

DSP

DIGITAL SIGNAL AND IMAGE PROCESSING SERIES



Data Analysis

Edited by Gérard Govaert

STE

 **WILEY**

0212
D232

Data Analysis



Edited by
Gérard Govaert

ISTE



E2010001112

 WILEY

First published in France in 2003 by Hermes Science/Lavoisier entitled: *Analyse des données*
© LAVOISIER, 2003

First published in Great Britain and the United States in 2009 by ISTE Ltd and John Wiley & Sons, Inc.

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms and licenses issued by the CLA. Enquiries concerning reproduction outside these terms should be sent to the publishers at the undermentioned address:

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

© ISTE Ltd, 2009

The rights of Gérard Govaert to be identified as the author of this work have been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

Library of Congress Cataloging-in-Publication Data

Analyse des données. English.
Data analysis / edited by Gérard Govaert.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-84821-098-1
1. Mathematical statistics. I. Govaert, Gérard. II. Title.
QA276.D325413 2009
519.5--dc22

2009016228

British Library Cataloguing-in-Publication Data

A CIP record for this book is available from the British Library
ISBN: 978-1-84821-098-1

Printed and bound in Great Britain by CPI/Antony Rowe, Chippenham and Eastbourne.

Data Analysis

Preface

Statistical analysis has traditionally been separated into two phases: an exploratory phase, drawing on a set of descriptive and graphical techniques, and a decisional phase, based on probabilistic models. Some of the tools employed as part of the exploratory phase belong to *descriptive statistics*, whose elementary exploratory methods consider only a very limited number of variables. Other tools belong to *data analysis*, the subject matter of this book. This topic comprises more elaborate exploratory methods to handle multidimensional data, and is often seen as stepping beyond a purely exploratory context.

The first part of this book is concerned with methods for obtaining the pertinent dimensions from a collection of data. The variables so obtained provide a synthetic description, often leading to a graphical representation of the data. A considerable number of methods have been developed, adapted to different data types and different analytical goals. Chapters 1 and 2 discuss two reference methods, namely Principal Components Analysis (PCA) and Correspondence Analysis (CA), which we illustrate with examples from statistical process control and sensory analysis. Chapter 3 looks at a family of methods known as Projection Pursuit (less well known, but with a promising future), that can be seen as an extension of PCA and CA, which makes it possible to specify the structures that are being sought. Multidimensional positioning methods, discussed in Chapter 4, seek to represent proximity matrix data in low-dimensional Euclidean space. Chapter 5 is devoted to functional data analysis where a function such as a temperature or rainfall graph, rather than a simple numerical vector, is used to characterize individuals.

The second part is concerned with methods of clustering, which seek to organize data into homogenous classes. These methods provide an alternative means, often complementary to those discussed in the first part, of synthesizing and analyzing data. In view of the clear link between clustering and discriminant analysis – in pattern recognition the former is termed unsupervised and the latter supervised learning – Chapter 6 gives a general introduction to discriminant analysis. Chapter 7 then

provides an overall picture of clustering. The statistical interpretation of clustering in terms of mixtures of probability distributions is discussed in Chapter 8 and Chapter 9 looks at how this approach can be applied to spatial data.

I would like to express my heartfelt thanks to all the authors who were involved in this publication. Without their expertise, their professionalism, their invaluable contributions and the wealth of their experience, it would not have been possible.

Gérard GOVAERT

Contents

Preface	xiii
Chapter 1. Principal Component Analysis: Application to Statistical Process Control	1
Gilbert SAPORTA, Ndèye NIANG	
1.1. Introduction	1
1.2. Data table and related subspaces	2
1.2.1. Data and their characteristics	2
1.2.2. The space of statistical units	5
1.2.3. Variables space	7
1.3. Principal component analysis	8
1.3.1. The method	8
1.3.2. Principal factors and principal components	8
1.3.3. Principal factors and principal components properties	10
1.4. Interpretation of PCA results	11
1.4.1. Quality of representations onto principal planes	11
1.4.2. Axis selection	12
1.4.3. Internal interpretation	13
1.4.4. External interpretation: supplementary variables and individuals	15
1.5. Application to statistical process control	18
1.5.1. Introduction	18
1.5.2. Control charts and PCA	20
1.6. Conclusion	22
1.7. Bibliography	23
Chapter 2. Correspondence Analysis: Extensions and Applications to the Statistical Analysis of Sensory Data	25
Jérôme PAGÈS	
2.1. Correspondence analysis	25

