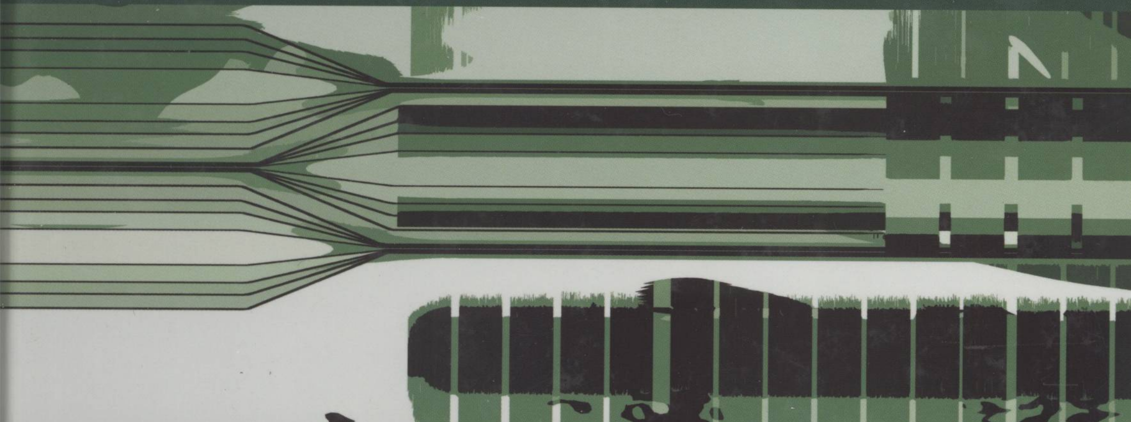
DSP

DIGITAL SIGNAL AND IMAGE PROCESSING SERIES



Data Analysis

Edited by Gérard Govaert

ISTE

WILEY

Data Analysis

Edited by
Gérard Govaert

ISTE

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.

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

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.

FSC
Mixed Sources
Product group from well-managed forests and other controlled sources
Cert no. SGS-COC-2953
www.fsc.org
© 1996 Forest Stewardship Council

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

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.

Chapter written by Gilbert SAPORTA and Ndèye NIANG.

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 $\boldsymbol{x}_i$ of p components with $\boldsymbol{x}_i = (x_i^1, \dots, x_i^p)'$.

Table 1.1 is an example of such a data matrix. Computations have been carried out using SPAD 5 software, version 5 [1], 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
Fiée 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 Bondoire 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 $\overline{\boldsymbol{x}}$ be the vector of arithmetic means of each of the p variables, defining the centroid:

$$\overline{\boldsymbol{x}} = (\overline{x}^1, \ldots, \overline{x}^p)'$$

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

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:

$$\mathbf{D}_p = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_n \end{pmatrix}.$$

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

1.2.1.2.2. Covariance matrix and correlation matrix

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

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

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