymbol{\Sigma}}_{11}^{-1} (\widehat{\boldsymbol{\mu}}_1^{(1)} - \widehat{\boldsymbol{\mu}}_2^{(1)}) \end{aligned}$$

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

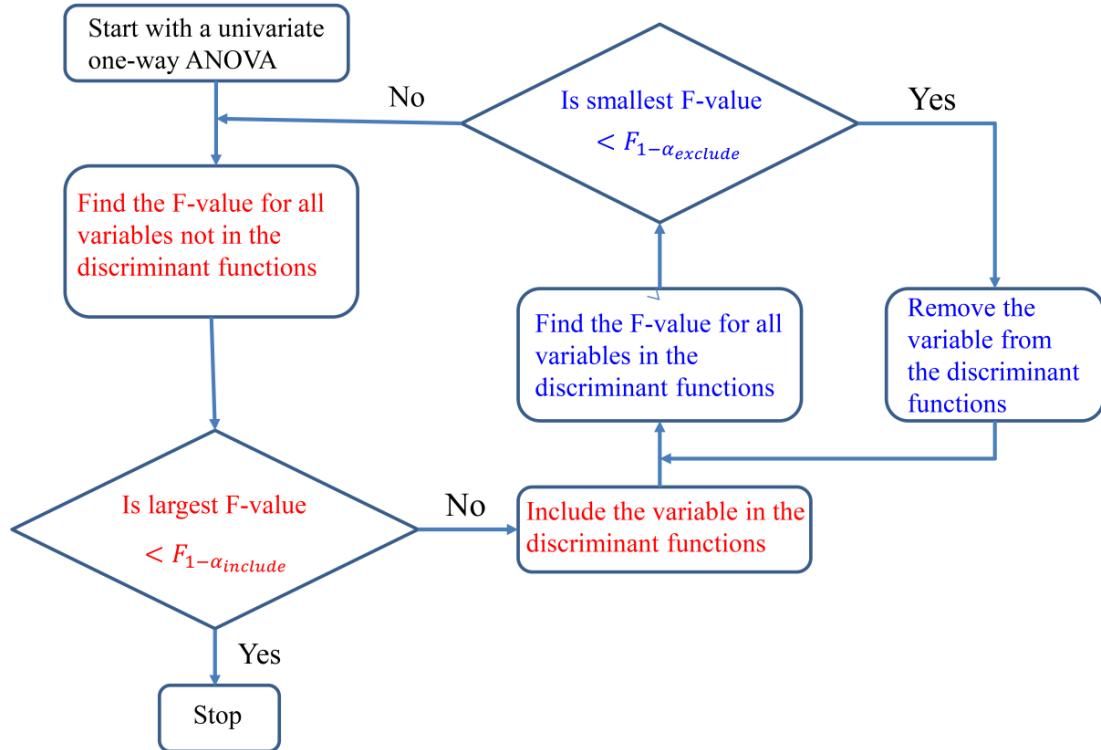


Figure 5.8 – Flow diagram for a stepwise discriminant analysis procedure. The red part basically corresponds to a forward selection method and the blue part to a backward elimination method. The F-values are computed as shown in theorem 5.21

||| Theorem 5.21

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 et al. (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

$$\frac{1}{N-1} \mathbf{T} = \frac{1}{N-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}})(\mathbf{X}_{ij} - \bar{\mathbf{X}})^T$$

as an “estimated” dispersion matrix. The eigenvalues and eigenvectors of this matrix are collected in two matrices

$$\boldsymbol{\Lambda}_T = \text{diag} [\lambda_{T1} \ \cdots \ \lambda_{Tp}]$$

$$\mathbf{P}_T = [\mathbf{p}_{T1} \ \cdots \ \mathbf{p}_{Tp}]$$

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

$$\mathbf{Y}_{ij} = \begin{bmatrix} Y_{ij1} \\ \vdots \\ Y_{ijm} \end{bmatrix} = \begin{bmatrix} (\mathbf{p}_{T1})^T \\ \vdots \\ (\mathbf{p}_{Tm})^T \end{bmatrix} \mathbf{X}_{ij}$$

The analysis may now be performed on the \mathbf{Y}_{ij} instead of the \mathbf{X}_{ij} .

See section 6.1 for more details regarding PCA.

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 $H_0 : \Sigma_1 = \dots = \Sigma_k$ 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 $y = d^T x$ so

$$\varphi(d) = \frac{d^T B d}{d^T W d}$$

is maximized. We note from theorem A.48 that the maximum value is the largest eigenvalue λ_1 and the corresponding eigenvector d_1 to

$$\det(B - \lambda W) = 0$$

or

$$\det(W^{-1}B - \lambda I) = 0$$

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). Since $\text{rk}(B) \leq k - 1$ we have at most $k - 1$ eigenvalues ≥ 0 .

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 (d1; d2)-plane.

The functions $d_1^T x$ are called ***Canonical Discriminant Functions (CDF)*** and the type of analysis a ***Canonical Discriminant Analysis (CDA)***.

We will illustrate the method in the following example.

||| Example 5.22

We use the famous Fisher data of Iris flower measurements. The complete data set is included in most statistical software including SAS. It has $n = 150$ observations, 4 variables: *Sepal length*, *Sepal width*, *Petal length*, and *Petal width*; and 3 groups: *Setosa*, *Versicolor*, and *Virginica*. We will assume that groups have the same dispersion. An example from the data set is given below

Observation	Species	Sepal length	Sepal width	Petal length	Petal width
1	Setosa	5.1	3.5	1.4	0.2
2	Setosa	4.9	3	1.4	0.20
:	:	:	:	:	:
51	Versicolor	7	3.2	4.7	1.4
52	Versicolor	6.4	3.2	4.5	1.5
:	:	:	:	:	:
101	Virginica	6.3	3.3	6	2.5
102	Virginica	5.8	2.7	5.1	1.9

The means and dispersion matrix are

Species \ Variable	Means of the 3 species and overall			
	Sepal length	Sepal width	Petal length	Petal width
Setosa	5.0060	3.4280	1.4620	0.2460
Versicolor	5.9360	2.7700	4.2600	1.3260
Virginica	6.5880	2.9740	5.5520	2.0260
Overall	5.8433	3.0573	3.7580	1.1993

$$\Sigma = \begin{bmatrix} 0.6857 & -0.0424 & 1.2743 & 0.5163 \\ -0.0424 & 0.1900 & -0.3297 & -0.1216 \\ 1.2743 & -0.3297 & 3.1163 & 1.2956 \\ 0.5163 & -0.1216 & 1.2956 & 0.5810 \end{bmatrix}$$

We further need the within class matrix

$$W = \begin{bmatrix} 102.17 & -6.32 & 189.87 & 76.92 \\ -6.323 & 28.31 & -49.12 & -18.12 \\ 189.87 & -49.12 & 464.33 & 193.05 \\ 76.92 & -18.12 & 193.05 & 86.57 \end{bmatrix},$$

and the between class matrix

$$\mathbf{B} = \begin{bmatrix} 63.21 & -19.95 & 165.25 & 71.28 \\ -19.95 & 11.34 & -57.24 & -22.93 \\ 165.25 & -57.24 & 437.10 & 186.77 \\ 71.28 & -22.93 & 186.77 & 80.41 \end{bmatrix}.$$

We then calculate the eigenvalues and corresponding eigenvectors

$$\det(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}) = 0$$

$$\det \begin{bmatrix} -3.0584 - \lambda & 1.0814 & -8.1119 & -3.4586 \\ -5.5616 & 2.1782 - \lambda & -14.9646 & -6.3077 \\ 8.0774 & -2.9427 & 21.5116 - \lambda & 9.1421 \\ 10.4971 & -3.4199 & 27.5485 & 11.8459 - \lambda \end{bmatrix} = 0.$$

Since we have 3 groups, only two eigenvalues are larger than zero. These eigenvalues and corresponding vectors are:

$$\lambda_1 = 32.20, \mathbf{d}_1^* = \begin{bmatrix} -0.2087 \\ -0.3862 \\ 0.5540 \\ 0.7074 \end{bmatrix} \quad \lambda_2 = 0.2854, \mathbf{d}_2^* = \begin{bmatrix} -0.0065 \\ -0.5866 \\ 0.2526 \\ -0.7695 \end{bmatrix}.$$

We then need to scale the eigenvectors using \mathbf{W}

$$\mathbf{d}_1^T \mathbf{W} \mathbf{d}_1 = 1 \Rightarrow \mathbf{d}_1 = \begin{bmatrix} -0.0684 \\ -0.1266 \\ 0.1815 \\ 0.2318 \end{bmatrix},$$

and for the second eigenvector

$$\mathbf{d}_2^T \mathbf{W} \mathbf{d}_2 = 1 \Rightarrow \mathbf{d}_2 = \begin{bmatrix} 0.0020 \\ 0.1785 \\ -0.0769 \\ 0.2341 \end{bmatrix}.$$

We can now transform the observations and plot them in the new canonical coordinate system by subtracting the overall mean, and multiplying with the eigenvectors

$$[\mathbf{d}_1^T(\mathbf{x}_{ij} - \bar{\mathbf{x}}), \mathbf{d}_2^T(\mathbf{x}_{ij} - \bar{\mathbf{x}})].$$

We show the procedure for the first observation. The value for the first canonical discriminant function becomes

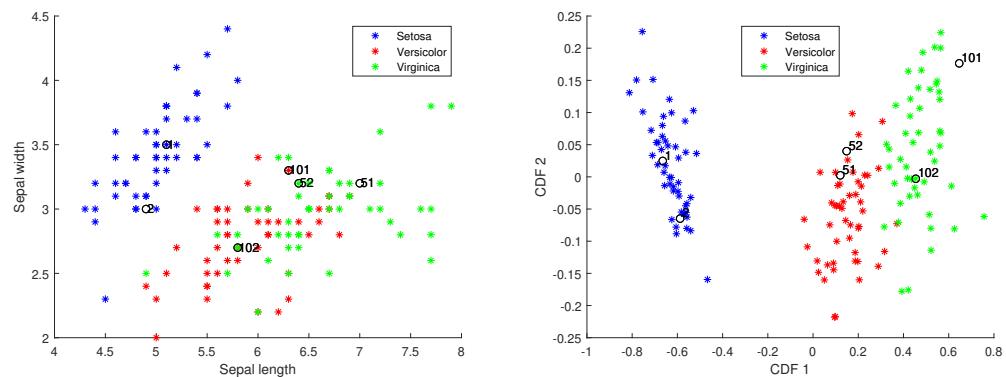
$$[\mathbf{d}_1^T(\mathbf{x}_1 - \bar{\mathbf{x}})] = [-0.0684 \ -0.1266 \ 0.1815 \ 0.2318] \begin{bmatrix} 5.1 - 5.8433 \\ 3.5 - 3.0573 \\ 1.4 - 3.7580 \\ 0.2 - 1.1993 \end{bmatrix} = -0.6649,$$

and for the second canonical discriminant function

$$[\mathbf{d}_1^T(\mathbf{x}_1 - \bar{\mathbf{x}})] = [0.0020 \quad 0.1785 \quad -0.0769 \quad 0.2341] \begin{bmatrix} 5.1 - 5.8433 \\ 3.5 - 3.0573 \\ 1.4 - 3.7580 \\ 0.2 - 1.1993 \end{bmatrix} = 0.0248$$

This procedure is continued for all the observations.

Finally, we show the the data plotted in the original coordinates as [*Sepal length*, *Sepal width*] compared to a plot of the CDF's



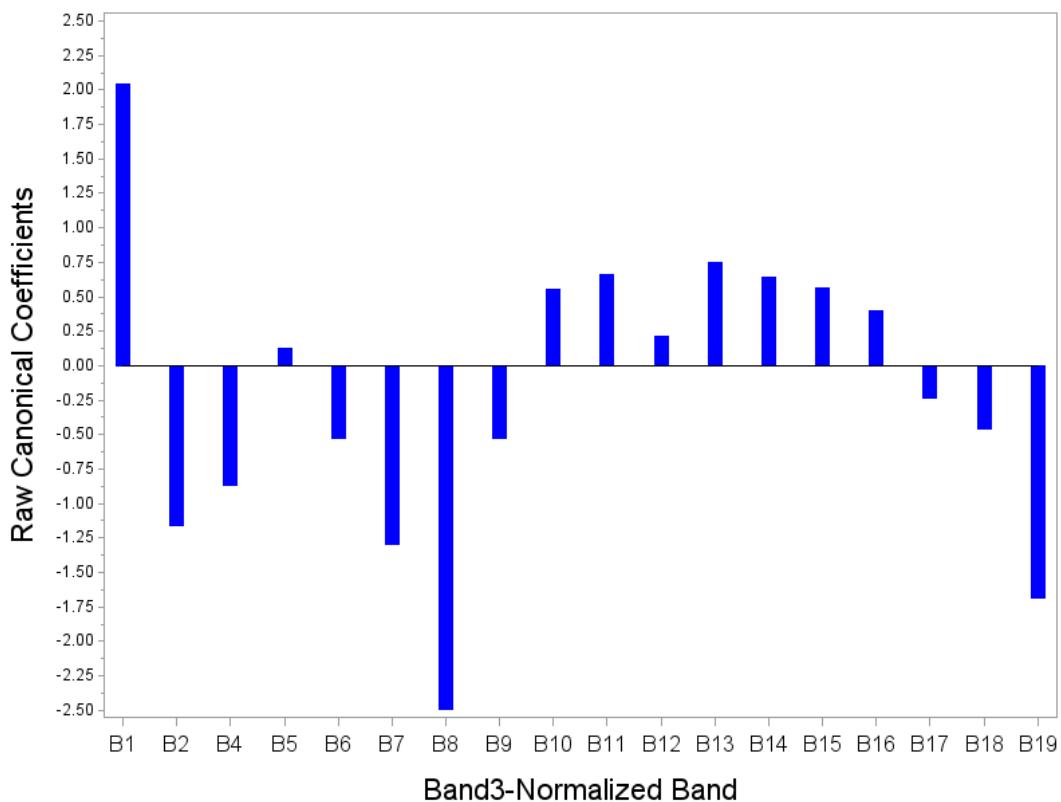
We see how the CDF's maximises the between group distance, while minimising the within group distance.

|||| Remark 5.23

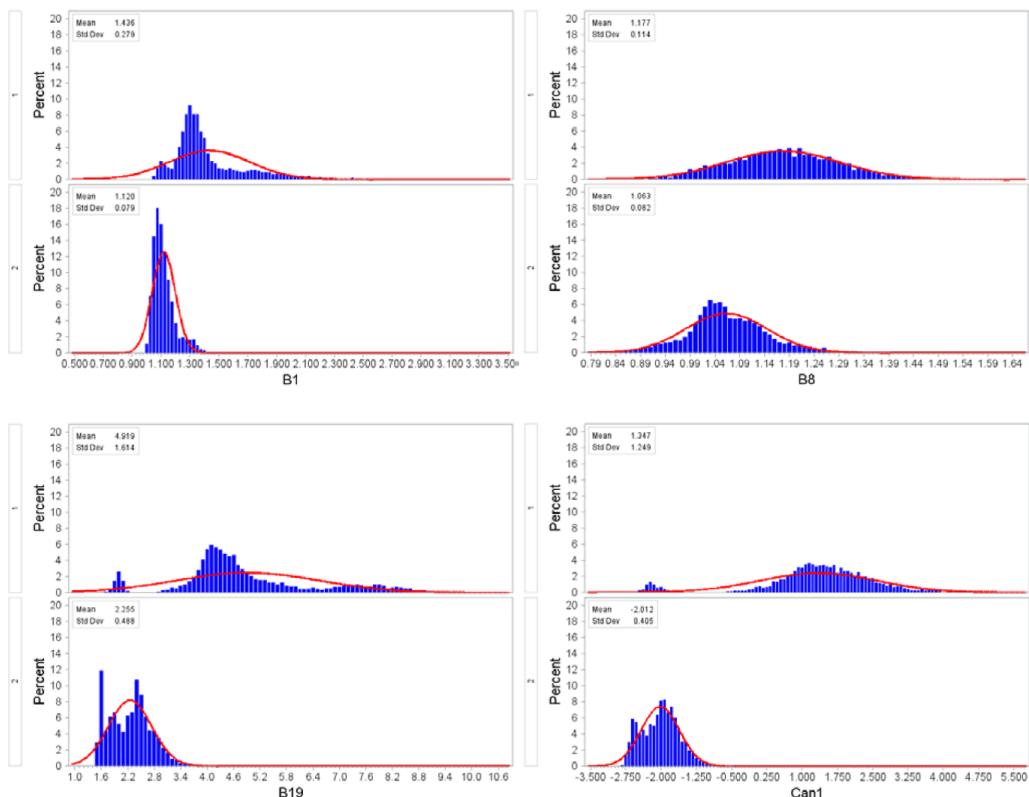
PROC CANDISC in SAS performs the analysis, by doing canonical correlation analysis on the data set, and a data set coded with dummy values for the different classes. The CDF's will thus differ in scale - but otherwise be identical - when compared to CDF's from the approach shown here.

|||| Example 5.24

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.

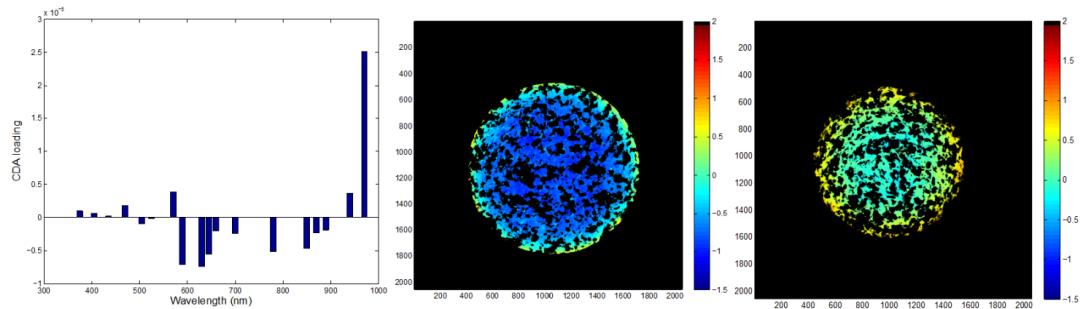


Loadings for computing the canonical discriminant variable in the salami case.



Histograms for the two salami classes fat and meat for selected bands (normalized) and for the canonical discriminant scores in the salami case.

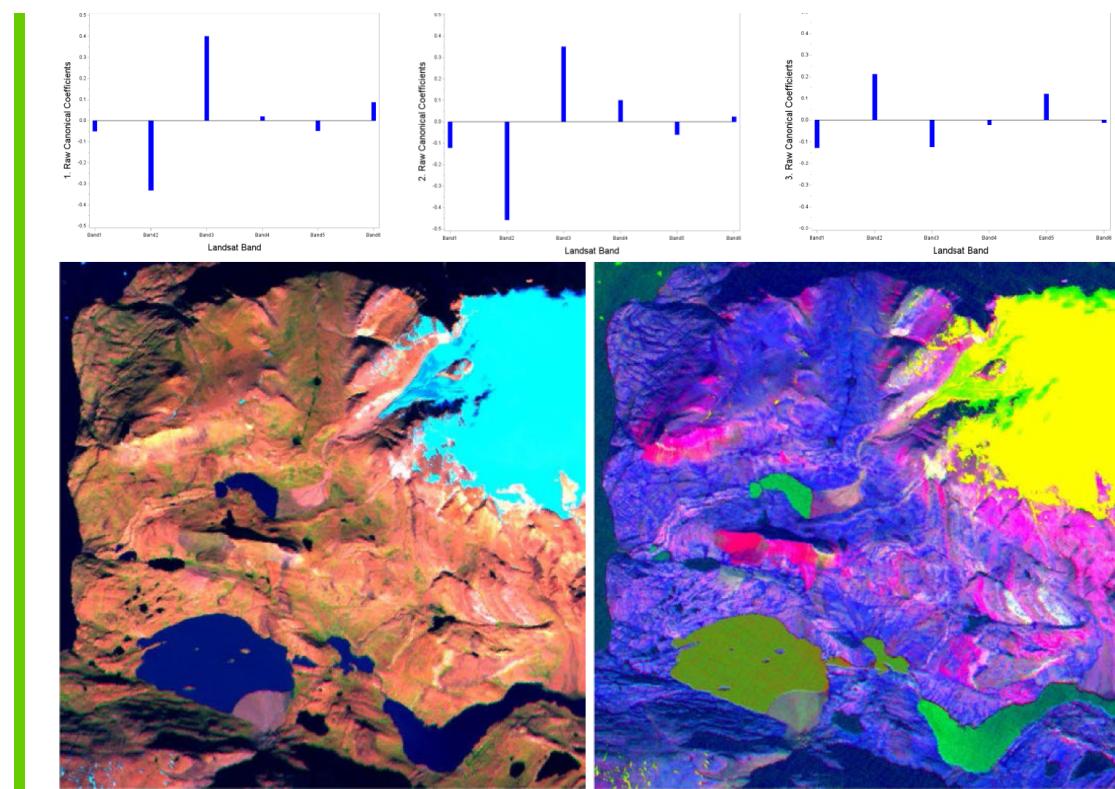
Based on the meat parts from days 2 and 42 we now make another CDA in order to distinguish between meat at those two days. In this way we obtain a statistical color scale shown in the figure below.