2.1.1. Data, example, notations	25
2.1.2. Questions: independence model	26
2.1.3. Intensity, significance and nature of a relationship between two qualitative variables	27
2.1.4. Transformation of the data	28
2.1.5. Two clouds	29
2.1.6. Factorial analysis of \mathbf{X}	30
2.1.7. Aid to interpretation	32
2.1.8. Some properties	33
2.1.9. Relationships to the traditional presentation	35
2.1.10. Example: recognition of three fundamental tastes	36
2.2. Multiple correspondence analysis	39
2.2.1. Data, notations and example	39
2.2.2. Aims	41
2.2.3. MCA and CA	41
2.2.4. Spaces, clouds and metrics	42
2.2.5. Properties of the clouds in CA of the CDT	43
2.2.6. Transition formulae	45
2.2.7. Aid for interpretation	45
2.2.8. Example: relationship between two taste thresholds	46
2.3. An example of application at the crossroads of CA and MCA	50
2.3.1. Data	50
2.3.2. Questions: construction of the analyzed table	51
2.3.3. Properties of the CA of the analyzed table	53
2.3.4. Results	54
2.4. Conclusion: two other extensions	63
2.4.1. Internal correspondence analysis	63
2.4.2. Multiple factor analysis (MFA)	63
2.5. Bibliography	64

Chapter 3. Exploratory Projection Pursuit 67

Henri CAUSSINUS, Anne RUIZ-GAZEN

3.1. Introduction	67
3.2. General principles	68
3.2.1. Background	68
3.2.2. What is an interesting projection?	69
3.2.3. Looking for an interesting projection	70
3.2.4. Inference	70
3.2.5. Outliers	71
3.3. Some indexes of interest: presentation and use	71
3.3.1. Projection indexes based on entropy measures	71
3.3.2. Projection indexes based on L^2 distances	73
3.3.3. Chi-squared type indexes	75

- 3.3.4. Indexes based on the cumulative empirical function 75
- 3.4. Generalized principal component analysis 76
 - 3.4.1. Theoretical background 76
 - 3.4.2. Practice 78
 - 3.4.3. Some precisions 79
- 3.5. Example 81
- 3.6. Further topics 86
 - 3.6.1. Other indexes, other structures 86
 - 3.6.2. Unsupervised classification 86
 - 3.6.3. Discrete data 87
 - 3.6.4. Related topics 88
 - 3.6.5. Computation 89
- 3.7. Bibliography 89

Chapter 4. The Analysis of Proximity Data 93

Gerard D'AUBIGNY

- 4.1. Introduction 93
- 4.2. Representation of proximity data in a metric space 97
 - 4.2.1. Four illustrative examples 97
 - 4.2.2. Definitions 100
- 4.3. Isometric embedding and projection 103
 - 4.3.1. An example of computations 105
 - 4.3.2. The additive constant problem 106
 - 4.3.3. The case of observed dissimilarity measures blurred by noise 108
- 4.4. Multidimensional scaling and approximation 108
 - 4.4.1. The parametric MDS model 109
 - 4.4.2. The Shepard founding heuristics 111
 - 4.4.3. The majorization approach 114
 - 4.4.4. Extending MDS to a semi-parametric setting 119
- 4.5. A fielded application 122
 - 4.5.1. Principal coordinates analysis 122
 - 4.5.2. Dimensionality for the representation space 123
 - 4.5.3. The scree test 125
 - 4.5.4. Recourse to simulations 127
 - 4.5.5. Validation of results 127
 - 4.5.6. The use of exogenous information for interpreting the output configuration 131
 - 4.5.7. Introduction to stochastic modeling in MDS 137
- 4.6. Bibliography 139

Chapter 5. Statistical Modeling of Functional Data 149

Philippe BESSE, Hervé CARDOT

- 5.1. Introduction 149

5.2. Functional framework	152
5.2.1. Functional random variable	152
5.2.2. Smoothness assumption	153
5.2.3. Smoothing splines	154
5.3. Principal components analysis	156
5.3.1. Model and estimation	156
5.3.2. Dimension and smoothing parameter selection	158
5.3.3. Some comments on discretization effects	159
5.3.4. PCA of climatic time series	160
5.4. Linear regression models and extensions	161
5.4.1. Functional linear models	162
5.4.2. Principal components regression	163
5.4.3. Roughness penalty approach	163
5.4.4. Smoothing parameters selection	164
5.4.5. Some notes on asymptotics	165
5.4.6. Generalized linear models and extensions	165
5.4.7. Land use estimation with the temporal evolution of remote sensing data	166
5.5. Forecasting	169
5.5.1. Functional autoregressive process	169
5.5.2. Smooth ARH(1)	171
5.5.3. Locally ARH(1) processes	172
5.5.4. Selecting smoothing parameters	173
5.5.5. Some asymptotic results	173
5.5.6. Prediction of climatic time series	173
5.6. Concluding remarks	176
5.7. Bibliography	177
Chapter 6. Discriminant Analysis	181
Gilles CELEUX	
6.1. Introduction	181
6.2. Main steps in supervised classification	182
6.2.1. The probabilistic framework	182
6.2.2. Sampling schemes	183
6.2.3. Decision function estimation strategies	184
6.2.4. Variables selection	185
6.2.5. Assessing the misclassification error rate	187
6.2.6. Model selection and resampling techniques	189
6.3. Standard methods in supervised classification	190
6.3.1. Linear discriminant analysis	191
6.3.2. Logistic regression	192
6.3.3. The K nearest neighbors method	195
6.3.4. Classification trees	197

6.3.5. Single hidden layer back-propagation network	199
6.4. Recent advances	204
6.4.1. Parametric methods	204
6.4.2. Radial basis functions	207
6.4.3. Boosting	208
6.4.4. Support vector machines	209
6.5. Conclusion	211
6.6. Bibliography	212

Chapter 7. Cluster Analysis 215
 Mohamed NADIF, Gérard GOVAERT

7.1. Introduction	215
7.2. General principles	217
7.2.1. The data	217
7.2.2. Visualizing clusters	218
7.2.3. Types of classification	218
7.2.4. Objectives of clustering	222
7.3. Hierarchical clustering	224
7.3.1. Agglomerative hierarchical clustering (AHC)	225
7.3.2. Agglomerative criteria	226
7.3.3. Example	227
7.3.4. Ward's method or minimum variance approach	227
7.3.5. Optimality properties	228
7.3.6. Using hierarchical clustering	231
7.4. Partitional clustering: the k -means algorithm	233
7.4.1. The algorithm	233
7.4.2. k -means: a family of methods	234
7.4.3. Using the k -means algorithm	236
7.5. Miscellaneous clustering methods	239
7.5.1. Dynamic cluster method	239
7.5.2. Fuzzy clustering	240
7.5.3. Constrained clustering	241
7.5.4. Self-organizing map	242
7.5.5. Clustering variables	243
7.5.6. Clustering high-dimensional datasets	244
7.6. Block clustering	245
7.6.1. Binary data	247
7.6.2. Contingency table	248
7.6.3. Continuous data	249
7.6.4. Some remarks	250
7.7. Conclusion	251
7.8. Bibliography	251

- Chapter 8. Clustering and the Mixture Model** 257
- Gérard GOVAERT
- 8.1. Probabilistic approaches in cluster analysis 257
 - 8.1.1. Introduction 257
 - 8.1.2. Parametric approaches 258
 - 8.1.3. Non-parametric methods 259
 - 8.1.4. Validation 260
 - 8.1.5. Notation 260
- 8.2. The mixture model 261
 - 8.2.1. Introduction 261
 - 8.2.2. The model 261
 - 8.2.3. Estimation of parameters 262
 - 8.2.4. Number of components 263
 - 8.2.5. Identifiability 263
- 8.3. EM algorithm 263
 - 8.3.1. Introduction 263
 - 8.3.2. Complete data and complete-data likelihood 264
 - 8.3.3. Principle 264
 - 8.3.4. Application to mixture models 265
 - 8.3.5. Properties 266
 - 8.3.6. EM: an alternating optimization algorithm 266
- 8.4. Clustering and the mixture model 267
 - 8.4.1. The two approaches 267
 - 8.4.2. Classification likelihood 267
 - 8.4.3. The CEM algorithm 268
 - 8.4.4. Comparison of the two approaches 269
 - 8.4.5. Fuzzy clustering 270
- 8.5. Gaussian mixture model 271
 - 8.5.1. The model 271
 - 8.5.2. CEM algorithm 272
 - 8.5.3. Spherical form, identical proportions and volumes 273
 - 8.5.4. Spherical form, identical proportions but differing volumes 274
 - 8.5.5. Identical covariance matrices and proportions 275
- 8.6. Binary variables 275
 - 8.6.1. Data 275
 - 8.6.2. Binary mixture model 276
 - 8.6.3. Parsimonious model 277
 - 8.6.4. Example of application 279
- 8.7. Qualitative variables 279
 - 8.7.1. Data 279
 - 8.7.2. The model 279
 - 8.7.3. Parsimonious model 281
- 8.8. Implementation 282