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 5.25

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, factor analysis, and regression and latent structures

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 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.

Finally, we will give a brief overview of various regression techniques that

take advantage of the latent structure of the data. These are especially used in chemometrics, but have applications in many fields.

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 of $\boldsymbol{\Sigma}$ in 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 X we mean the direction of the eigenvector \mathbf{p}_i corresponding to the i 'th largest eigenvalue.

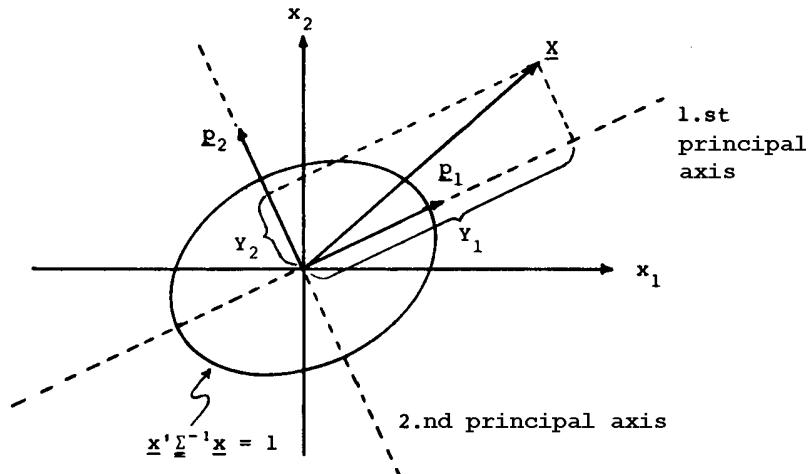


Figure 6.1

||| Definition 6.2

By the i 'th *principal component* of X we will understand X 's projection $Y_i = p_i^T X$ on the i 'th principal axis.

The vector

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} = P^T X$$

is called the *vector of principal components*.

The situation has been sketched geometrically in figure 6.1 where we have drawn the unit ellipsoid corresponding to the variance-covariance structure i.e. the ellipsoid with the equation

$$x^T \Sigma^{-1} x = 1.$$

It is seen that the principal axes are the main axes in this ellipsoid.

A number of theorems about the characteristics of the principal components are statistical reformulations of a number of the results corresponding to symmetrical positive semidefinite matrices which are given in appendix 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.24 (p. 457) we have

$$\begin{aligned} D(\mathbf{Y}) &= D(\mathbf{P}^T \mathbf{X}) = \mathbf{P}^T \boldsymbol{\Sigma} \mathbf{P} = \boldsymbol{\Lambda} \\ &= \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}, \end{aligned}$$

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. 58 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:

$$\sup_{\|\mathbf{b}\|=1} V(\mathbf{b}^T \mathbf{X}) = \lambda_1,$$

and the supremum is obtained for $\mathbf{b} = \mathbf{p}_1$. Further we have

$$\sup_{\substack{\mathbf{b} \perp \mathbf{p}_1, \dots, \mathbf{p}_{m-1} \\ \|\mathbf{b}\|=1}} V(\mathbf{b}^T \mathbf{X}) = \lambda_m,$$

and the supremum is obtained for $\mathbf{b} = \mathbf{p}_m$

|||| Proof

Since

$$\text{V}(\mathbf{b}^T \mathbf{X}) = \mathbf{b}^T \boldsymbol{\Sigma} \mathbf{b},$$

and

$$\begin{aligned}\text{Cov}(Y_i, \mathbf{b}^T \mathbf{X}) &= \text{Cov}(\mathbf{p}_i^T \mathbf{X}, \mathbf{b}^T \mathbf{X}) = \mathbf{p}_i^T \boldsymbol{\Sigma} \mathbf{b} \\ &= \lambda_i \mathbf{p}_i^T \mathbf{b},\end{aligned}$$

so that

$$\text{Cov}(Y_i, \mathbf{b}^T \mathbf{X}) = 0 \Leftrightarrow \mathbf{p}_i \perp \mathbf{b},$$

the theorem is just a reformulation of theorem A.33 p. 463.

■

|||| **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

$$\frac{\lambda_1 + \dots + \lambda_m}{\lambda_1 + \dots + \lambda_m + \dots + \lambda_k}$$

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 \mathbf{X} from the vector*

$$\mathbf{Y}^* = (Y_1, \dots, Y_m, 0, \dots, 0)^T.$$

Since $\mathbf{Y} = \mathbf{P}^T \mathbf{X}$ it implies

$$\mathbf{X} = \mathbf{P} \mathbf{Y} = Y_1 p_1 + \dots + Y_m p_m + \dots + Y_k p_k$$

It is tempting to try with

$$\mathbf{X}^* = \mathbf{P} \mathbf{Y}^* = Y_1 p_1 + \dots + Y_m p_m$$

We find

$$\begin{aligned} D(\mathbf{X}^*) &= \mathbf{P} D(\mathbf{Y}^*) \mathbf{P}^T \\ &= (\mathbf{p}_1 \dots \mathbf{p}_k) \begin{pmatrix} \lambda_1 & & \cdots & 0 \\ & \ddots & & \\ \vdots & & \lambda_m & \vdots \\ 0 & \cdots & & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p}_1^T \\ \vdots \\ \mathbf{p}_k^T \end{pmatrix} \\ &= \lambda_1 \mathbf{p}_1 \mathbf{p}_1^T + \dots + \lambda_m \mathbf{p}_m \mathbf{p}_m^T. \end{aligned}$$

The spectral decomposition of Σ is (p. 458)

$$\Sigma = \lambda_1 \mathbf{p}_1 \mathbf{p}_1^T + \dots + \lambda_m \mathbf{p}_m \mathbf{p}_m^T + \lambda_{m+1} \mathbf{p}_{m+1} \mathbf{p}_{m+1}^T + \dots + \lambda_k \mathbf{p}_k \mathbf{p}_k^T,$$

which means that

$$\Sigma - D(\mathbf{X}^*) = \lambda_{m+1} \mathbf{p}_{m+1} \mathbf{p}_{m+1}^T + \dots + \lambda_k \mathbf{p}_k \mathbf{p}_k^T.$$

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 was 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. [Anderson \(1958\)](#)).

There is, however, a very common problem by using the principal components since they 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.

$$Y_{i\ell} = \frac{X_{i\ell} - \bar{X}_\ell}{\sqrt{\sum_i (\bar{X}_{i\ell} - \bar{X}_\ell)^2 / (n-1)}},$$

where

$$\mathbf{X}_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ik} \end{pmatrix}, \quad i = 1, \dots, n.$$

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.

|||| **Theorem 6.8**

If we are using the estimated *variance-covariance matrix* $\hat{\Sigma}$, the test statistic for testing the hypothesis above becomes

$$Z_1 = -n^* \log \frac{\det \hat{\Sigma}}{\hat{\lambda}_1 \cdot \dots \cdot \hat{\lambda}_m \cdot \hat{\lambda}_*^{k-m}} = -n^* \log \frac{\hat{\lambda}_{m+1} \cdot \dots \cdot \hat{\lambda}_k}{\hat{\lambda}_*^{k-m}},$$

where

$$n^* = n - m - \frac{1}{6}(2(k - m) + 1 + \frac{2}{k - m}),$$

and

$$\hat{\lambda}_* = (\text{tr } \hat{\Sigma} - \hat{\lambda}_1 - \dots - \hat{\lambda}_m)/(k - m) = (\hat{\lambda}_{m+1} + \dots + \hat{\lambda}_k)/(k - m).$$

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

$$Z_2 = -n \log \frac{\det \hat{R}}{\hat{\lambda}_1 \cdot \dots \cdot \hat{\lambda}_m \cdot \hat{\lambda}_*^{k-m}} = -n \log \frac{\hat{\lambda}_{m+1} \cdot \dots \cdot \hat{\lambda}_k}{\hat{\lambda}_*^{k-m}},$$

where

$$\hat{\lambda}_* = (k - \hat{\lambda}_1 - \dots - \hat{\lambda}_m)/(k - m) = (\hat{\lambda}_{m+1} + \dots + \hat{\lambda}_k)/(k - m).$$

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.

|||| **Proof omitted**

A discussion of the above mentioned tests can be found in [Lawley \(1940\)](#).

We now give an example.

||| Example 6.9

The example is based on an example by John C. Davis (see [Davis \(1973\)](#) 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 features of 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 eigenvalues 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 slightly 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, it has numerically large 5th and 6th coordinates. This means that the first principal component

$$Y_1 = 0.164X_1 + \dots + 0.546X_5 + 0.768X_6 + 0.073X_7$$

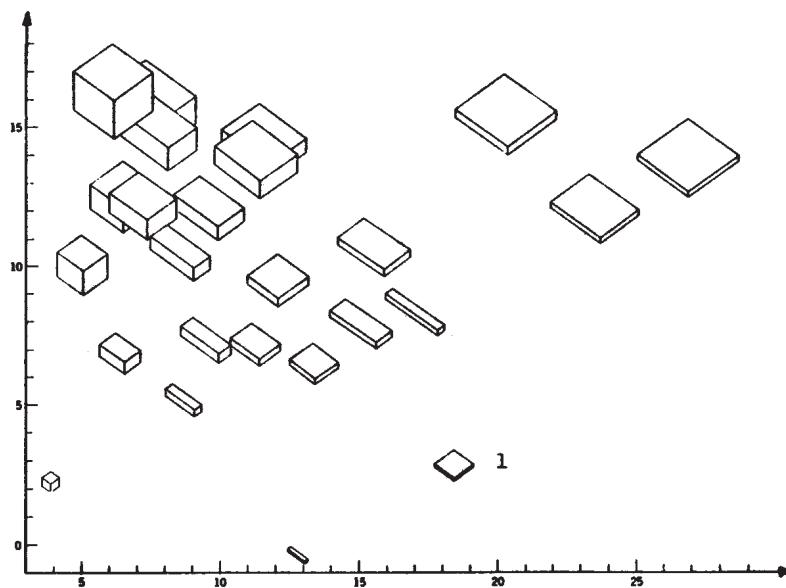
is especially sensitive to variations in X_5 and X_6 . These two variables: The ratio between the radius in the circumscribed sphere and the radius in the inscribed sphere and the ratio between the sum of the two longest sides and the shortest side both have something to do with how “flat” a box is. 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.



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 $X_1 >> X_2$ 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 [Agterberg \(1973\)](#) p. 128).

||| Example 6.10

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. Since we are analysing variables which are measured in non-comparable units we must standardise the numbers.

The eigenvalues and the corresponding eigenvectors are

$$\begin{aligned}\hat{\lambda}_1 &= 2.25; \quad \hat{p}_1 = \begin{bmatrix} 0.43 \\ 0.55 \\ 0.57 \\ -0.44 \end{bmatrix} \\ \hat{\lambda}_2 &= 0.74; \quad \hat{p}_2 = \begin{bmatrix} -0.66 \\ 0.49 \\ 0.37 \\ 0.44 \end{bmatrix} \\ \hat{\lambda}_3 &= 0.70; \quad \hat{p}_3 = \begin{bmatrix} 0.60 \\ -0.02 \\ 0.16 \\ 0.78 \end{bmatrix} \\ \hat{\lambda}_4 &= 0.31; \quad \hat{p}_4 = \begin{bmatrix} -0.14 \\ -0.68 \\ 0.72 \\ -0.06 \end{bmatrix}\end{aligned}$$

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

$$Z = -n \log \frac{0.74 \cdot 0.70 \cdot 0.31}{[(0.74 + 0.70 + 0.31)/3]^3} = 0.2120n,$$

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 $30 \text{ m} \times 30 \text{ m}$.

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

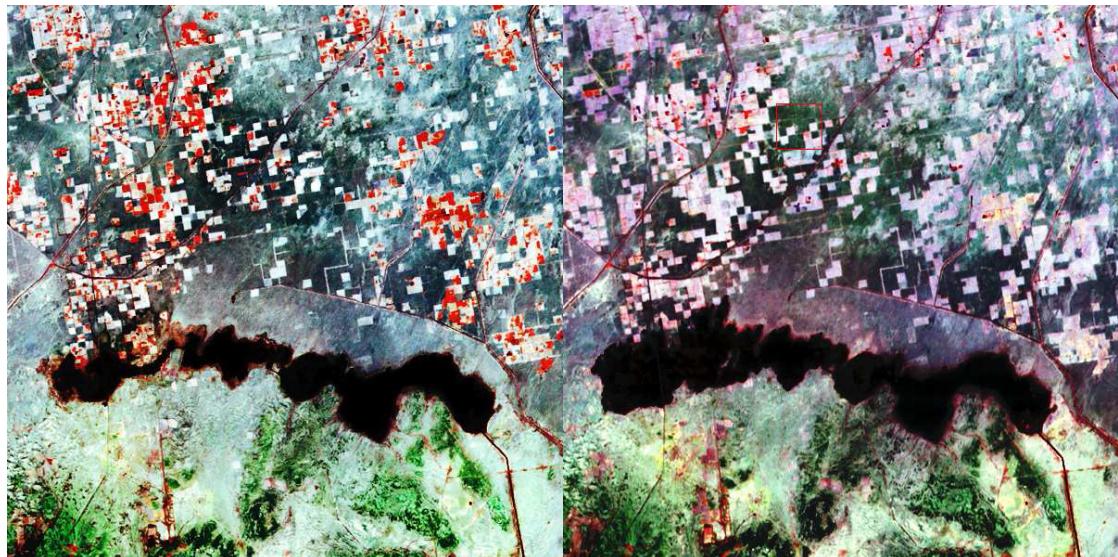


Figure 6.2 – Landsat images from India observed in May and in March.

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:

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{Y}_i \\ \mathbf{X}_i \end{bmatrix}, \quad \mathbf{Y}_i = \begin{bmatrix} b1_{May,i} \\ \vdots \\ b6_{May,i} \end{bmatrix}, \quad \mathbf{X}_i = \begin{bmatrix} b1_{March,i} \\ \vdots \\ b6_{March,i} \end{bmatrix}$$

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 6.11

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*.

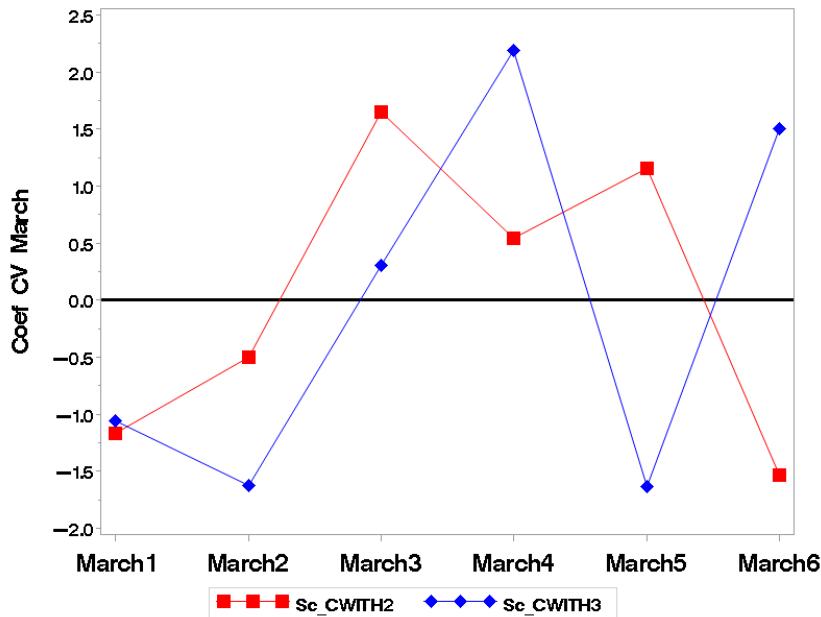


Figure 6.3 – The standardized canonical coefficients for canonical variables 2 and 3 of the March variables (the WITH variables).

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 $\mathbf{A}_{\text{std}} = \sigma(\mathbf{Y}) \mathbf{A}$ and $\mathbf{B}_{\text{std}} = \sigma(\mathbf{X}) \mathbf{B}$ 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. 6.4

We shall now show a relation between the canonical correlations and independence between the two set of variables.

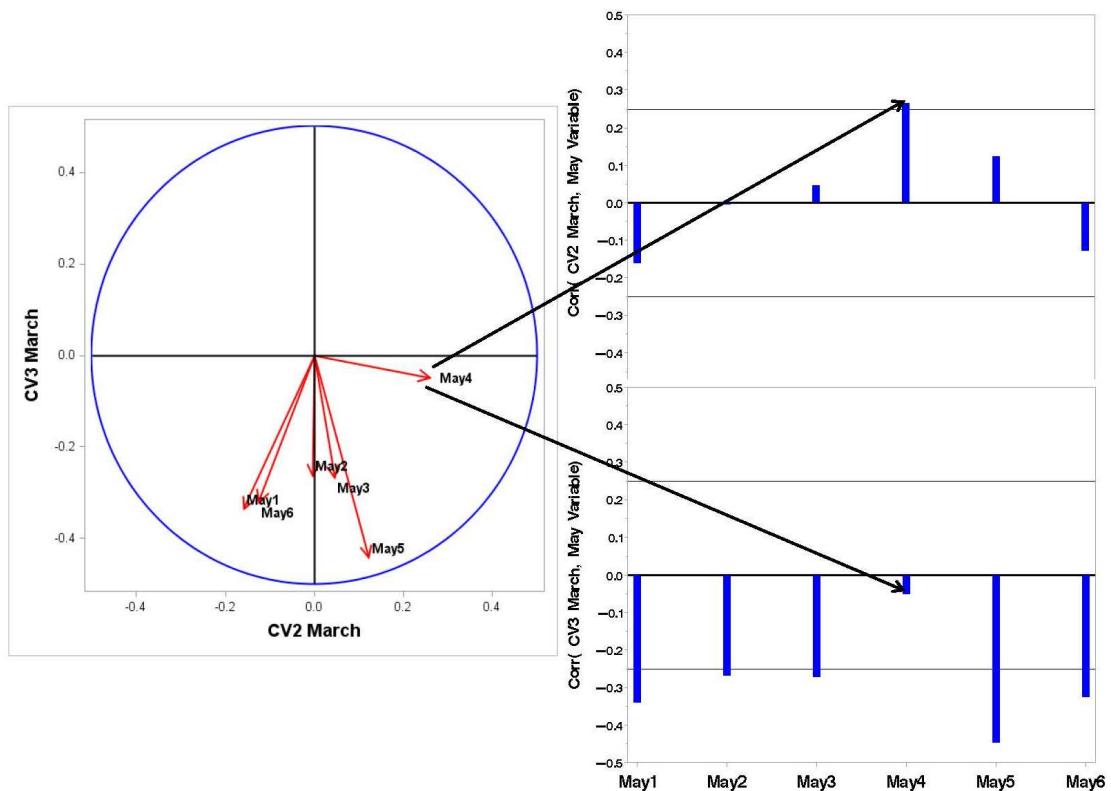


Figure 6.4 – On the right hand side we show the correlations between the May values and the second and the third canonical correlation of the March variables, each mapped against the May variable name. On the left hand side we have shown the same correlations mapped as vectors, one for each May variable.

||| Theorem 6.12

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) $\begin{bmatrix} -\rho_r \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & -\rho_r \Sigma_{xx} \end{bmatrix} \begin{bmatrix} \mathbf{a}_r \\ \mathbf{b}_r \end{bmatrix} = \mathbf{0}$
- (ii) $\mathbf{a}_r^T \Sigma_{yy} \mathbf{a}_r = 1$
- (iii) $\mathbf{b}_r^T \Sigma_{xx} \mathbf{b}_r = 1$

||| Proof

We have a maximization problem with constraints and one can solve the problem by using a Lagrange multiplier technique, see e.g. [Anderson \(1958\)](#) p. 289.