8.8.1. Choice of model and of the number of classes	282
8.8.2. Strategies for use	283
8.8.3. Extension to particular situations	283
8.9. Conclusion	284
8.10. Bibliography	284
Chapter 9. Spatial Data Clustering	289
Christophe AMBROISE, Mo DANG	
9.1. Introduction	289
9.1.1. The spatial data clustering problem	289
9.1.2. Examples of applications	290
9.2. Non-probabilistic approaches	293
9.2.1. Using spatial variables	293
9.2.2. Transformation of variables	293
9.2.3. Using a matrix of spatial distances	293
9.2.4. Clustering with contiguity constraints	294
9.3. Markov random fields as models	295
9.3.1. Global methods and Bayesian approaches	295
9.3.2. Markov random fields	297
9.3.3. Markov fields for observations and classes	300
9.3.4. Supervised segmentation	301
9.4. Estimating the parameters for a Markov field	305
9.4.1. Supervised estimation	305
9.4.2. Unsupervised estimation with EM	307
9.4.3. Classification likelihood and inertia with spatial smoothing	310
9.4.4. Other methods of unsupervised estimation	312
9.5. Application to numerical ecology	313
9.5.1. The problem	313
9.5.2. The model: Potts field and Bernoulli distributions	314
9.5.3. Estimating the parameters	315
9.5.4. Resulting clustering	315
9.6. Bibliography	316
List of Authors	319
Index	323

Chapter 1

Principal Component Analysis: Application to Statistical Process Control

1.1. Introduction

Principal component analysis (PCA) is an exploratory statistical method for graphical description of the information present in large datasets. In most applications, PCA consists of studying p variables measured on n individuals. When n and p are large, the aim is to synthesize the huge quantity of information into an easy and understandable form.

Unidimensional or bidimensional studies can be performed on variables using graphical tools (histograms, box plots) or numerical summaries (mean, variance, correlation). However, these simple preliminary studies in a multidimensional context are insufficient since they do not take into account the eventual relationships between variables, which is often the most important point.

Principal component analysis is often considered as the basic method of factor analysis, which aims to find linear combinations of the p variables called components used to visualize the observations in a simple way. Because it transforms a large number of correlated variables into a few uncorrelated principal components, PCA is a dimension reduction method. However, PCA can also be used as a multivariate outlier detection method, especially by studying the last principal components. This property is useful in multidimensional quality control.

1.2. Data table and related subspaces

1.2.1. Data and their characteristics

Data are generally represented in a rectangular table with n rows for the individuals and p columns corresponding to the variables. Choosing individuals and variables to analyze is a crucial phase which has an important influence on PCA results. This choice has to take into account the aim of the study; in particular, the variables have to describe the phenomenon being analyzed.

Usually PCA deals with numerical variables. However, ordinal variables such as ranks can also be processed by PCA. Later in this chapter, we present the concept of supplementary variables which afterwards integrates nominal variables.

1.2.1.1. Data table

Let \mathbf{X} be the (n, p) matrix of observations:

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \vdots & \vdots \\ x_i^1 & x_i^j & x_i^p \\ \vdots & \vdots & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$$

where x_i^j is the value of individual i for variable j (denoted \mathbf{x}^j) which is identified with a vector of n components $(x_1^j, \dots, x_n^j)'$. In a similar way, an individual i is identified to a vector \mathbf{x}_i of p components with $\mathbf{x}_i = (x_i^1, \dots, x_i^p)'$.