■

One can also determine the correlations and the coefficients by solving an eigenvalue problem since we have

|||| Theorem 6.13

Let the situation be as in the previous theorem. Then we have

$$(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \rho_r^2\Sigma_{yy})\mathbf{a}_r = \mathbf{0}$$

$$\det(\Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy} - \rho_r^2\Sigma_{yy}) = 0$$

respectively

$$(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \rho_r^2\Sigma_{xx})\mathbf{b}_r = \mathbf{0}$$

$$\det(\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} - \rho_r^2\Sigma_{xx}) = 0$$

|||| Proof omitted

|||| Theorem 6.14

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

$$\begin{aligned}\Sigma_{yx} = \mathbf{0} &\Leftrightarrow \forall \mathbf{a}, \mathbf{b} : \rho(\mathbf{a}^T \mathbf{Y}, \mathbf{b}^T \mathbf{X}) = 0 \\ &\Leftrightarrow \forall \mathbf{a}, \mathbf{b} : \rho^2(\mathbf{a}^T \mathbf{Y}, \mathbf{b}^T \mathbf{X}) = 0 \\ &\Leftrightarrow \max_{\mathbf{a}, \mathbf{b}} : \rho^2(\mathbf{a}^T \mathbf{Y}, \mathbf{b}^T \mathbf{X}) = 0,\end{aligned}$$

which concludes the proof. ■

|||| Remark 6.15

Consequently, we may test whether \mathbf{Y} and \mathbf{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 \mathbf{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 \mathbf{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 data matrices \mathbf{Y} and \mathbf{X} should not be confused with the indexless general notation for the vector 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 6.16

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 6.17**

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.

||| Case 6.18 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,	(CaO) ₃ SiO ₂	52-63
Belite	Dicalcium silicate	C2S,	(CaO) ₂ SiO ₂	11-24
Aluminate	Tricalcium aluminate	C3A,	(CaO) ₃ Al ₂ O ₃	4-8
Ferrite	Tetracalcium aluminoferrite	C4AF,	(CaO) ₄ Al ₂ O ₃ Fe ₂ O ₃	1-11
Gypsum	Calciumsulfatedihydrate	CSH ₂	(CaO)SO ₃ ·2H ₂ 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 1.35.

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

$$MS = \frac{\text{Al}_2\text{O}_3}{\text{Fe}_2\text{O}_3},$$

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 [Pedersen and Skjøth \(1978\)](#) (see pp. 178). We are analyzing data on 198 samples of Portland cement. Means, standard deviations, and correlations are presented in table 3 and 4. 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 1.

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 Na2O	GL TAB	FRI CAO	X SO3	MS	MA	LSF	BLA INE	LSR
C3S	1.00	-.89	-.11	-.27	-.03	-.31	-.27	-.33	0.64	-.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.

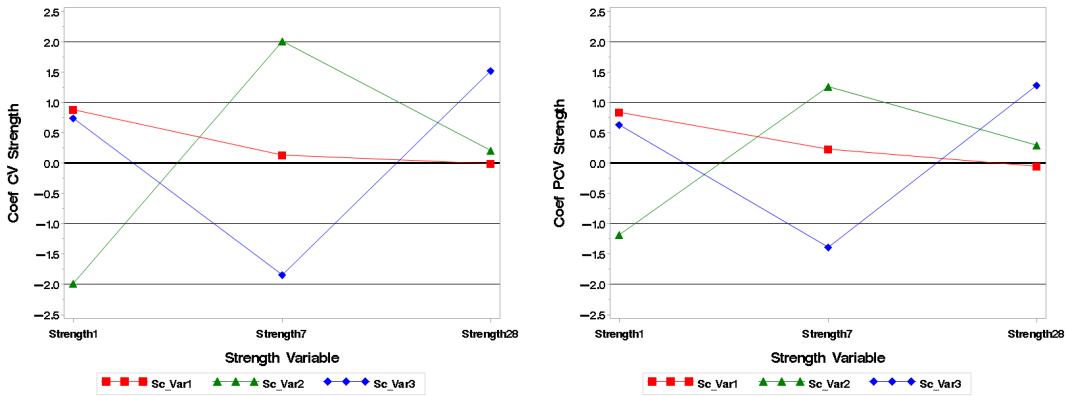


Figure 1: Standardized coefficients for computing the strength canonical variables in the full variable case and in the partial case.

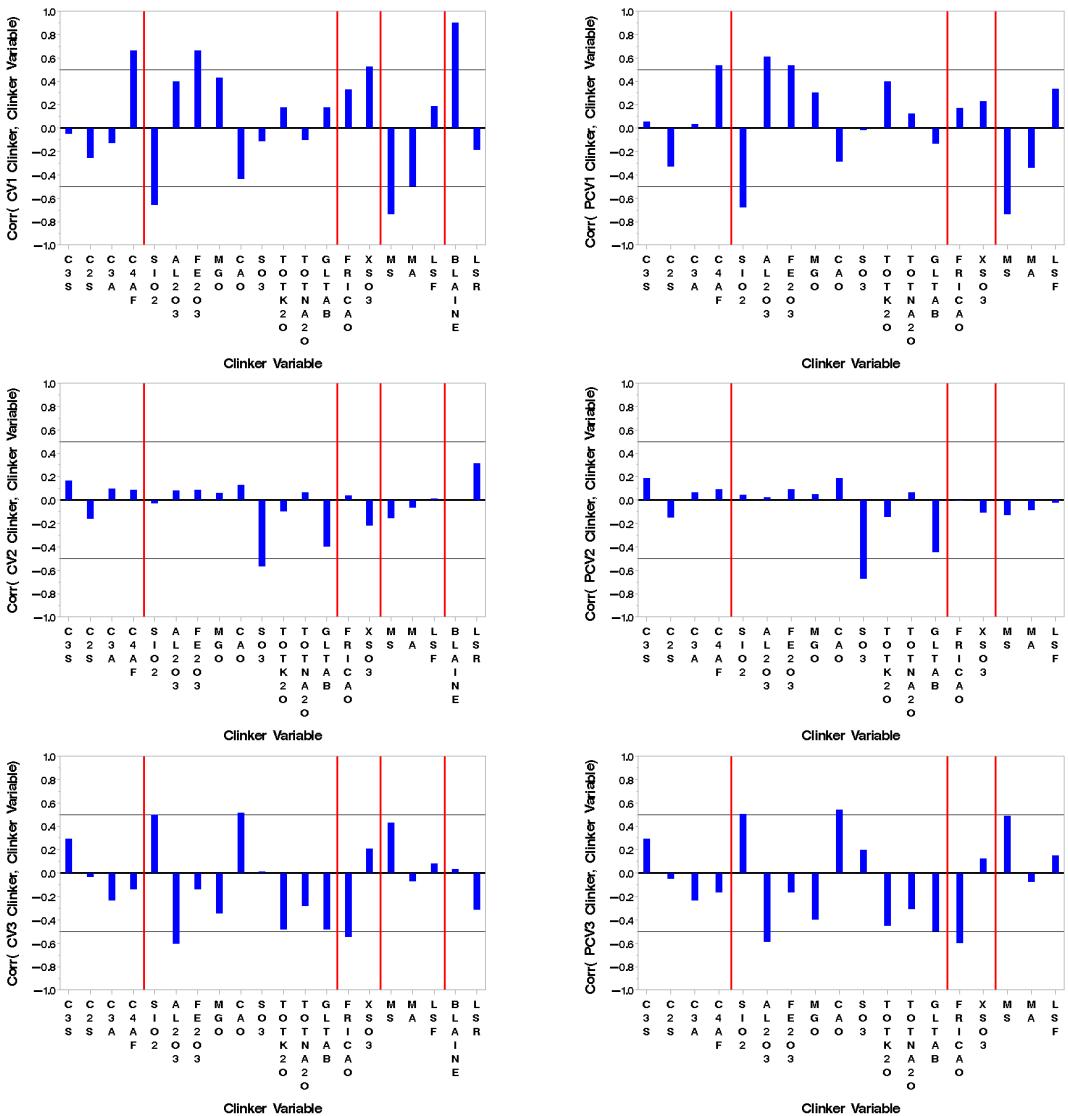


Figure 2: Correlations between the canonical variables for the clinkers and the clinker variables. Left: Ordinary analysis, right: Analysis partialled on Blaine and LSR.

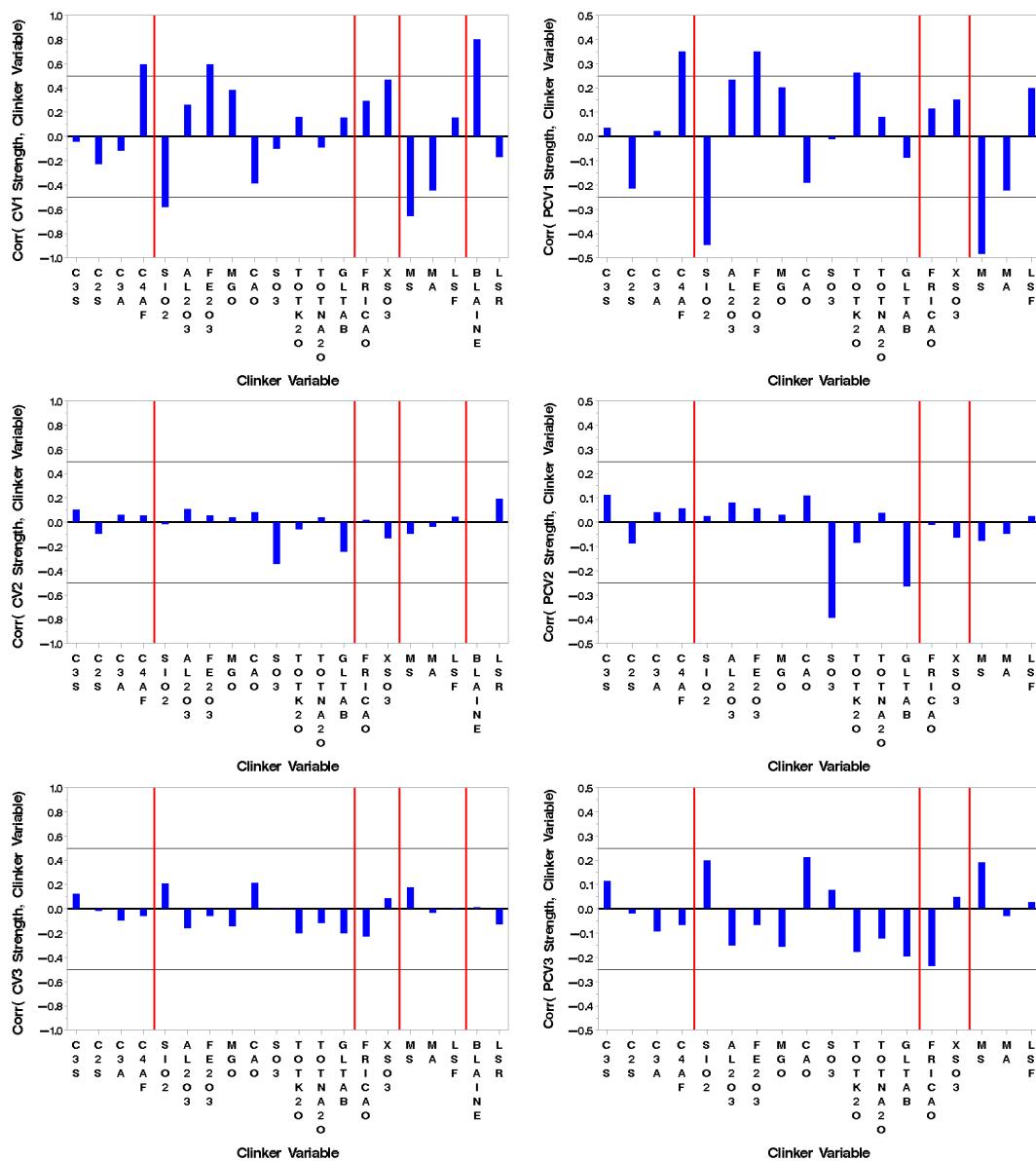


Figure 3: Correlations between the canonical variables for the strength measurements and the clinker variables. Left: Ordinary analysis, right: Analysis partialled on Blaine and LSR.

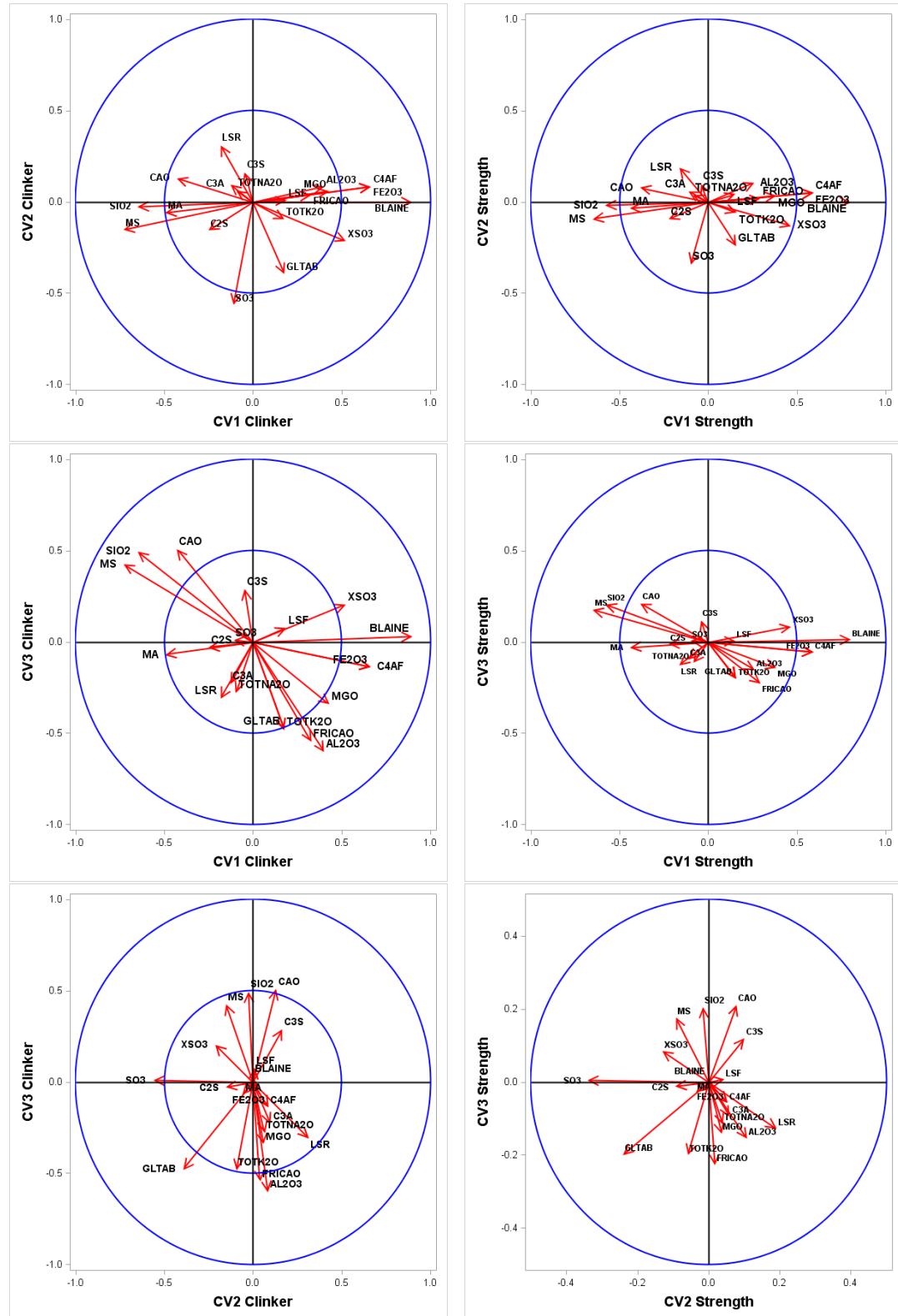


Figure 4: The correlations from Figures 2 and 3 (left part) presented as vector plots.

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 (\sim CV2 strength), and if we condition on the fineness it is also positively correlated with the development from day 7 to day 28 (\sim 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 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 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

$$(outcome \text{ of } \text{geometric} \text{ test}) \sim \\ (\text{large } a_{11}) \times (\text{unknown } F_1) + \dots + (\text{small } a_{1m}) \times (\text{unknown } F_m) \quad (6-1)$$

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 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 X and F and 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

$$\begin{bmatrix} X_{i1} \\ \vdots \\ X_{ik} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \begin{bmatrix} F_{i1} \\ \vdots \\ F_{im} \end{bmatrix} + \begin{bmatrix} G_{i1} \\ \vdots \\ G_{ik} \end{bmatrix},$$

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

$$\begin{bmatrix} X_{11} \cdots X_{n1} \\ \vdots \\ X_{k1} \cdots X_{nk} \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \\ a_{k1} \cdots a_{km} \end{bmatrix} \begin{bmatrix} F_{11} \cdots F_{n1} \\ \vdots \\ F_{1m} \cdots F_{nm} \end{bmatrix} + \begin{bmatrix} G_{11} \cdots G_{n1} \\ \vdots \\ G_{1k} \cdots G_{nk} \end{bmatrix}.$$

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

$$R = AA^T + \Delta.$$

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.

$$\text{Cov}(X_i, F_j) = \text{Cov}\left(\sum_{\nu} a_{i\nu} F_{\nu} + G_i, F_j\right) = a_{ij}.$$

It can be shown [Dwyer \(1939\)](#) 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

$$\mathbf{A} \mathbf{A}^T = \mathbf{R} - \Delta.$$

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.24 p. 457, that

$$\mathbf{P}^T \mathbf{V} \mathbf{P} = \Lambda = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}.$$

Since \mathbf{P} is orthogonal we get

$$\mathbf{V} = \mathbf{P} \Lambda \mathbf{P}^T = (\mathbf{P} \Lambda^{\frac{1}{2}})(\mathbf{P} \Lambda^{\frac{1}{2}})^T,$$

where

$$\Lambda^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda}_k \end{pmatrix}.$$

We now define

$$\Lambda_*^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ & \ddots & \vdots \\ \vdots & & \sqrt{\lambda}_m \\ 0 & \cdots & 0 \end{pmatrix}.$$

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

$$\begin{aligned} (\mathbf{P} \Lambda_*^{\frac{1}{2}})(\mathbf{P} \Lambda_*^{\frac{1}{2}})^T &= \mathbf{P} \Lambda_*^{\frac{1}{2}} \Lambda_*^{\frac{1}{2}}' \mathbf{P}^T \\ &= \mathbf{P} \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \lambda_m & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \mathbf{P}^T \\ &\simeq \mathbf{V}, \end{aligned}$$

cf. the analogous considerations p. 373.