Table 1.1 is an example of such a data matrix. Computations have been carried out using SPAD 5 software, version 5¹, kindly provided by J.-P. Gauchi.

The data file contains 57 brands of mineral water described by 11 variables defined in Table 1.2. The data come from the bottle labels. Numerical variables are homogenous; they are all active variables (see section 1.4.3). A variable of a different kind such as price would be considered as a supplementary variable. On the other hand, qualitative variables such as country, type and whether still or sparkling (PG) are necessarily supplementary variables.

1. DECISIA (former CISIA-CERESTA), Building Hoche, 13 rue Auger, 93697 Pantin cedex.

Name	Country	Type	PG	CA	MG	NA	K	SUL	NO3	HCO3	CL
Evian	F	M	P	78	24	5	1	10	3.8	357	4.5
Montagne des Pyrénées	F	S	P	48	11	34	1	16	4	183	50
Cristaline-St-Cyr	F	S	P	71	5.5	11.2	3.2	5	1	250	20
Ficée des Lois	F	S	P	89	31	17	2	47	0	360	28
Volcania	F	S	P	4.1	1.7	2.7	0.9	1.1	0.8	25.8	0.9
Saint Diéry	F	M	G	85	80	385	65	25	1.9	1350	285
Luchon	F	M	P	26.5	1	0.8	0.2	8.2	1.8	78.1	2.3
Volvic	F	M	P	9.9	6.1	9.4	5.7	6.9	6.3	65.3	8.4
Alpes/Moulettes	F	S	P	63	10.2	1.4	0.4	51.3	2	173.2	1
Orée du bois	F	M	P	234	70	43	9	635	1	292	62
Arvie	F	M	G	170	92	650	130	31	0	2195	387
Alpes/Roche des Ecrins	F	S	P	63	10.2	1.4	0.4	51.3	2	173.2	10
Ondine	F	S	P	46.1	4.3	6.3	3.5	9	0	163.5	3.5
Thonon	F	M	P	108	14	3	1	13	12	350	9
Aix les Bains	F	M	P	84	23	2	1	27	0.2	341	3
Contrex	F	M	P	486	84	9.1	3.2	1187	2.7	403	8.6
La Bondaire Saint Hippolite	F	S	P	86	3	17	1	7	19	256	21
Dax	F	M	P	125	30.1	126	19.4	365	0	164.7	156
Quézac	F	M	G	241	95	255	49.7	143	1	1685.4	38
Salvetat	F	M	G	253	11	7	3	25	1	820	4
Stamna	GRC	M	P	48.1	9.2	12.6	0.4	9.6	0	173.3	21.3
Iolh	GR	M	P	54.1	31.5	8.2	0.8	15	6.2	267.5	13.5
Avra	GR	M	P	110.8	9.9	8.4	0.7	39.7	35.6	308.8	8
Rouvas	GRC	M	P	25.7	10.7	8	0.4	9.6	3.1	117.2	12.4
Alisea	IT	M	P	12.3	2.6	2.5	0.6	10.1	2.5	41.6	0.9
San Benedetto	IT	M	P	46	28	6.8	1	5.8	6.6	287	2.4
San Pellegrino	IT	M	G	208	55.9	43.6	2.7	549.2	0.45	219.6	74.3
Levissima	IT	M	P	19.8	1.8	1.7	1.8	14.2	1.5	56.5	0.3
Vera	IT	M	P	36	13	2	0.6	18	3.6	154	2.1
San Antonio	IT	M	P	32.5	6.1	4.9	0.7	1.6	4.3	135.5	1
La Française	F	M	P	354	83	653	22	1055	0	225	982
Saint Benoit	F	S	G	46.1	4.3	6.3	3.5	9	0	163.5	3.5
Plancoët	F	M	P	36	19	36	6	43	0	195	38
Saint Alix	F	S	P	8	10	33	4	20	0.5	84	37
Puits Saint Georges/Casino	F	M	G	46	33	430	18.5	10	8	1373	39
St-Georges/Corse	F	S	P	5.2	2.43	14.05	1.15	6	0	30.5	25
Hildon bleue	B	M	P	97	1.7	7.7	1	4	26.4	236	16
Hildon blanche	B	M	G	97	1.7	7.7	1	4	5.5	236	16
Mont Roucous	F	M	P	1.2	0.2	2.8	0.4	3.3	2.3	4.9	3.2
Ogeu	F	S	P	48	11	31	1	16	4	183	44
Highland Spring	B	M	P	35	8.5	6	0.6	6	1	136	7.5
Parot	F	M	G	99	88.1	968	103	18	1	3380.51	88
Vernière	F	M	G	190	72	154	49	158	0	1170	18
Terres de Flein	F	S	P	116	4.2	8	2.5	24.5	1	333	15
Courmayeur	IT	M	P	517	67	1	2	1371	2	168	1
Pyrénées	F	M	G	48	12	31	1	18	4	183	35
Puits Saint Georges/Monoprix	F	M	G	46	34	434	18.5	10	8	1373	39
Prince Noir	F	M	P	528	78	9	3	1342	0	329	9
Montcalm	F	S	P	3	0.6	1.5	0.4	8.7	0.9	5.2	0.6
Chantereine	F	S	P	119	28	7	2	52	0	430	7
18 Carats	F	S	G	118	30	18	7	85	0.5	403	39
Spring Water	B	S	G	117	19	13	2	16	20	405	28
Vals	F	M	G	45.2	21.3	453	32.8	38.9	1	1403	27.2
Vernand	F	M	G	33.5	17.6	192	28.7	14	1	734	6.4
Sidi Harazem	MO	S	P	70	40	120	8	20	4	335	220
Sidi Ali	MO	M	P	12.02	8.7	25.5	2.8	41.7	0.1	103.7	14.2
Montclar	F	S	P	41	3	2	0	2	3	134	3