Since \mathbf{V} is an estimate of $\mathbf{A}\mathbf{A}^T$, we then have

$$\mathbf{A}\mathbf{A}^T \simeq (\mathbf{P}\boldsymbol{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\boldsymbol{\Lambda}_*^{\frac{1}{2}})^T,$$

so it would be natural to choose $\mathbf{P}\boldsymbol{\Lambda}_*^{\frac{1}{2}}$ 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 6.19

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

$$\mathbf{P}\boldsymbol{\Lambda}_*^{\frac{1}{2}} = (\sqrt{\lambda_1}\mathbf{p}_1, \dots, \sqrt{\lambda_m}\mathbf{p}_m),$$

where λ_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 6.20

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

$$\mathbf{A} \mathbf{A}^T \simeq (\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}})(\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}})^T$$

If \mathbf{Q} is an arbitrary $m \times m$ orthogonal matrix i.e. $\mathbf{Q} \mathbf{Q}' = \mathbf{I}$ then we have

$$\begin{aligned} (\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}} \mathbf{Q})(\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}} \mathbf{Q})^T &= (\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}}) \mathbf{Q} \mathbf{Q}^T (\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}})^T \\ &= (\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}})(\mathbf{P} \boldsymbol{\Lambda}_*^{\frac{1}{2}})^T \\ &= \mathbf{A} \mathbf{A}^T. \end{aligned}$$

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 orthogonal 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 6.21

We will now perform a factor analysis on the data given in example 6.9.

First we determine the correlation matrix. From the estimate of the variance-covariance matrix p. 376 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 6.9 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 6.19 the estimated principal factor solution to the problem is $(\sqrt{\hat{\lambda}_1}\hat{p}_1, \sqrt{\hat{\lambda}_2}\hat{p}_2)$, where

$$\begin{pmatrix} \sqrt{\hat{\lambda}_1}\hat{p}_1^T \\ \sqrt{\hat{\lambda}_2}\hat{p}_2^T \end{pmatrix} = \begin{pmatrix} 0.747 & 0.795 & 0.710 & 0.910 & -0.235 & -0.178 & -0.886 \\ 0.491 & 0.373 & -0.596 & 0.389 & 0.963 & 0.971 & 0.218 \end{pmatrix}.$$

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 $\hat{\delta}_j = 1 - \hat{h}_j^2$ 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. 403). We find

$$\hat{\delta}^T = [0.202 \ 0.229 \ 0.140 \ 0.021 \ 0.017 \ 0.024 \ 0.167].$$

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

$$\hat{\mathbf{Z}} = \hat{\mathbf{R}} - \hat{\mathbf{A}}\hat{\mathbf{A}}^T,$$

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 6.9 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. 410) \hat{a}_{ij} in a two-dimensional coordinate system. They are shown in figure 1

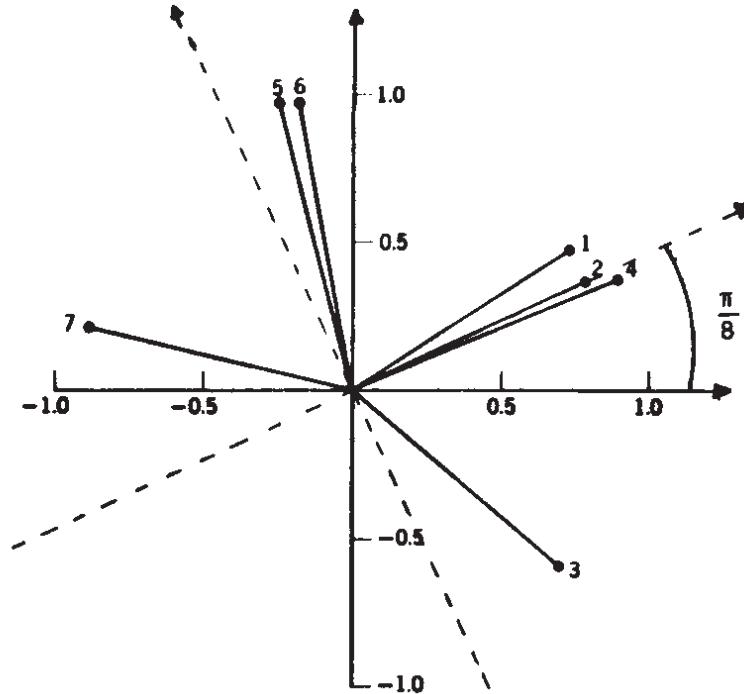


Figure 1

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 in example 6.22 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. [Harman \(1967\)](#) and [Cattell \(1965\)](#)).

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 [Harman \(1967\)](#) and [Cattell \(1965\)](#).

6.3.4 Computation of the factor scores

If we in the above mentioned example 6.21 wish to make a diagram analogous to the one mentioned in example 6.9 p. 378 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. 402)

$$\mathbf{X} = \mathbf{A}\mathbf{F} + \mathbf{G},$$

where

$$\begin{aligned} D(\mathbf{F}) &= \mathbf{I} \\ D(\mathbf{G}) &= \Delta, \end{aligned}$$

and where \mathbf{F} and \mathbf{G} are uncorrelated.

Therefore we have

$$D \left(\begin{array}{c} \mathbf{X} \\ \mathbf{F} \end{array} \right) = \left(\begin{array}{cc} \mathbf{A}\mathbf{A}^T + \Delta & \mathbf{A} \\ \mathbf{A}^T & \mathbf{I} \end{array} \right).$$

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

$$\begin{bmatrix} \hat{\mathbf{A}}\hat{\mathbf{A}}^T + \hat{\Delta} & \hat{\mathbf{A}} \\ \hat{\mathbf{A}}^T & \mathbf{I} \end{bmatrix}.$$

Assuming that the underlying distributions are normal, the conditional distribution of \mathbf{F} given \mathbf{X} has the mean value

$$\boldsymbol{\mu}_F + \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \Delta)^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)$$

(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

$$\mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \Delta)^{-1}\mathbf{x}$$

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

$$\hat{F}_i = \hat{\mathbf{A}}^T(\hat{\mathbf{A}}\hat{\mathbf{A}}^T + \hat{\Delta})^{-1}\mathbf{X}_i. \quad (6-2)$$

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

$$(\mathbf{A}\mathbf{A}^T + \Delta)^{-1}\mathbf{A} = \Delta^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}^T\Delta^{-1}\mathbf{A})^{-1},$$

which gives

$$\hat{F}_i = (\mathbf{I} + \hat{\mathbf{A}}^T\hat{\Delta}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}^T\hat{\Delta}^{-1}\mathbf{X}_i. \quad (6-3)$$

The validity of the identity is found by the following relationships

$$\begin{aligned} (\mathbf{A}\mathbf{A}^T + \Delta)^{-1}\mathbf{A} &= \Delta^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}^T\Delta^{-1}\mathbf{A})^{-1} \\ \Leftrightarrow \mathbf{A} &= (\mathbf{A}\mathbf{A}^T + \Delta)\Delta^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}^T\Delta^{-1}\mathbf{A})^{-1} \\ &= \mathbf{A}(\mathbf{A}^T\Delta^{-1}\mathbf{A} + \mathbf{I})(\mathbf{I} + \mathbf{A}^T\Delta^{-1}\mathbf{A})^{-1}, \end{aligned}$$

and the last relationship is trivially fulfilled.

Now $\mathbf{I} + \mathbf{A}^T\Delta^{-1}\mathbf{A}$ 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 $(\mathbf{A}\mathbf{A}^T + \Delta)$ 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

$$\mathbf{R} = \mathbf{A}\mathbf{A}^T + \Delta$$

(cf. p. 404). This gives the expression which is equivalent to (6-2)

$$\hat{\mathbf{F}}_i = \hat{\mathbf{A}}^T \hat{\mathbf{R}}^{-1} \mathbf{X}_i \quad (6-4)$$

Finally we must emphasize that there are a number of other methods of determining the factor scores see e.g. Harman (1967) or Morrison (1967). 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 6.22

In example 6.21, p. 409 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. 404)

$$\hat{h}_j^2 = \hat{a}_{j1}^2 + \hat{a}_{j2}^2 = 1 - \hat{\delta}_j.$$

We note that the given communalities are equal to those we found on p. 410 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

$$(\mathbf{I} + \hat{\mathbf{A}}^T \hat{\Delta}^{-1} \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T \hat{\Delta}^{-1} =$$

$$\begin{bmatrix} 0.0669 & 0.0597 & 0.0593 & 0.7839 & 0.0244 & 0.0510 & -0.0750 \\ -0.0002 & -0.0059 & 0.0655 & -0.0943 & 0.5770 & 0.3641 & 0.0415 \end{bmatrix}$$

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

$$\mathbf{z} = (-1.4373 \quad -0.4603 \quad -1.0860 \quad -1.2787 \quad 1.3167 \quad 1.4422 \quad 1.5124)^T,$$

where e.g. the second value is found as

$$z_2 = \frac{3.660 - 4.773}{2.4178} = -0.4603.$$

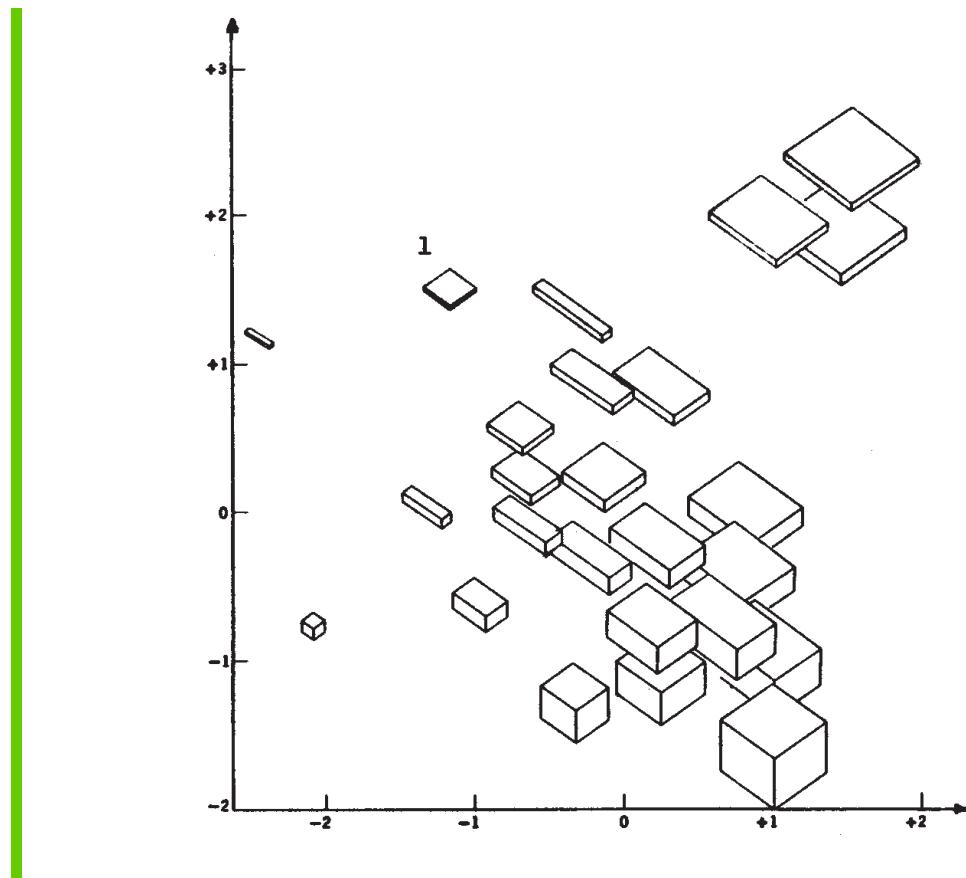
We now easily find the factor scores corresponding to the first box as

$$\hat{\mathbf{F}}_1 = (\mathbf{I} + \hat{\mathbf{A}}^T \hat{\Delta}^{-1} \hat{\mathbf{A}})^{-1} \hat{\mathbf{A}}^T \hat{\Delta}^{-1} \mathbf{z} = \begin{pmatrix} -1.20 \\ 1.40 \end{pmatrix}.$$

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. 378).

We note (cf. example 6.21) 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 6.9.



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):

$$\log L(\Sigma) = -\frac{1}{2}(n-1)\log(\det \Sigma) - \frac{1}{2}(n-1)\text{tr}(\mathbf{S}\Sigma^{-1}).$$

Here we now introduce the usual m factor model

$$\mathbf{D}(X) = \Sigma = \mathbf{A}\mathbf{A}^T + \Delta,$$

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

$$\log L(\mathbf{A}, \Delta) = -\frac{1}{2}(n-1)\log(\det(\mathbf{A}\mathbf{A}^T + \Delta)) - \frac{1}{2}(n-1)\text{tr}(\mathbf{S}(\mathbf{A}\mathbf{A}^T + \Delta)^{-1}).$$

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 Jöreskog (1967).

By partial differentiation of the logarithm of the likelihood function, and after long and tedious algebraic manipulations, one obtains the equation:

$$\hat{\mathbf{A}} = (\hat{\Delta} + \hat{\mathbf{A}}\hat{\mathbf{A}}^T)\mathbf{S}^{-1}\hat{\mathbf{A}}, \quad (6-5)$$

see e.g. Morrison (1967).

If we perform a scale-transformation of the X 's i.e. we introduce

$$\mathbf{Z}_i = \mathbf{C}X_i,$$

we then have

$$\mathbf{S}_z = \mathbf{C}\mathbf{S}_x\mathbf{C}^T$$

where z and x as subscripts shows whether the different quantities have been computed on the base of the Z_i 's or the 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

$$\mathbf{C}^{-1}\hat{\mathbf{A}}_z = [\mathbf{C}^{-1}\hat{\Delta}_z\mathbf{C}^{T-1} + \mathbf{C}^{-1}\hat{\mathbf{A}}_z(\mathbf{C}^{-1}\hat{\mathbf{A}}_z)^T]\mathbf{S}_x^{-1}\mathbf{C}^{-1}\hat{\mathbf{A}}_z. \quad (6-6)$$

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 s (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

$$H_0 : \Sigma = \Delta + \mathbf{A}\mathbf{A}^T \quad \text{against} \quad H_1 : \Sigma \text{ arbitrary.}$$

The ratio test will then be equivalent to the test given by the test statistic

$$Z = (n - 1 - \frac{1}{6}(2k + 5) - \frac{2}{3}m) \ln \frac{|\hat{\Delta} + \hat{\mathbf{A}}\hat{\mathbf{A}}^T|}{|\mathbf{S}|}$$

and we will reject for

$$Z > \chi^2(\frac{1}{2}\{(k - m)^2 - k - m\}).$$

||| Example 6.23

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 [Kaiser \(1958\)](#)).

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

$$\mathbf{X}' = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kn} \end{bmatrix},$$

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 1.31. In a **dual** way we could of course define

$$\mathbf{Q} = \mathbf{X} \mathbf{X}^T,$$

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.

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 $\mathbf{X} \mathbf{X}^T$ 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.

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

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

$$\hat{\mathbf{Y}}_{OLS} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{X}\hat{\mathbf{B}}_{OLS}$$

i.e. the regression coefficients coefficients for *ordinary least squares regression* are

$$\hat{\mathbf{B}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

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

$$\hat{\mathbf{Y}}_{PCA} = \mathbf{X}\mathbf{P}_f \left((\mathbf{X}\mathbf{P}_f)^T(\mathbf{X}\mathbf{P}_f) \right)^{-1} (\mathbf{X}\mathbf{P}_f)^T\mathbf{Y} = \mathbf{X}\mathbf{P}_f \Lambda_f^{-1} \mathbf{P}_f^T \mathbf{X}^T \mathbf{Y}$$

i.e. the regression coefficients coefficients for *principal component regression* are

$$\hat{\mathbf{B}}_{PCA} = \mathbf{P}_f \Lambda_f^{-1} \mathbf{P}_f^T \mathbf{X}^T \mathbf{Y}$$

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

In this section we consider the SIMPLS solution to the partial least squares problem. SIMPLS is a “statistical” alternative to the usual algorithmic method NIPALS.

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 estimated *model effect (X) loadings* and $\widehat{\mathbf{C}} = \mathbf{Y}^T \mathbf{T} (\mathbf{T}^T \mathbf{T})^{-1}$ the estimated *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.

$$[\text{Cov}(\mathbf{r}^T \mathbf{X}_i, Y_{i1}) \ \dots \ \text{Cov}(\mathbf{r}^T \mathbf{X}_i, Y_{im})] = \text{Cov}(\mathbf{r}^T \mathbf{X}_i, \mathbf{Y}_i) = \mathbf{r}^T \Sigma_{xy}$$

The **sum of those m covariances squared** becomes

$$\mathbf{r}^T \Sigma_{xy} \Sigma_{yx} \mathbf{r}$$

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

$$S_1^T S_1 = Y^T X X^T Y.$$

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

$$\mathbf{P} = \mathbf{X}^T \mathbf{T} (\mathbf{T}^T \mathbf{T})^{-1} = \mathbf{X}^T \mathbf{T}_f \mathbf{D}^{-1},$$

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 Y *factor scores*

$$\mathbf{U}_f.$$

By design $\mathbf{U}_f^T \mathbf{T}$ is lower triangular implying that the $Y - score_i$ and the $X - score_j$ values are (empirically) uncorrelated for $i > j$, i.e. the $Y - score$ values are uncorrelated with the previous $X - score$ values.

The *model effect weights* (X *block factor weights*) are

$$\mathbf{R} = \mathbf{R}_f \mathbf{D}$$

and the *dependent variable weights* (Y *block factor weights*) are

$$\mathbf{Q}_f$$

If we want to express the predicted value $\widehat{\mathbf{Y}}$ as a function of the X -variables, we see that

$$\mathbf{X} \mathbf{R}_f \mathbf{H}_f^T = \mathbf{X} \mathbf{R}_f \mathbf{T}_f^T \mathbf{Y} = \mathbf{T}_f \mathbf{T}_f^T \mathbf{Y} = \mathbf{T} (\mathbf{D}^{-1} \mathbf{D}^{-1}) \mathbf{T}^T \mathbf{Y} = \widehat{\mathbf{Y}}$$

and thus the *partial least squares (SIMPLS) regression coefficients* are

$$\widehat{\mathbf{B}}_{PLS} = \mathbf{B}_f = \mathbf{R}_f \mathbf{H}_f^T$$

|||| 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. Therefore 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) $\forall u, v \in V : u + v = v + u$
commutative law for vector addition
- ii) $\forall u, v, x \in V : u + (v + x) = (u + v) + x$
associative law for vector addition
- iii) $\exists \mathbf{0} \in V \forall u \in V : u + \mathbf{0} = u$
existence of a *neutral element*
- iv) $\forall u \in V \exists -u \in V : u + (-u) = \mathbf{0}$
existence of an *inverse element*
- v) $\forall \lambda \in R \forall u, v \in V : \lambda(u + v) = \lambda u + \lambda v$
distributive law for scalar multiplication
- vi) $\forall \lambda_1, \lambda_2 \in R \forall u \in V : (\lambda_1 + \lambda_2)u = \lambda_1 u + \lambda_2 u$
distributive law for scalar multiplication
- vii) $\forall \lambda_1, \lambda_2 \in R \forall u \in V : (\lambda_1 \lambda_2)u = \lambda_1(\lambda_2 u)$
associative law for scalar multiplication
- viii) $\forall u \in V : 1u = u$

||| 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 \mathbb{R}^n

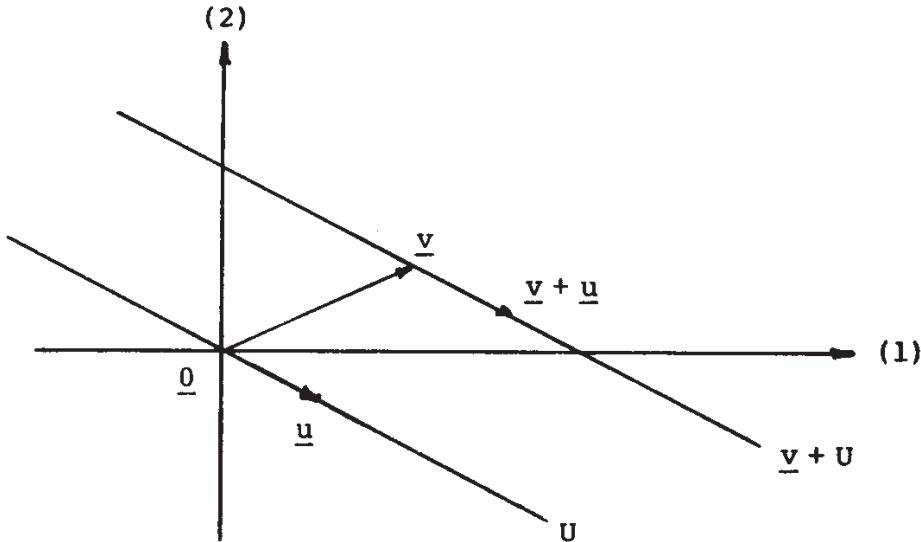


Figure A.1 – Sub-space and corresponding side-subspace in \mathbb{R}^2 .

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 $v_1, \dots, v_k \in V$, we can define the *linear span* of those vectors

$$\text{span}\{v_1, \dots, v_k\}$$

as the smallest subspace of V , which contains $\{v_1, \dots, v_k\}$. It is easily shown that

$$\text{span}\{v_1, \dots, v_k\} = \left\{ \sum_{i=1}^k \alpha_i v_i \mid \alpha_i \in \mathbb{R}, \quad i = 1, \dots, k \right\}.$$

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 basises 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

$$v = \sum_{i=1}^n \alpha_i v_i$$

where $\{v_1, \dots, v_n\}$ is a basis for V , we call the set $\alpha_1, \dots, \alpha_n$ v 's **coordinates** with respect to the basis $\{v_1, \dots, v_n\}$.

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

$$V = U_1 \oplus \cdots \oplus U_k = \bigoplus_{i=1}^k U_i,$$

if an arbitrary vector $v \in V$ in exactly one way can be expressed like

$$v = u_1 + \cdots + u_k, \quad u_1 \in U_1, \dots, u_k \in U_k \tag{A-1}$$

This condition is equivalent to that for vectors $u_i \in U_i$ the following holds true

$$u_1 + \cdots + u_k = \mathbf{0} \Rightarrow u_1 = \cdots = u_k = \mathbf{0}.$$

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 equation A-1. It is the unambiguousness 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

$$V = U_1 \oplus \cdots \oplus U_k$$

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 .

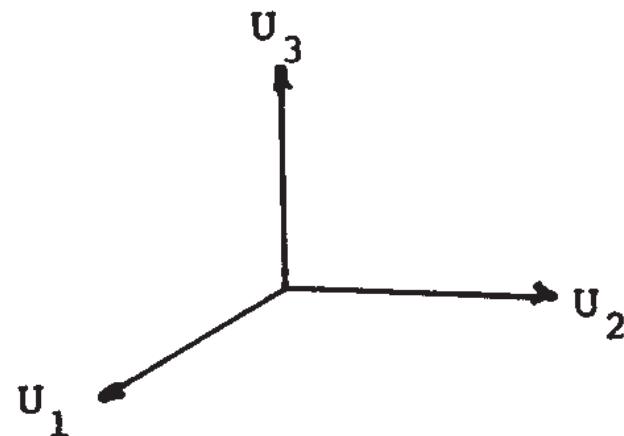
A.2 Linear transformations and matrices

We start with a section on *linear transformations* (or *linear mappings*).

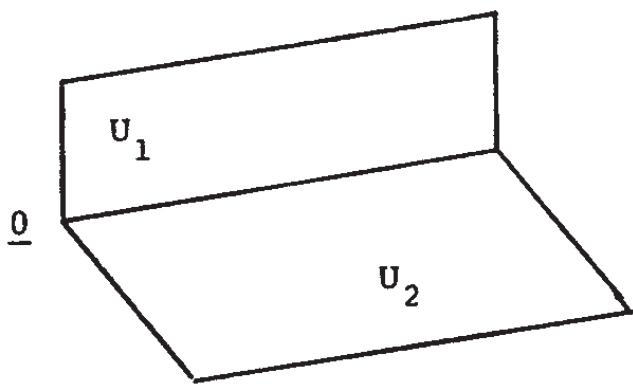
A.2.1 Linear transformations

A transformation (or mapping) $A : U \rightarrow V$, where U and V are vector spaces is said to be *linear* if

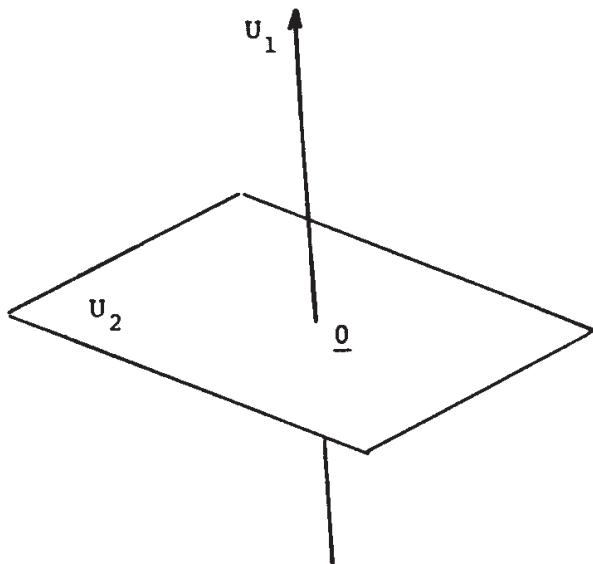
$$\begin{aligned} \forall \lambda_1, \lambda_2 \in R \quad & \forall \mathbf{u}_1, \mathbf{u}_2 \in U : A(\lambda_1 \mathbf{u}_1 + \lambda_2 \mathbf{u}_2) = \\ & \lambda_1 A(\mathbf{u}_1) + \lambda_2 A(\mathbf{u}_2) \end{aligned}$$



$U_1 \oplus U_2 \oplus U_3 = \mathbb{R}^3$ The sum is direct because for instance $\dim U_1 + \dim U_2 + \dim U_3 = 3$



\mathbb{R}^3 is not a direct sum of U_1 and U_2 ; because $\dim U_1 + \dim U_2 = 4$



Here $U_1 \oplus U_2 = \mathbb{R}^3$ because for instance U_1 and U_2 besides spanning \mathbb{R}^3 also satisfy $U_1 \cap U_2 = \mathbf{0}$

Figure A.2

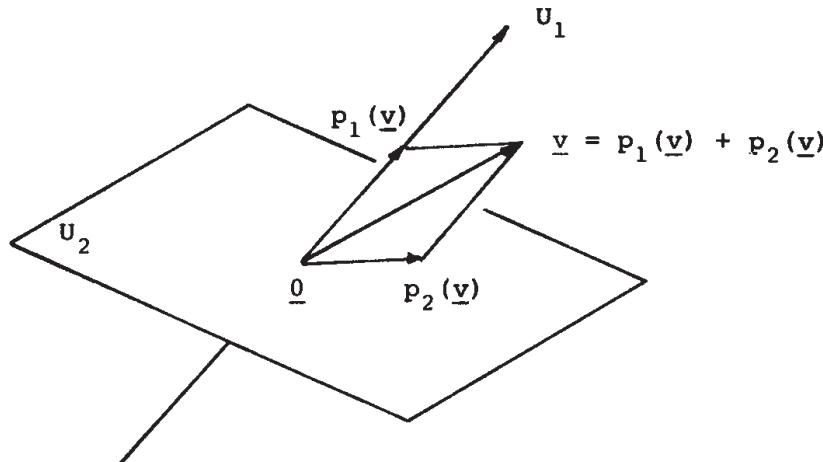
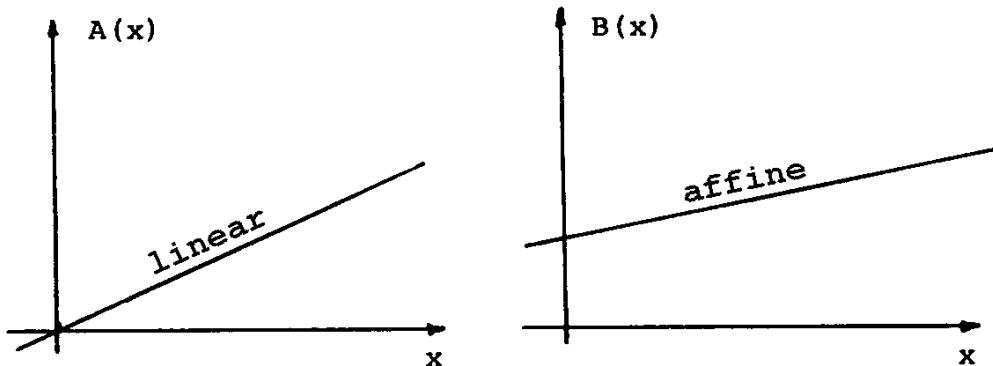


Figure A.3 – Projection of a vector.

||| Example A.4

A transformation $A : \mathbb{R} \rightarrow \mathbb{R}$ 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*.

Graphs for a linear and an affine transformation $\mathbb{R} \rightarrow \mathbb{R}$.

By the *null-space* $N(A)$ of a linear transformation $A : U \rightarrow V$ we mean the sub-space

$$A^{-1}(\mathbf{0}) = \{u | A(u) = \mathbf{0}\}$$

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 \mathbb{R}^n . In the following we will therefore often identify an n -dimensional vector space with \mathbb{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^T , i.e.

$$A^T = \begin{bmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{mn} \end{bmatrix}$$

An $m \times n$ matrix is *square* if $n = m$. A square matrix for which $A = A^T$ 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

$$\Delta = \mathbf{diag}(\delta_1, \dots, \delta_n) = \begin{bmatrix} \delta_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \delta_n \end{bmatrix}.$$

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

$$\begin{aligned} (A + B)^T &= A^T + B^T \\ (cA)^T &= cA^T \\ (CD)^T &= D^T C^T \end{aligned}$$

A.2.3 Linear transformations using matrix-formulation

It can be shown that for any *linear transformation* $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ there is a corresponding $m \times n$ matrix A , such that

$$\forall x \in \mathbb{R}^n : A(x) = Ax$$

Conversely an A defined by this relation is a linear transformation. A is easily found as the matrix which as columns has the coordinates of the transformation of the unit vectors in \mathbb{R}^n . E.g. we have

$$A e_2 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} = a_2$$

If we also have a linear transformation $B : \mathbb{R}^m \rightarrow \mathbb{R}^k$ with corresponding matrix B ($k \times m$), then we have that $B \circ A \leftrightarrow BA$ i.e.

$$\forall x \in \mathbb{R}^n : B \circ A(x) = B(A(x)) = BAx$$

Here we note, that an $n \times n$ matrix 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 A^{-1} , which satisfies

$$A A^{-1} = A^{-1} A = I$$

where I is the identity matrix of order n . Furthermore we have

$$\begin{aligned}(A^{-1})^{-1} &= A \\ (kA)^{-1} &= \frac{1}{k} A^{-1} \\ (A^T)^{-1} &= (A^{-1})^T\end{aligned}$$

and for invertible matrices A, B we have

$$(AB)^{-1} = B^{-1} A^{-1}$$

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

$$A A = A$$

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.

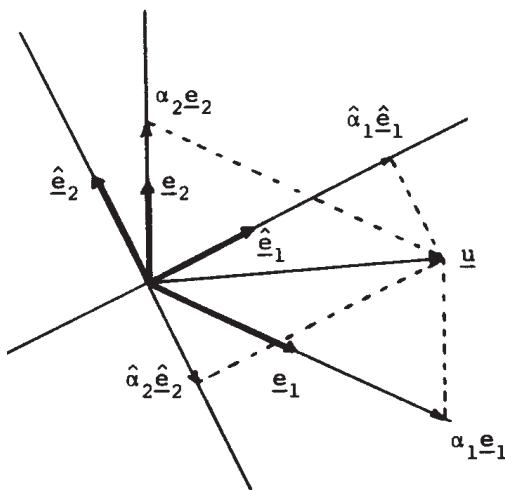
We first consider the change of coordinates going from one coordinate system to another. Normally, we choose not to distinguish between a vector \mathbf{u} and its set of coordinates. This gives a simple notation and does not lead to confusion. However, when several coordinate systems are involved we do need to be able to make this distinction. In \mathbb{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 \mathbf{u} in each of the two coordinate systems is denoted respectively $(\alpha_1, \dots, \alpha_n)^T$ and $(\hat{\alpha}_1, \dots, \hat{\alpha}_n)^T$, 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.

$$\hat{e}_i = s_{1i} e_1 + \dots + s_{ni} e_n, \quad i = 1, \dots, n.$$



$$, u = \begin{cases} \alpha_1 e_1 + \alpha_2 e_2 = [e_1 \ e_2] \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \\ \hat{\alpha}_1 \hat{e}_1 + \hat{\alpha}_2 \hat{e}_2 = [\hat{e}_1 \ \hat{e}_2] \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \end{cases}$$

Figure A.4 – Sketch of the coordinate transformation problem.

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 $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$, 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 $\mathbb{R}^n \rightarrow \mathbb{R}^n$ and we use the same coordinate transformation, then we get the relation

$$\hat{A} = S^{-1}AS.$$

The matrices A and $\hat{A} = S^{-1}AS$ are then called *similar matrices*.

A.2.5 Rank of a matrix

By *rank of a linear projection* $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ we mean the dimension of the image space, i.e.

$$\text{rk}(A) = \text{rank}(A) = \dim A(\mathbb{R}^n).$$

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

$$\text{rk}(A) \leq n.$$

If we introduce the transposed matrix A^T it is easily shown that $\text{rk}(A) = \text{rk}(A^T)$ i.e. we have

$$\text{rk}(A) \leq \min(m, n).$$

If A and B are two $m \times n$ matrices, then

$$\text{rk}(A + B) \leq \text{rk}(A) + \text{rk}(B).$$

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

$$\text{rk}(BA) \leq \text{rk}(A).$$

If B is regular $(m \times m)$ we have

$$\text{rk}(BA) = \text{rk}(A).$$

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 :

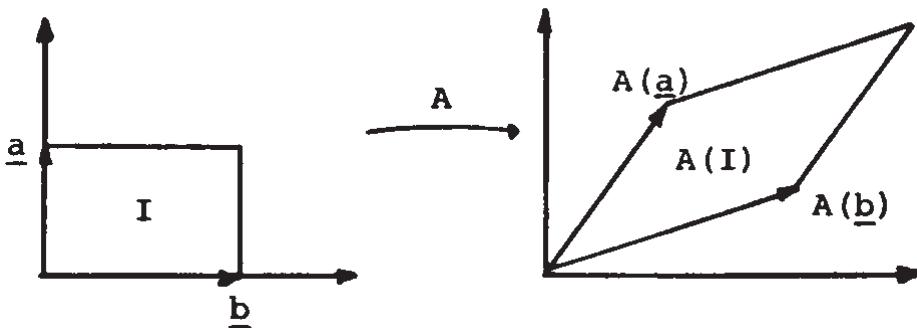


Figure A.5 – A rectangle and its image after a linear projection.

$$\text{rk}(A C) \leq \text{rk}(A)$$

with equality if C is a regular $(n \times n)$ -matrix. From these we can deduce for regular B and C that

$$\text{rk}(B A C) = \text{rk}(A).$$

Finally we mention that an $(n \times n)$ -matrix A is regular if $\text{rk}(A) = n$.

A.2.6 Determinant of a matrix

The abstract definition of the *determinant* of a square $p \times p$ matrix A is

$$\det(A) = \sum_{\text{all } \sigma} \pm a_{1\sigma(1)} \cdots a_{p\sigma(p)},$$

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

$$\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = aei + bfg + cdh - gec - hfa - idb.$$

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 determinant of matrix A_{ij} is called the **minor** of the matrix 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 is 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

$$\det(A^T) = \det(A)$$

For square matrices A and B we have

$$\det(A B) = \det(A) \det(B)$$

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

$$\det(C) = c_1 \cdots c_n$$

By means of determinants one can directly state the inverse of a regular matrix A . We have

$$A^{-1} = \frac{1}{\det(A)} (A_{ij})^T,$$

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

$$Ax = b$$

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 A.8

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.9

For the symmetrical matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

we have

$$\Sigma^{-1} = \begin{bmatrix} B^{-1} & -B^{-1}A^T \\ -AB^{-1} & \Sigma_{22}^{-1} + AB^{-1}A^T \end{bmatrix},$$

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

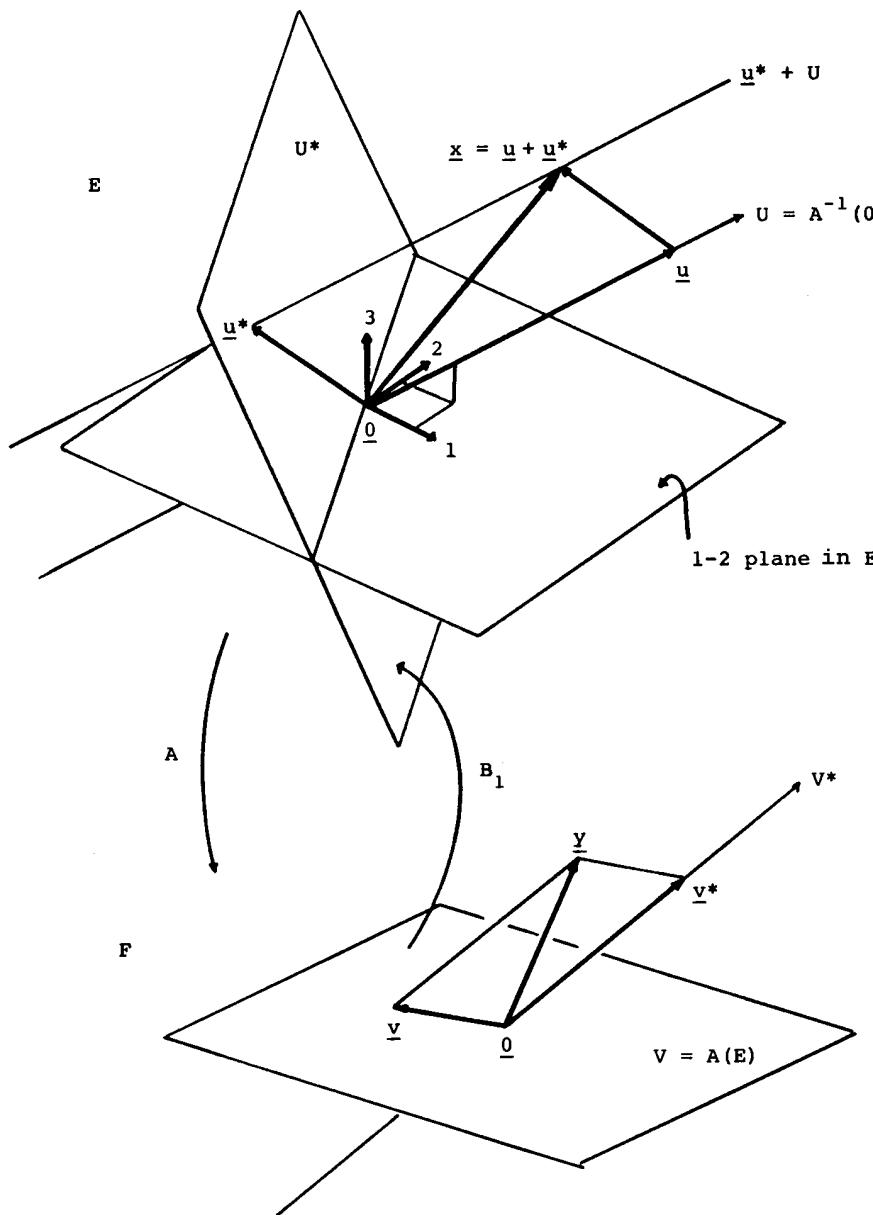
$$V = A(E)$$

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

$$x = u + u^*, \quad u \in U \quad \text{og} \quad u^* \in U^*,$$



10

Figure A.6 – Sketch showing pseudoinverse transformation.