Table 1.1. Data table

1.2.1.2. Summaries

1.2.1.2.1. Centroid

Let \bar{x} be the vector of arithmetic means of each of the p variables, defining the centroid:

$$\bar{x} = (\bar{x}^1, \dots, \bar{x}^p)'$$

4 Data Analysis

Name	Complete water name as labeled on the bottle
Country	Identified by the official car registration letters; sometimes it is necessary to add a letter, for example Crete: GRC (Greece Crete)
Type	M for mineral water, S for spring water
PG	P for still water, G for sparkling water
CA	Calcium ions (mg/litre)
MG	Magnesium ions (mg/litre)
NA	Sodium ions (mg/litre)
K	Potassium ions (mg/litre)
SUL	Sulfate ions (mg/litre)
NO3	Nitrate ions (mg/litre)
HCO3	Carbonate ions (mg/litre)
CL	Chloride ions (mg/litre)

Table 1.2. Variable description

where $\bar{x}^j = \sum_{i=1}^n p_i x_i^j$.

If the data are collected following a random sampling, the n individuals all have the same importance in the computations of the sample characteristics. The same weight $p_i = 1/n$ is therefore allocated to each observation.

However, it can be useful for some applications to use weight p_i varying from one individual to another as grouped data or a reweighted sample. These weights, which are positive numbers summing to 1, can be viewed as frequencies and are stored in a diagonal matrix of size n :

$$D_p = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_n \end{pmatrix}.$$

We then have the matrix expression $\bar{x} = X'D_p \mathbf{1}_n$ where $\mathbf{1}_n$ represents the vector of \mathbb{R}^n with all its components equal to 1. The centered data matrix associated with X is then Y with $y_i^j = x_i^j - \bar{x}^j$ and $Y = X - \mathbf{1}_n \bar{x}' = (\mathbf{I}_n - \mathbf{1}_n \mathbf{1}_n' D_p) X$, where \mathbf{I}_n is the unity matrix of dimension n .

1.2.1.2.2. Covariance matrix and correlation matrix

Let $s_j^2 = \sum_{i=1}^n p_i (x_i^j - \bar{x}^j)^2$ and $s_{kl} = \sum_{i=1}^n p_i (x_i^k - \bar{x}^k)(x_i^\ell - \bar{x}^\ell)$, the variance of variable j and the covariance between variables k and ℓ , respectively. They are stored in the covariance matrix $S = X'D_p X - \bar{x}\bar{x}' = Y'D_p Y$.

We define the linear correlation coefficient between variables k and ℓ by:

$$r_{kl} = \frac{s_{kl}}{s_k s_\ell}.$$