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

$$\mathbf{u}^* + U = \{\mathbf{u}^* + \mathbf{u} \mid \mathbf{u} \in U\}$$

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

$$B_1(\mathbf{v}) = \mathbf{u}^* \Leftrightarrow A(\mathbf{u}^*) = \mathbf{v}$$

We are now able to formulate the definition of the pseudoinverse transformation.

||| Definition A.10

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.11

The pseudoinverse is thus the combined transformation onto V along V^* and the inverse of A 's restriction to U^* .

||| Remark A.12

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.13

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.14

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.15

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.16

Let A and B be as in definition A.15. The general solution of the equation

$$A x = \mathbf{0}$$

is

$$(I - B A) z, \quad z \in \mathbb{R}^n,$$

and the general solution of the equation (which is assumed to be consistent)

$$A x = y,$$

is

$$B y + (I - B A) z, \quad z \in \mathbb{R}^n.$$

||| 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 $B A$ 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

$$x = (I - BA)z, \quad z \in \mathbb{R}^n.$$

The statement regarding the homogeneous equation has now been proved.

The equation $Ax = y$ 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.13.

The result for the complete solution follows readily.

■

In order to illustrate the concept we now give

||| Example A.17

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

$$\begin{aligned} \mathbf{x} \in U &\Leftrightarrow A\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow x_1 + x_2 + 2x_3 = 0 \quad \wedge \quad 2x_1 + x_2 + x_3 = 0 \\ &\Leftrightarrow x_1 = x_3 \quad \wedge \quad -3x_1 = x_2 \\ &\Leftrightarrow \mathbf{x}^T = x_1(1, -3, 1). \end{aligned}$$

The null-space is then

$$U = \left\{ t \cdot \begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix} \mid t \in \mathbb{R} \right\} = \{t \cdot \mathbf{u}_3 \mid t \in \mathbb{R}\}$$

As complementary sub-space we choose to consider the orthogonal complement U^* . This has the equation

$$(1, -3, 1)\mathbf{x} = 0,$$

or

$$U^* = \{\mathbf{x} \mid x_1 - 3x_2 + x_3 = 0\}$$

We now consider a new basis for E , namely $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. 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

$$\mathbf{x} = S\mathbf{z}.$$

The columns of the S matrix are known to be the \mathbf{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

$$\mathbf{y} = T\mathbf{w}.$$

We will now find coordinate expressions for A in z - and w -coordinates. Since

$$\mathbf{y} = A\mathbf{x}$$

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

$$z = \mathbf{S}^{-1} \mathbf{x} = \mathbf{C} \mathbf{w} = \mathbf{C} \mathbf{T}^{-1} \mathbf{y}$$

we get

$$\mathbf{x} = \mathbf{S} \mathbf{C} \mathbf{T}^{-1} \mathbf{y},$$

where \mathbf{C} is the matrix in formula A-1.

We therefore have

$$\begin{aligned} \mathbf{B} &= \mathbf{S} \mathbf{C} \mathbf{T}^{-1} \\ &= \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{3}{22} & \frac{1}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= \frac{1}{22} \begin{bmatrix} -8 & 7 & 7 \\ 2 & 1 & 1 \\ 14 & -4 & -4 \end{bmatrix} \end{aligned}$$

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.18

Let the $m \times n$ matrix A have rank s and let

$$\mathbf{A} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix},$$

where \mathbf{C} is regular with dimension $s \times s$. A (possible) pseudoinverse of A is then

$$\mathbf{A}^- = \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

where the 0-matrices have dimensions such that \mathbf{A}^- has the dimension $n \times m$.

||| Proof

We have

$$\mathbf{A} \mathbf{A}^- \mathbf{A} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{E} \mathbf{C}^{-1} \mathbf{D} \end{bmatrix}.$$

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

$$\begin{bmatrix} \mathbf{D} \\ \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ \mathbf{E} \end{bmatrix} \mathbf{H}$$

or

$$\begin{aligned} \mathbf{D} &= \mathbf{C} \mathbf{H} \\ \mathbf{F} &= \mathbf{E} \mathbf{H} \end{aligned}$$

From this we find

$$\mathbf{F} = \mathbf{E} \mathbf{C}^{-1} \mathbf{D}.$$

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 449.

■

We illustrate the use of the theorem in the following

||| Example A.19

We consider the matrix given in example A.17

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

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.17 instead of the far more simple one given in example A.19, is that one obtains a precise geometrical description of the situation.

||| Remark A.20

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) $(A A^+)^T = A A^+$
- iv) $(A^+ A)^T = A^+ A$

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. [Rao and Mitra \(1971\)](#).

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.21

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. $x_i^T x_j = 0$.

||| 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

$$\begin{aligned} x_j^T A x_i &= \lambda_i x_j^T x_i \\ x_i^T A x_j &= \lambda_j x_i^T x_j. \end{aligned}$$

We transpose the first relationship and get

$$x_i^T A^T x_j = \lambda_i x_i^T x_j.$$

Since A is symmetric this implies that

$$\lambda_i x_i^T x_j = \lambda_j x_i^T x_j,$$

and since $\lambda_i \neq \lambda_j$ then $x_i^T x_j = 0$ i.e. $x_i \perp x_j$.

■

The result in theorem A.21 can be supplemented with the following theorem given without proof.

||| Theorem A.22

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.23

For an arbitrary symmetric matrix A a basis exists for \mathbb{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

$$P^T P = I$$

P is therefore by definition an *orthogonal matrix*, and

$$A P = P \Lambda$$

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.24

Let A be a symmetric matrix. Then an orthogonal matrix P exists, such that

$$P^T A P = \Lambda$$

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.25

Let A be a symmetric matrix with non-negative eigenvalues. Then a regular matrix B exists such that

$$B^T A B = E,$$

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.24 is equivalent to

$$\mathbf{A} = \mathbf{P} \Lambda \mathbf{P}^T$$

or

$$\mathbf{A} = (\mathbf{p}_1 \dots \mathbf{p}_n) \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{p}_1^T \\ \vdots \\ \mathbf{p}_n^T \end{bmatrix},$$

i.e. we have the following partitioning of the matrix

$$\mathbf{A} = \lambda_1 \mathbf{p}_1 \mathbf{p}_1^T + \cdots + \lambda_n \mathbf{p}_n \mathbf{p}_n^T.$$

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

$$\mathbf{A} = (\mathbf{P} \Lambda^{\frac{1}{2}})(\mathbf{P} \Lambda^{\frac{1}{2}})^T = \mathbf{G} \mathbf{G}^T.$$

||| Remark A.26

Here we mention that if \mathbf{A} is positive definite, then there is a relation

$$\mathbf{A} = \mathbf{L} \mathbf{L}^T,$$

where \mathbf{L} is a lower triangular matrix. This relation is called the *Cholesky decomposition* of \mathbf{A} (see e.g. Wilkinson (1961)). The decomposition is unique.

Finally we have

||| Theorem A.27

Let A be a regular symmetrical matrix. Then A and A^{-1} have the same eigenvectors corresponding to reciprocal eigenvalues.

||| Proof

Let λ be an eigenvalue of A and x be a corresponding eigenvector, i.e.

$$Ax = \lambda x.$$

Since A is regular then this is equivalent to

$$A^{-1}x = \frac{1}{\lambda}x,$$

which concludes the proof. ■

Finally, we note that

$$\det(A) = \prod_i \lambda_i.$$

||| Example A.28

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.

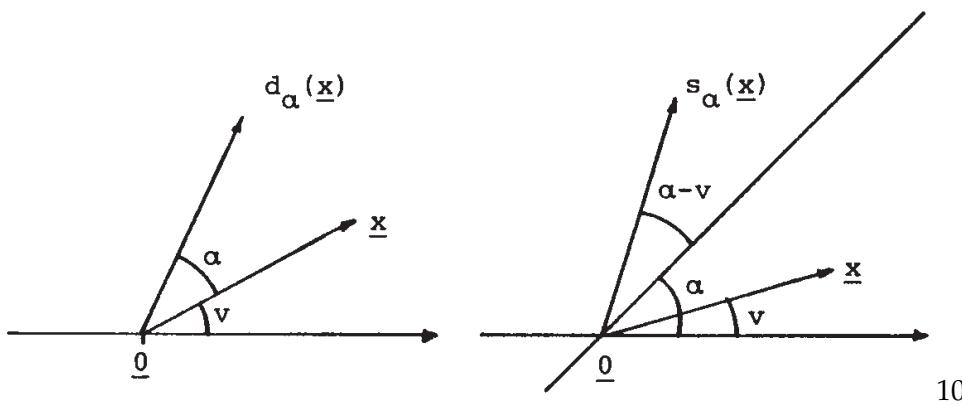
By utilising the orthogonality conditions $P^T P = I$ we readily see, that the only orthogonal 2×2 -matrices are matrices of the form

$$\begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} \cos \alpha & \sin \alpha \\ \sin \alpha & -\cos \alpha \end{bmatrix}.$$

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.28.



Rotation and reflection as determined by the angle α .

Since $x = r(\cos v, \sin v)'$, where r is equal to 1, we have

$$\begin{aligned} d_\alpha(x) &= \begin{bmatrix} \cos(\alpha + v) \\ \sin(\alpha + v) \end{bmatrix} = \begin{bmatrix} \cos \alpha \cos v - \sin \alpha \sin v \\ \sin \alpha \cos v + \cos \alpha \sin v \end{bmatrix} \\ &= \begin{bmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}. \end{aligned}$$

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

$$\begin{aligned} s_\alpha(x) &= \begin{bmatrix} \cos(2\alpha - v) \\ \sin(2\alpha - v) \end{bmatrix} = \begin{bmatrix} \cos 2\alpha \cos v + \sin 2\alpha \sin v \\ \sin 2\alpha \cos v - \cos 2\alpha \sin v \end{bmatrix} \\ &= \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}. \end{aligned}$$

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.29

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

$$x = V \Gamma U^T = [v_1 \dots v_r] \begin{bmatrix} \gamma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \gamma_r \end{bmatrix} \begin{bmatrix} u_1^T \\ \vdots \\ u_r^T \end{bmatrix} = \gamma_1 v_1 u_1^T + \dots + \gamma_r v_r u_r^T,$$

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. [Johnson \(1963\)](#).

||| Remark A.30

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 xx^T ($n \times n$) and x^Tx ($p \times p$).

However, first we will state

||| Theorem A.31

For an arbitrary (real valued) matrix \mathbf{x} it holds that $\mathbf{x}^T \mathbf{x}$ and $\mathbf{x} \mathbf{x}^T$ have non-negative eigenvalues and

$$\text{rk}(\mathbf{x}^T \mathbf{x}) = \text{rk}(\mathbf{x} \mathbf{x}^T) = \text{rk}(\mathbf{x})$$

||| Proof

It suffices to prove the results for $\mathbf{x}^T \mathbf{x}$. It is obvious that $\mathbf{x}^T \mathbf{x}$ is symmetric, so an orthogonal matrix \mathbf{P} , exists such that

$$\mathbf{P}^T \mathbf{x}^T \mathbf{x} \mathbf{P} = \Lambda$$

i.e.

$$(\mathbf{x} \mathbf{P})^T (\mathbf{x} \mathbf{P}) = \Lambda.$$

By letting $\mathbf{x} \mathbf{P} = \mathbf{B} = (b_{ij})$, we find $\mathbf{B}^T \mathbf{B} = \Lambda$, i.e.

$$\lambda_i = \sum_j b_{ij}^2 > 0,$$

i.e. $\mathbf{x}^T \mathbf{x}$ has non-negative eigenvectors. Furthermore we see that

$$\begin{aligned} \text{rk}(\mathbf{x}^T \mathbf{x}) &= \text{card}(\lambda_i \neq 0) \\ &= \text{card}\{\text{columns } \mathbf{b}_j \text{ in } \mathbf{B}, \text{ which are } \neq \mathbf{0}\} \end{aligned}$$

Since $\mathbf{b}_i^T \mathbf{b}_j = 0$ for $i \neq j$ (due to equation A-1) we have

$$\text{rk}(\mathbf{x}^T \mathbf{x}) = \text{rk}(\mathbf{B})$$

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.32

Let Σ be symmetrical and positive definite. Then for an arbitrary matrix \mathbf{x} it holds that

$$\text{rk}(\mathbf{x}^T \Sigma^{-1} \mathbf{x}) = \text{rk}(\mathbf{x}),$$

under the condition that the involved products exist.

||| Proof

Since Σ^{-1} is also regular and positive definite, an orthogonal matrix P exists, such that

$$P^T \Sigma^{-1} P = \Lambda,$$

where Λ is a diagonal matrix. This implies

$$\Sigma^{-1} = P \Lambda P^T = P \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} P^T = P \Lambda^{\frac{1}{2}} (P \Lambda^{\frac{1}{2}})^T = B B^T.$$

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

$$x^T \Sigma^{-1} x = x^T B B^T x = (B^T x)^T B^T x,$$

i.e.

$$\text{rk}(x^T \Sigma^{-1} x) = \text{rk}(B^T x) = \text{rk}(x),$$

which concludes the proof. ■

Using the notation from theorem A.31 we have.

||| Theorem A.33

The matrix $x x^T$ ($n \times n$) 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 x . The corresponding eigenvectors are v_1, \dots, v_r .

Similarly $x^T x$ ($p \times p$) 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 $x x^T$ and $x^T x$ are therefore equal and the relationship between the corresponding eigenvectors is ($m = 1, \dots, r$)

$$v_m = \frac{1}{\gamma_m} x u_m \quad \text{and} \quad u_m = \frac{1}{\gamma_m} x^T v_m,$$

or in a more compact notation

$$V = x U \Gamma^{-1} \quad \text{og} \quad U = x^T V \Gamma^{-1}$$

||| Proof omitted

Follows by use of Eckart-Young's theorem.

||| Remark A.34

Analysis of the matrix $\mathbf{x}^T \mathbf{x}$ is called **R-mode** analysis and the analysis of $\mathbf{x} \mathbf{x}^T$ is called **Q-mode** analysis. These names originate from factor analysis, cf. chapter 6.3.

||| Remark A.35

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 $\mathbf{x}^T \mathbf{x}$ and $\mathbf{x} \mathbf{x}^T$ 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 \mathbf{A} we mean the mapping

$$\mathbf{x} \rightarrow \mathbf{x}' \mathbf{A} \mathbf{x} = \sum a_{ii} x_i^2 + 2 \sum_{1 < j} a_{ij} x_i x_j.$$

We say that a symmetrical matrix \mathbf{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

$$\forall \mathbf{x} \neq \mathbf{0} : \mathbf{x}' \mathbf{A} \mathbf{x} > 0,$$

respectively

$$\forall \mathbf{x} \neq \mathbf{0} : \mathbf{x}' \mathbf{A} \mathbf{x} \geq 0.$$

We then also say the quadratic form is **positive definite** respectively **positive semi-definite**.

We have the following

||| Theorem A.36

The symmetrical matrix \mathbf{A} is positive definite respectively semi-definite, if all \mathbf{A} 's eigenvalues are positive respectively non-negative.

||| Proof

With \mathbf{P} as in theorem A.24 we have

$$\begin{aligned} \mathbf{x}' \mathbf{A} \mathbf{x} &= \mathbf{x}' \mathbf{P}' \mathbf{P} \mathbf{A} \mathbf{P} \mathbf{P}' \mathbf{x} = (\mathbf{P}' \mathbf{x})' \mathbf{A} (\mathbf{P}' \mathbf{x}) \\ &= \mathbf{y}' \mathbf{\Lambda} \mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2. \end{aligned}$$

Another useful result is

||| Theorem A.37

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.38

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 $R(x)$ as follows, then $R(x)$ is called *Rayleigh's coefficient* or *quotient*

$$R(x) = \frac{x^T A x}{x^T x},$$

and

$$M_k = \{x | x^T p_i = 0, \quad i = 1, \dots, k-1\}.$$

Then it holds that

$$\begin{aligned} \sup_x R(x) &= R(p_1) = \lambda_1, \\ \inf_x R(x) &= R(p_n) = \lambda_n, \\ \sup_{x \in M_k} R(x) &= R(p_k) = \lambda_k. \end{aligned}$$

The coefficient above should not be confused with the *general Rayleigh quotient* defined in theorem A.48.

||| Proof

An arbitrary vector \mathbf{x} can be written

$$\mathbf{x} = \alpha_1 \mathbf{p}_1 + \cdots + \alpha_n \mathbf{p}_n.$$

If $\mathbf{p}_i^T \mathbf{x} = 0$, $i = 1, \dots, k-1$, we find $\alpha_1 = \cdots = \alpha_{k-1} = 0$, i.e.

$$\mathbf{x} = \alpha_k \mathbf{p}_k + \cdots + \alpha_n \mathbf{p}_n.$$

Therefore we have

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \alpha_k^2 \lambda_k + \cdots + \alpha_n^2 \lambda_n,$$

and

$$R(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{A} \mathbf{x}}{\mathbf{x}^T \mathbf{x}} = \frac{\alpha_k^2 \lambda_k + \cdots + \alpha_n^2 \lambda_n}{\alpha_k^2 + \cdots + \alpha_n^2}$$

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.39

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.

We will now describe the level sets for positive definite forms.

||| Theorem A.40

Let \mathbf{A} be positive definite. Then the set of solutions for the equation

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = c, \quad c > 0,$$

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

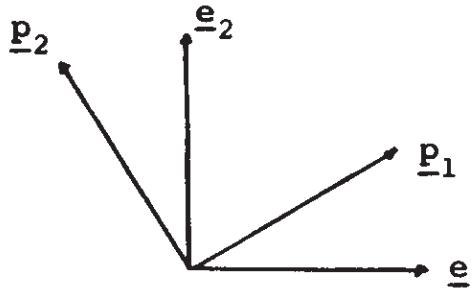
We consider the matrix $P = (p_1, \dots, p_n)$, whose columns are the coordinates of orthonormalized eigenvectors of A . Assuming $y = P^T x$ the following holds

$$\begin{aligned} x^T A x &= y^T \Lambda y \\ &= \lambda_1 y_1^2 + \dots + \lambda_n y_n^2 \\ &= \frac{y_1^2}{(1/\sqrt{\lambda_1})^2} + \dots + \frac{y_n^2}{(1/\sqrt{\lambda_n})^2} \end{aligned} \quad (\text{A-4})$$

The matrix equation

$$y = P^T x \Leftrightarrow x = P y$$

corresponds to a change of basis from the original orthonormal basis $\{e_1, \dots, e_n\}$ to the orthonormal basis $\{p_1, \dots, p_n\}$.



- $S = \begin{cases} (p_1, \dots, p_n) - \text{coordinates } y \\ (e_1, \dots, e_n) - \text{coordinates } x \end{cases}$

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.

$$Ix = Py,$$

where I is a unit matrix.

The expression in equation A-4 therefore shows the equation of the set of solutions in y -coordinates corresponding to the coordinate system consisting of orthonormalized 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.41

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

$$\begin{aligned} \mathbf{x}^T \mathbf{A} \mathbf{x} = c &\Leftrightarrow \lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 + 0y_{r+1}^2 + \cdots + 0y_n^2 = c \\ &\Leftrightarrow \lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 = c. \end{aligned}$$

This leads to the statement. If we consider the restriction of the quadratic form to the subspace spanned by the eigenvectors corresponding to eigenvalues > 0 , then the set of solutions becomes an ellipsoid.

||| Example A.42

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

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = 3x_1^2 + 2x_2^2 + 2\sqrt{2}x_1x_2,$$

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

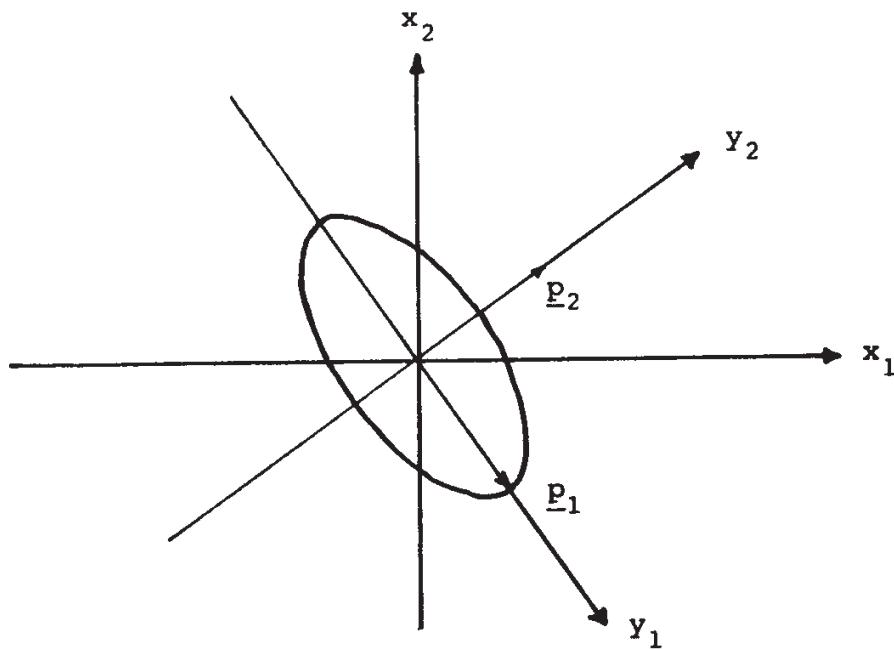
$$\begin{aligned} (\mathbf{A} - \lambda \mathbf{I}) = 0 &\Leftrightarrow \lambda^2 - 5\lambda + 4 = 0 \\ &\Leftrightarrow \lambda = 1 \quad \vee \quad \lambda = 4. \end{aligned}$$

Eigenvectors 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.42. The ellipse has the equation

$$\frac{y_1^2}{1^2} + \frac{y_2^2}{\frac{1}{2}^2} = 1.$$

It is illustrated in the figure above

Since

$$\begin{aligned} p_1 &= \begin{bmatrix} \frac{\sqrt{3}}{3} \\ -\frac{\sqrt{6}}{3} \end{bmatrix} = \begin{bmatrix} 0.577 \\ -0.820 \end{bmatrix} \\ &\approx \begin{bmatrix} \cos(-54.7^\circ) \\ \sin(-54.7^\circ) \end{bmatrix}, \end{aligned}$$

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.43

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

$$A x = \lambda B x.$$

Such a vector x is called an *eigenvector for A w.r.t. B* .

||| Remark A.44

The concepts given above can be traced back to eigenvalues and eigenvectors for the *non-symmetrical matrix $B^{-1}A$* .

||| Theorem A.45

We consider again the situation in the definition A.43 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.25 there is a regular matrix T where

$$T^T B T = I.$$

Let

$$D = T^T A T$$

D is obviously symmetrical, and since

$$x^T D x = (T x)^T A (T x),$$

we see that D and A are at the same time respectively positive semi-definite and positive definite.

Now we have

$$\begin{aligned} (D - \lambda I)v = 0 &\Leftrightarrow (T^T A T - \lambda T^T B T)v = 0 \\ &\Leftrightarrow (A - \lambda B)(T v) = 0 \end{aligned}$$

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.46

Let the situation be as above. Then a basis exists for \mathbb{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.

$$\mathbf{u}_i^T \mathbf{A} \mathbf{u}_j = \mathbf{u}_i^T \mathbf{B} \mathbf{u}_j = 0.$$

||| Proof

Follows from the proof of the above theorem and of the corollary to theorem A.22, remembering that

$$0 = \mathbf{v}_i^T \mathbf{v}_j = (\mathbf{v}_i^T \mathbf{T}^T) \mathbf{T}^{T-1} \mathbf{T}^{-1}(\mathbf{T} \mathbf{v}_j) = \mathbf{u}_i^T \mathbf{B} \mathbf{u}_j,$$

where $\mathbf{v}_1, \dots, \mathbf{v}_m$ is an orthonormal basis for \mathbb{R}^m consisting of eigenvectors of \mathbf{D} .

Finally we have

$$\mathbf{u}_i^T \mathbf{A} \mathbf{u}_j = \lambda_j \mathbf{u}_i^T \mathbf{B} \mathbf{u}_j = 0$$

■

||| Theorem A.47

Let \mathbf{A} be symmetrical and let \mathbf{B} be positive definite. Then a regular matrix \mathbf{R} exists with

$$\mathbf{R}^T \mathbf{A} \mathbf{R} = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_n),$$

and

$$\mathbf{R}^T \mathbf{B} \mathbf{R} = \mathbf{I},$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of \mathbf{A} w.r.t. \mathbf{B} . If the i 'th column in $(\mathbf{R}^T)^{-1}$ is termed \mathbf{s}_i then these relations can be written

$$\mathbf{A} = \lambda_1 \mathbf{s}_1 \mathbf{s}_1^T + \dots + \lambda_m \mathbf{s}_m \mathbf{s}_m^T,$$

and

$$\mathbf{B} = \mathbf{s}_1 \mathbf{s}_1^T + \dots + \mathbf{s}_m \mathbf{s}_m^T.$$

||| Proof

From the proof of theorem A.45 we consider the $\mathbf{D} = \mathbf{T}^T \mathbf{A} \mathbf{T}$. Since \mathbf{D} is symmetrical, according to theorem A.24 there exists an orthogonal matrix \mathbf{C} with

$$\mathbf{C}^T \mathbf{D} \mathbf{C} = \Lambda,$$

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

$$R^T B R = C^T T^T B T C = C^T C = I,$$

and

$$R^T A R = C^T T^T A T C = C^T D C = \Lambda.$$

■

Finally we state an analogue of theorem A.38 in the following

||| Theorem A.48

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 $v_i^T B v_j = 0 \quad i \neq j$. We define the *general Rayleigh quotient*

$$R(x) = \frac{x^T A x}{x^T B x}$$

and put

$$M_k = \{x | x^T B v_1 = \dots = x^T B v_{k-1} = 0\}.$$

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 $v_i^T B v_i = 1$, and since an arbitrary vector x can be written

$$x = \alpha_1 v_1 + \dots + \alpha_m v_m,$$

we find

$$R(x) = \frac{\sum \alpha_i^2 v_i^T A v_i}{\sum \alpha_i^2 v_i^T B v_i} = \frac{\sum \lambda_i \alpha_i^2}{\sum \alpha_i^2}.$$

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

$$\text{tr}(\mathbf{A}^T) = \text{tr}(\mathbf{A}).$$

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 equation A-5 and theorem A.24

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$. 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.49

For a symmetrical $(n \times n)$ -matrix \mathbf{A} and an arbitrary n -dimensional vector \mathbf{b} it holds that

- i) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^T \mathbf{x}) = \mathbf{b}$
- ii) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T \mathbf{x}) = 2\mathbf{x}$
- iii) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2\mathbf{A} \mathbf{x}$.

|||| 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

$$(\mathbf{x} + \mathbf{h})^T \mathbf{A} (\mathbf{x} + \mathbf{h}) = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{h}^T \mathbf{A} \mathbf{h} + 2\mathbf{h}^T \mathbf{A} \mathbf{x}$$

By choosing $\mathbf{h} = (0, \dots, h, \dots, 0)^T$ we see that

$$\frac{\partial}{\partial x_i}(\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2 \sum_{j=1}^h a_{ij} x_j,$$

and the result follows readily.

■

We will illustrate the use of the theorem in the following

||| Example A.50

We want to find the minimum of the function

$$g(\theta) = (\mathbf{y} - \mathbf{A}\theta)^T \mathbf{B} (\mathbf{y} - \mathbf{A}\theta),$$

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

$$\begin{aligned} g(\theta) &= \mathbf{y}^T \mathbf{B} \mathbf{y} - \theta^T \mathbf{A}^T \mathbf{B} \mathbf{y} + \theta^T \mathbf{A}^T \mathbf{B} \mathbf{A} \theta - \mathbf{y}^T \mathbf{B} \mathbf{A} \theta \\ &= \mathbf{y}^T \mathbf{B} \mathbf{y} - 2\mathbf{y}^T \mathbf{B} \mathbf{A} \theta + \theta^T \mathbf{A}^T \mathbf{B} \mathbf{A} \theta. \end{aligned}$$

Here we have used that

$$\theta^T \mathbf{A}^T \mathbf{B} \mathbf{y} = \mathbf{y}^T \mathbf{B} \mathbf{A} \theta$$

(both 1×1 matrices, i.e. a scalar, and each others transposed). From this follows that

$$\frac{\partial g}{\partial \theta} = -2\mathbf{A}^T \mathbf{B} \mathbf{y} + 2\mathbf{A}^T \mathbf{B} \mathbf{A} \theta,$$

and it is seen that

$$\frac{\partial g}{\partial \theta} = \mathbf{0} \leftrightarrow \mathbf{A}^T \mathbf{B} \mathbf{A} \theta = \mathbf{A}^T \mathbf{B} \mathbf{y}.$$

This equation has as mentioned always at least one root. If $\mathbf{A}^T \mathbf{B} \mathbf{A}$ is regular then we have

$$\theta_{\min} = (\mathbf{A}^T \mathbf{B} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{B} \mathbf{y}.$$

If the matrix is singular, then we can write

$$\theta_{\min} = (\mathbf{A}^T \mathbf{B} \mathbf{A})^- \mathbf{A}^T \mathbf{B} \mathbf{y},$$

where $(\mathbf{A}^T \mathbf{B} \mathbf{A})^-$ denotes a pseudoinverse of $\mathbf{A}^T \mathbf{B} \mathbf{A}$.

We are now able to find an alternative description of the principle axes in an ellipsoid, due to

||| Theorem A.51

Let \mathbf{A} be a positive definite symmetrical matrix. The principle directions of the ellipsoid E_c with the equation

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = c, \quad c > 0$$

are those directions where $\mathbf{x}^T \mathbf{x}$, $\mathbf{x} \in E_c$, has stationary points.

||| Proof

We may assume that $x = 1$. We then need to find the stationary points for

$$f(x) = x^T x$$

with the condition that

$$x^T A x = 1$$

We apply a Lagrange multiplier technique and define

$$\varphi(x, \lambda) = x^T x - \lambda(x^T A x - 1).$$

By differentiation we obtain

$$\frac{\partial \varphi}{\partial x} = 2x - 2\lambda A x.$$

If this quantity is to equal 0, then

$$x = \lambda A x$$

or

$$A x = \frac{1}{\lambda} x,$$

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.52

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. Bourbaki (1967)). If this is introduced in coordinate form then we can either use equation 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.

||| Remark A.53

Tensor product rules.

- i) $\mathbf{0} \otimes A = A \otimes \mathbf{0} = \mathbf{0}$
- ii) $(A_1 + A_2) \otimes B = A_1 \otimes B + A_2 \otimes B$
- iii) $A \otimes (B_1 + B_2) = A \otimes B_1 + A \otimes B_2$
- iv) $\alpha A \otimes \beta B = \alpha\beta A \otimes B$
- v) $A_1 A_2 \otimes B_1 B_2 = (A_1 \otimes B_1)(A_2 \otimes B_2)$
- vi) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$ if the inverses exist
- vii) $(A \otimes B)^- = A^- \otimes B^-$
- viii) $(A \otimes B)^T = A^T \otimes B^T$
- ix) Let A be symmetrical and $p \times p$, have eigenvalues $\alpha_1, \dots, \alpha_p$ and eigenvectors x_i , and let B , be symmetrical and $q \times q$, have eigenvalues β_1, \dots, β_q and eigenvectors y_1, \dots, y_q . Then $A \otimes B$ will have the eigenvalues $\alpha_i \beta_j$, $i = 1, \dots, p$, $j = 1, \dots, q$, with corresponding eigenvectors.

$$\mathbf{x}_i \otimes \mathbf{y}_j = \begin{bmatrix} x_{1i}y_j \\ \vdots \\ x_{pi}y_j \end{bmatrix}$$

- x) $\det(A \otimes B) = (\det A)^q (\det B)^p$

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

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = (x_1 \dots x_n) \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^n x_i y_i,$$

and we note that \mathbf{x} and \mathbf{y} are orthogonal if and only if

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}^T \mathbf{y} = 0.$$

The corresponding norm is

$$\|\underline{x}\| = (\underline{x} \cdot \underline{x})^{\frac{1}{2}} = (\underline{x}^T \underline{x})^{\frac{1}{2}} = \sqrt{x_1^2 + \cdots + x_n^2}$$

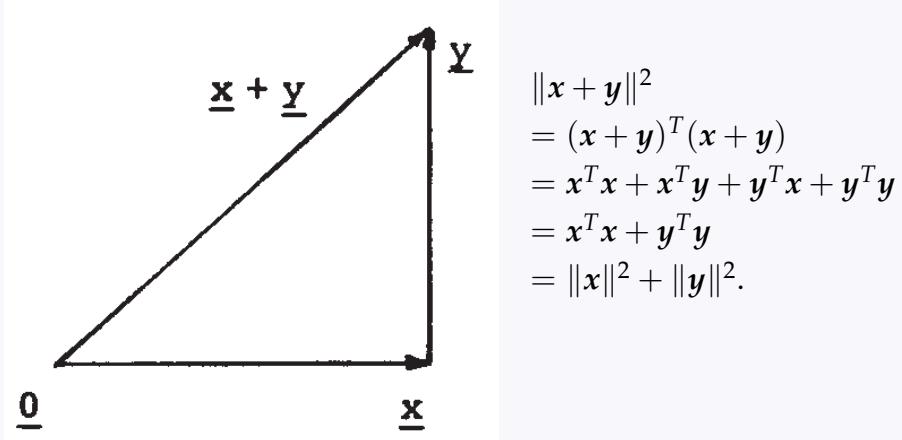
We note that $\|\underline{x} - \underline{y}\|$ represents the euclidian distance between the points \underline{x} and \underline{y} .

||| Theorem A.54

For orthogonal vectors \underline{x} and \underline{y} (i.e. $\underline{x} \perp \underline{y}$) we have the *pythagorean theorem*

$$\|\underline{x} + \underline{y}\|^2 = \|\underline{x}\|^2 + \|\underline{y}\|^2;$$

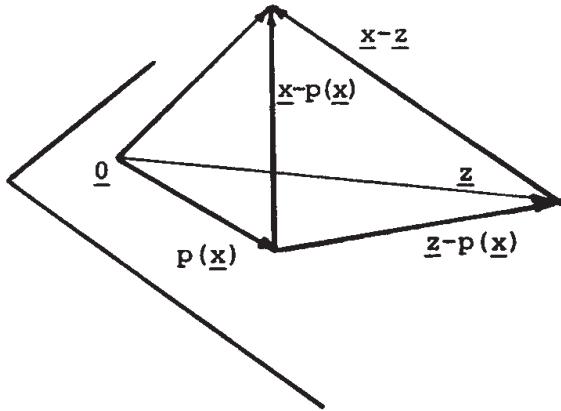
see the figure below



Further we note that the (orthogonal) projection $p(\underline{x})$ of a vector \underline{x} onto the sub-space U can be determined by means of the norm, since we have that $p(\underline{x})$ is given by

$$\|\underline{x} - p(\underline{x})\| = \min_{z \in U} \|\underline{x} - z\|$$

||| Proof



Due to the Pythagorean theorem we have that

$$\begin{aligned} \|x - p(x)\|^2 + \|z - p(x)\|^2 \\ = \|x - z\|^2, \end{aligned}$$

i.e. the minimal value of
 $\|x - z\|^2$, and therefore of
 $\|x - z\|$ is achieved for
 $z = p(x)$.

■

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 : $(x|y) = (y|x)$
- IP2 : $(x + y|z) = (x|z) + (y|z)$
- IP3 : $(kx|y) = k(x|y)$
- IP4 : $x \neq 0 \Rightarrow (x|x) > 0.$

For an arbitrary bi-linear form $(\cdot|\cdot)$, which satisfies the above one can define a concept of orthogonality by

$$x \perp y \quad \stackrel{d}{\Leftrightarrow} \quad (x|y) = 0.$$

For an arbitrary positive definite symmetrical matrix A we can define an inner product by

$$(x|y)_A = x^T A y.$$

It is trivial to prove that IP 1-4 are satisfied. for this inner product and the corresponding norm given by

$$\|x\|_A = \sqrt{(x|x)_A} = \sqrt{x^T A x},$$

we will - whenever it does not lead to confusion - use the terms $(x|y)$ and $\|x\|$.

We note that the set of points with constant A -norm equal to 1 is the set

$$\{x | \|x\|^2 = 1\} = \{x | x^T A x = 1\},$$

i.e. the points on an ellipsoid.

Conversely, to any non-degenerate ellipsoid there is a corresponding positive definite matrix A , so

$$E = \{x | x^T A x = 1\} = \{x | \|x\|_A^2 = 1\}.$$

In this way we have brought about a connection between the set of possible inner products and the set of ellipsoids.

Two vectors x and y are *orthogonal (with respect to A)*, if

$$x^T A y = 0,$$

i.e. if x and 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

$$\cos(\angle a, b) = \frac{(a|b)}{\|a\| \|b\|}.$$

We now give a lemma which we will need for the theorems of independence of projections of normally distributed stochastic variables.

||| Lemma A.55

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.

$$x \perp y \Leftrightarrow x^T \Sigma^{-1} y = 0.$$

For $i = 1, \dots, k$ we let the projection p_i onto U_i be given by the matrix C_i . Then

$$C_i \Sigma C_j^T = 0$$

for all $i \neq j$. Furthermore, we have

$$\Sigma^{-1} C_i = C_i^T \Sigma^{-1} = C_i^T \Sigma C_i.$$

||| 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

$$p_i(x)^T \Sigma^{-1} (x - p_i(x)) = 0,$$

i.e.

$$x C_i^T \Sigma^{-1} [x - C_i x] = 0.$$

This holds for all x , and therefore

$$C_i^T \Sigma^{-1} (I - C_i) = 0,$$

or

$$C_i^T \Sigma^{-1} = C_i^T \Sigma^{-1} C_i.$$

The right hand side of the equation is obviously symmetrical, so that

$$C_i^T \Sigma^{-1} = \Sigma^{-1} C_i.$$

By pre- and post-multiplication with Σ we get

$$\Sigma C_i^T = C_i \Sigma,$$

so

$$C_i \Sigma C_i^T = C_i C_i \Sigma = C_i \Sigma.$$

This gives

$$C_i \Sigma C_j^T = C_i \Sigma C_i^T C_j^T = C_i \Sigma \mathbf{0} = \mathbf{0}.$$

The second-last equal sign follows from the fact that the sum is direct, so for all x it holds that

$$p_j(p_i(x)) = \mathbf{0},$$

i.e.

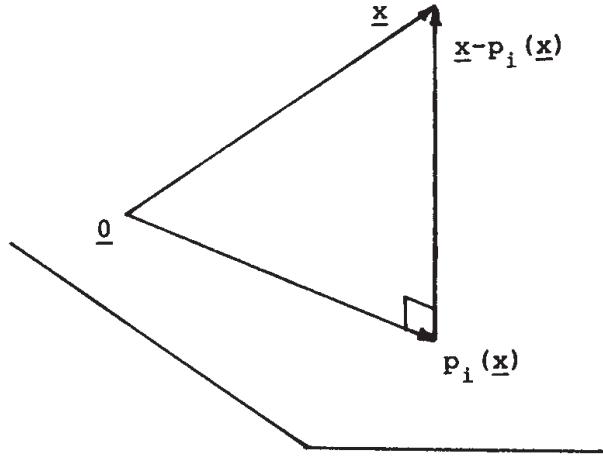
$$C_j C_i x = \mathbf{0}.$$

Since x - as was mentioned previously - is arbitrary, then this implies

$$C_j C_i = \mathbf{0},$$

or

$$C_i^T C_j^T = \mathbf{0}.$$



Bibliography

- Agterberg, F. P. (Amsterdam 1973). *Geomathematics. Mathematical background and geo-science applications*. Elsevier.
- Andersen, H. H., Hojbjerre, M., Sorensen, D., and Eriksen, P. S. (1995). *Linear and graphical models: for the multivariate complex normal distribution*, volume 101. Springer Science & Business Media.
- Anderson, T. W. (New York 1958). *An Introduction to Multivariable Statistical Analysis*. John Wiley & sons.
- Bjerre, T. (2013). *Automated Image-Based Procedures for Adaptive Radiotherapy*. PhD thesis.
- Bourbaki, N. (Paris 1967). *Algebre*, chapter 2 Algebre Lineaire. Hermann.
- Cattell, R. (1965). 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.
- Christensen, E. L., Skou, N., Dall, J., Woelders, K. W., Jorgensen, J. H., Granholm, J., and Madsen, S. N. (1998). Emisar: An absolutely calibrated polarimetric l-and c-band sar. *IEEE Transactions on Geoscience and Remote Sensing*, 36(6):1852–1865.
- Conradsen, K., A. N. B. N. n. E. J. P. T. T. (1987). The use of structural and spectral enhancement of remote sensing data in ore prospecting. east greenland case study. Technical report, Technical University of Denmark, IMSOR.
- Conradsen, K., Nielsen, A. A., Schou, J., and Skriver, H. (2003). A test statistic in the complex wishart distribution and its application to change detection in polarimetric sar data. *IEEE Transactions on Geoscience and Remote Sensing*, 41(1):4–19.
- Cortes, C. and Mohri, M. (2005). Confidence intervals for the area under the roc curve. In *Advances in neural information processing systems*, pages 305–312.
- Davies, O. L., editor (London 1967). *Design and Analysis of Industrial Experiments*. Oliver and Boyd, second edition.
- Davis, J. C. (New York 1973). *Statistics and data analysis in geology*. John Wiley.
- Dwyer, P. S. (1939). *The contribution of an orthogonal multiple facto solution to multiple correlation*, volume 4. John Wiley & Sons.

- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.
- Emerson, M. J., Wang, Y., Withers, P. J., Conradsen, K., Dahl, A. B., and Dahl, V. A. (2018). Quantifying fibre reorientation during axial compression of a composite through time-lapse x-ray imaging and individual fibre tracking. *Composites Science and Technology*, 168:47–54.
- Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874.
- Frisch, R. and Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica: Journal of the Econometric Society*, pages 387–401.
- Gallager, R. G. (2008). *Principles of digital communication. Supplementary Notes*. Technical Publications.
- Goodman, N. R. (1963). Statistical analysis based on a certain multivariate complex gaussian distribution (an introduction). *The Annals of mathematical statistics*, 34(1):152–177.
- Graybill, F. A. (1976). *Theory And Application of the Linear Model*, volume 1. Duxbury Press.
- Harman, H. H. (Chicago 1967). *Modern Factor Analysis*. The University of Chicago Press, second edition.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression. biased estimation for nonorthogonal problems. In *Technometrics*, volume 12, No.1, pages 55–67.
- Johnson, R. M. (1963). On a theorem stated by Eckart and Young. In *Psychometrika*, volume 28, pages 259–263.
- Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. In *Psychometrika*, volume 32.
- Kaiser, H. F. (1958). *The varimax criterion for analytic rotation in factor analysis*.
- Kendall, M. G. and Stuart, A. (London 1967). *The Advanced Theory of Statistics*, volume 2. Charles Griffin & Co.
- Knudsen, J. G. (1975). En statistisk analyse af cementstyrke. Eksamensprojekt, IMSOR, DTU.
- Lawley, D. N. (1940). *The estimation of factor loadings by the method of maximum likelihood*.
- Marquardt, D. W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. In *Technometrics*, volume 12, No.3, pages 591–612.
- McLachlan, G. (2004). *Discriminant analysis and statistical pattern recognition*, volume 544. John Wiley & Sons.
- Milliken, G. and Johnson, D. (1993). *Analysis of Messy Data: Designed Experiments*. Number vb. 1. Taylor & Francis.
- Morrison, D. F. (New York 1967). *Multivariate Statistical Methods*. McGraw-Hill.
- Ottosen, A. L. (2015). Personal communication.

- Pandey, S. and Elliott, W. (2010). Suppressor variables in social work research: Ways to identify in multiple regression models. *Journal of the Society for Social Work and Research*, 1(1):28–40.
- Pedersen, S. T. and Skjøth, P. (1976). Statistisk analyse af data fra cementfabrikation. Ek-samsensprojekt, IMSOR, DTU.
- Pedersen, S. T. and Skjøth, P. (1978). Statistical analysis of data from manufacturing. Technical report, Technical University of Denmark, IMSOR.
- Rao, C. R. (1955). *Estimation and tests of significance in factor analysis*.
- Rao, C. R. and Mitra, S. K. (New York 1971). *Generalized Inverse of Matrices and Its Applications*. John Wiley.
- Rao, C. R., Rao, C. R., Statistiker, M., Rao, C. R., and Rao, C. R. (1973). *Linear statistical inference and its applications*, volume 2. Wiley New York.
- Salomonsen, E. (København 1977). Fjernelse af klorider fra forhistorisk jern. Master's thesis, Konservatorskolen.
- Sjöstrand, K., Clemmensen, L. H., Larsen, R., Einarsson, G., and Ersbøll, B. K. (2018). Spasm: A matlab toolbox for sparse statistical modeling. *Journal of Statistical Software*, 84(10).
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.
- Trinderup, C. H., Dahl, A., Jensen, K., Carstensen, J. M., and Conradsen, K. (2015). Comparison of a multispectral vision system and a colorimeter for the assessment of meat color. *Meat science*, 102:1–7.
- Trinderup, C. H., Møller, F., Dahl, A. B., and Conradsen, K. (2018a). Investigation of pausing fermentation of salamis with multispectral imaging for optimal sensory evaluations. *Meat science*, 146:9–17.
- Trinderup, C. H., Møller, F., Dahl, A. B., and Conradsen, K. (2018b). Investigation of pausing fermentation of salamis with multispectral imaging for optimal sensory evaluations. *Meat science*, 146:9–17.
- Velicer, W. F. (1978). Suppressor variables and the semipartial correlation coefficient. *Educational and Psychological Measurement*, 38(4):953–958.
- Wahba, G. (1990). Spline models for observational data. *Society for Industrial and Applied Mathematics*.
- Wessel Lindberg, A.-S., Conradsen, K., Larsen, R., Friis Lippert, M., Røge, R., and Vyberg, M. (2017). Quantitative tumor heterogeneity assessment on a nuclear population basis. *Cytometry Part A*, 91(6):574–584.
- Wilkinson, J. H. (1961). Error analysis of direct methods of matrix inversion. *Journal of the Association of Computing Machinery*, 8:281–330.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.

Index

- accuracy, 45
- affine support, 14
- algebra, 429
- analysis of variance, 302
 - one-way, multidimensional, 302
 - two-way, multidimensional, 304
- Analysis of variance, table, 130
- Andersons U, 295, 296
- ANOVA, table, 130
- Autoregressive model, 163
- backwards elimination, 229
- Beta Distribution
 - complex, 77
- BLUE - Best Linear Unbiased Estimators, 104
- canonical
 - correlation, 382
 - pair of variables, 382
 - standardized coefficients, 383
- canonical correlation, 381
- canonical variable, 381
- central limit therorem, 24
- Cholesky decomposition, 458
- Circularly-symmetric, 67
- classification
 - area under curve, 353
 - AUC, 353
 - Bayes solution, 322
 - canonical discriminant analysis, 361
 - canonical discriminant functions, 361
 - CDA, 361
 - CDF, 361
 - confusion matrix, 347
 - cross validation, 347
 - decision function, 321
 - discriminant score, 336
 - discriminant value, 336
 - k-fold cross validation, 348
 - LDF, 338
- linear discriminant function, 338
- linear discriminant functions, 324
- linear discriminator, 325
- linear discriminator, distribution, 328
- loss function, 319
- minimax solution, 322
- posterior distribution, 321
- prior distribution, 321
- QDF, 338
- quadratic discriminant function, 338
- Receiver Operating Characteristic, 350
- resubstitution, 347
- ROC, 350
- set aside, 347
- column-vector, 436
- Complex Multivariate Beta distribution, 77
- complex random matrix, 61
- complex random variable, 60
- Complex Wishart Distribution, 72
- Compound symmetry, 162
- condition number, 208
- conditional distribution, 22, 46
- confidence region for
 - mean value, 282, 285
- confidenceinterval for
 - correlation-coefficient, 40
 - partial correlation-coefficient, 40
- confidenceintervals for
 - estimated value, 123
- conjugated vectors, 471
- constrained estimation, 117
- contour ellipsoid, 17, 21
- Cook's D, 204
- coordinate transformation, 438
- correlation
 - partial, 189
 - semi-partial, 196
 - squared multiple, 194
- correlation matrix, 7

- correlation-coefficient, 27, 29
correspondence analysis, 419
Covariance
 complex, 60
covariance, 8
Covariance Matrix
 Pseudo, 64
Covariance maximization, 424
covariance-matrix, 5
COVRATIO, 206
Cramér's theorem, 443
Cramér-Rao's inequality, 92
cross validation, 249

data-matrix, 25
deletion formula, 204
determinant, 441, 444
DFBETAS, 207
DFFITS, 206
diag, 436
differentiation of
 linear form, 474
 quadratic form, 474
dim, 432
direct sum, 480
Discrimination between two populations, 319
dispersion
 complex matrix, 62
dispersion matrix, 5
 pooled estimate, 333, 341
distance
 Mahalanobis', 333
 squared generalised from mean to population, 342
dot product, 477

Eckart-Young's theorem, 461
eigenvalue, 455, 463, 473
 multiplicity, 456
eigenvalue problem, general, 469
eigenvalue w.r.t. matrix, 470
eigenvector, 455, 463
 orthogonality, 456
eigenvector w.r.t. matrix, 470
Elastic net, 254
ellipsoid, 466
elliptical cylinder, 468
empirical generalised variance, 58

empirical partial correlation, 39
empirical variance-covariance matrix, 27
estimation of
 variance-covariance matrix, 25
estimation of/in
 factor scores, 412
eigenvalue of variance-covariance matrix,
 374
factor loadings, 405
 variance-covariance matrix, 278, 283, 292
euclidian distance, 478
Expectation
 complex, 60
expectation, 3
Expected value
 complex matrix, 61
expected value, 3

Factor analysis
 common factors, 403
 communalities, 404
 estimation of loadings, 405
 factor loadings, 403
 factor scores, 403
 maximum likelihood analysis, 416
 principle factor solution, 407
 Q-mode analysis, 419
 quartimax rotation, 408
 test for model, 418
 unique factors, 403
 uniqueness, 404
 Varimax rotation, 408
factor analysis, 402
forward selection, 230
functional relation, 273

Gauss-Markov's theorem, 104, 291
general linear model, 128
 multidimensional, 289
generalised variance, 54, 370
generalized variance, 58
geodesy, 122
GLM
 confidence interval, 124
 intercept, 146
 least squares estimate, 104
 no intercept, 136
 normal equation, 103

- prediction interval, 125
residual, 107
residual sum of squares, 107
- hat matrix, 134
hessian matrix, 89
Hotelling-Lawley's Trace, 297
Hotellings T^2
 one-sample situation, 277
 two-sample situation, 283
Huyn-Feldt condition, 162
- idempotent matrix, 473
identity matrix, 457
independence, 17
influence statistics, 204
information matrix, Fisher, 88
inner product, 477
inner product, fundamental properties, 479
inverse matrix, 443, 459
- Jacobian, 90
- k-nearest neighbor, 344
k-NN, 344
- Lasso, 253
left singular vectors, 461
leverage, 206
likelihood
 conditional, 99
 marginal, 99
 partial, 99
 profile, 98
 quasi, 99
likelihood equations, 85
likelihood ratio test, 129
linear equations, solution, 449
linear functional relation, 273
linear regression analyse, 179
linear restrictions., 109
linear transformation, 437
linearity in parameters, 180
Little Jiffy, 418
logistic curve, 263
Logistic regression, 262
 odds, 266
logit, 264
- mapping - see *transformation*, 433
- matrix
 block matrix, 443
 blocks, 443
 Cholesky decomposition, 458
 cofactor, 442
 coordinate transformation matrix, 439
 definition, 436
 determinant, 441
 diag, 436
 diagonal element, 436
 diagonal matrix, 436
 eigenvalue, 455
 eigenvalue w.r.t. matrix, 470
 eigenvector, 455
 eigenvector w.r.t. matrix, 470
 g2 inverse, 455
 generalised inverse, 449
 hat matrix, 134
 idempotent, 438
 inverse, 438
 Kronecker product, 476
 linear transformation, 437
 minor, 442
 Moore-Penrose inverse, 455
 non-symmetrical, 470
 orthogonal, 457
 partitioned, 443
 positive definite, 464
 positive semi-definite, 464
 principal minor, 465
 product, 437
 pseudoinverse, 446, 449
 rank, 440
 regular, 438
 scalar multiplication, 437
 Schur complement, 445
 semi-definite, 464
 similar matrices, 440
 spectral decomposition, 458
 spectrum, 458
 square, 436
 sum, 436
 symmetric, 436
 tensor product, 476
 tensor product rules, 477
 trace, 473
 transposed, 436
- Maximum likelihood estimate, 88

- Mean
 complex, 60
 complex matrix, 61
 mean squared error, 241
 Mean Squared Error (MSE), 45
 mean value, 3
 Minimum Mean Squared Error, 46
 Model specification
 CContinuous-nesting-class effect, 149
 class variable, 148
 continous variable, 148
 Continuous-by-class effect, 149
 crossed effect, 148
 crossing, 148
 effect, 148
 main effect, 148
 nested effect, 149
 nesting, 148
 polynomial effect, 148
 regressor effect, 148
 Moore-Penrose inverse, 455
 MSE, 241
 multicollinearity, 207, 240
 multidimensional analysis of variance, 302, 304, 306
 multidimensional general linear model, 289, 291, 302, 304, 306
 multiple correlation-coefficient, 41
 multivariate complex random variable, 61
 multivariate general linear model
 multiple correlation-coefficient, 43
 normal distribution, 25
 partial correlation-coefficient, 35
 $N_p(\mu, \Sigma)$, 12
 Newton-Raphson algorithm, 90
 norm, 477
 normal distribution
 multi-dimensional, 11
 multivariate, 11
 two-dimensional, 27
 normal equation, 184
 observed information matrix, 89
 odds, 266
 orthogonal polynomials, 212
 orthogonal regression, 49, 273
 orthogonal vectors, 456, 477
 orthonormal basis, 457
 partial correlation coefficient, 33
 partition theorem, 49, 50
 Pillai's Trace, 297
 precision, 14, 45
 prediction, 226, 240
 prediction interval, 123
 prediction variance, 46
 predictor, 45
 principal axis, 368
 principal component, 369, 374
 principal components, 276, 368
 principal coordinate analysis, 419
 PROC
 GLM, 307
 projection, 478
 Pseudo Covariance Matrix, 64
 pseudoinverse matrix, 446, 474
 pseudoinverse transformation, 447
 pythagorean theorem, 478
 Q-mode, 419, 464
 quadratic form, 464, 474
 \mathbb{R}^n , 430, 432, 435
 R-mode, 464
 random matrix, 3
 random variable
 complex, 60
 rank of matrix, 440, 462
 rank of projection, 440
 Rayleigh quotient, general, 472
 Rayleigh's coefficient, 465
 Rayleigh's quotient, 465
 Regression
 canonical correlation, 422
 ordinary least squares, 421
 Partial least squares, 425
 principal components, 421
 reduced rank, 423
 regression, 45, 46
 confidence interval, 124
 least squares estimate, 104
 normal equation, 103
 orthogonal, 273
 prediction interval, 125
 residual, 107
 residual sum of squares, 107

- weighted, 180
- regression analysis, 179
 - by orthogonal polynomials, 212
 - multidimensional, 290, 299
- regression equation
 - all possible regressions, 228
 - backwards elimination, 229
 - choice of best, 225
 - forward selection, 230
 - stepwise regression, 232
- regular matrix, 438, 440, 441
- regularization, 246
- Relation Matrix, 64
- REML, 95
- Repeated Measurements Models, 158
- reproductivity theorem for
 - normal distribution, 24
- residual, 200
- residual plot, 200
- Restricted Maximum Likelihood, 95
- ridge estimator, 246
- ridge regression, 240
- ridge trace, 249
- right singular vectors, 461
- RMM (Repeated Measurements Models), 158
- row-vector, 436
- Roy's Maximum Root, 297
- RSTUDENT, 205
- SAS
 - PROC CANDISC, 363
 - PROC DISCRIM, 344, 349
 - PROC GLM, 136, 148, 302
 - PROC GLMSELECT, 259, 262, 348
 - PROC LOGISTIC, 266
 - PROC MIXED, 164, 171
 - PROC NLIN, 270
 - PROC REG, 223, 251
- scalar product, 477
- scale-invariance, 416
- Schur complement, 445
- score vector, 85
- scoring equations, 91
- similar matrices, 440
- similarity measures, 419
- singular value, 461
- Singular Value Decomposition, 461
- SS, 304, 306
 - type I - IV, 150
- standard error of estimate, 133
- stepwise regression, 232
- stochastic matrix, 3
- STUDENT RESIDUAL, 205
- studentised residual, 205
- successive testing, 154
- Sums of squares
 - type I - IV, 150
- support, 14
- suppressor variable, 199
- surveying, 122
- SVD, 461
- symmetric matrix, 457
- test for/in
 - assumptions in regr. analysis, 199
 - best discriminator, 335
 - correlation, 39
 - diagonal structure of variance-covariance matrix, 312
 - eigenvalue of correlation matrix, 375
 - eigenvalue of variance-covariance matrix, 375
 - equal variance-covariance matrices, 315
 - equality of means, 334
 - factor model, 418
 - independence, 312
 - McNemar's test statistic, 354
 - mean value, 277, 283
 - multidimensional analysis of variance, 302
 - multidimensional general linear model, 295
 - multiple correlation, 44
 - partial correlation, 39
 - performance of classifiers, 354
 - proportional variance-covariance, 313
- Tikhonov regularization, 246
- TOL, 208
- tolerance, 208
- total variance, 371
- tr, 473
- transformation
 - affine, 435
 - generalised inverse, 448
 - isomorphism, 435
 - linear, 433
 - null-space, 435

- orthogonal, 459
- pseudoinverse, 448
- reflection, 459
- rotation, 459
- transpose, 437
- U, 295
- U-distribution, 296
- uncorrelated, 10
- variance components, 96
- variance inflation, 208
- Variance-Covariance
 - Complex matrix, 62
- variance-covariance matrix, 5
- variation
 - between groups, 303
 - partitioning of total, 303
 - partitioning total, 215, 306
 - within groups, 303
- VC, 10
- vector
 - addition, 430
 - angle, 480
 - associative law, 430
 - basis, 432
 - basis coordinates, 432
 - commutative law, 430
 - conjugate directions, 480
 - dimension, 432
 - distributive law, 430
 - inverse element, 430
 - linear combination, 431
 - linear dependency, 432
 - linear independency, 431
 - neutral element, 430
 - orthogonal w.r.t. matrix, 480
 - orthonormal basis, 457
 - projection, 433
 - projection, idempotent, 433
 - scalar multiplication, 430
 - side-subspace, 431
 - space, 430
 - span, 431
 - subspace, 431
 - sum of sub-spaces, 432
- vector of principal components, 369
- VIF, 208
- $W(n, \Sigma)$, 54
- Wilks' Λ , 297
- Wilks' Λ , 295
- Wilks' complex Lambda distribution, 77
- Wilks' Lambda complex, 77
- Wishart Distribution complex, 72
- Wishart distribution, 54