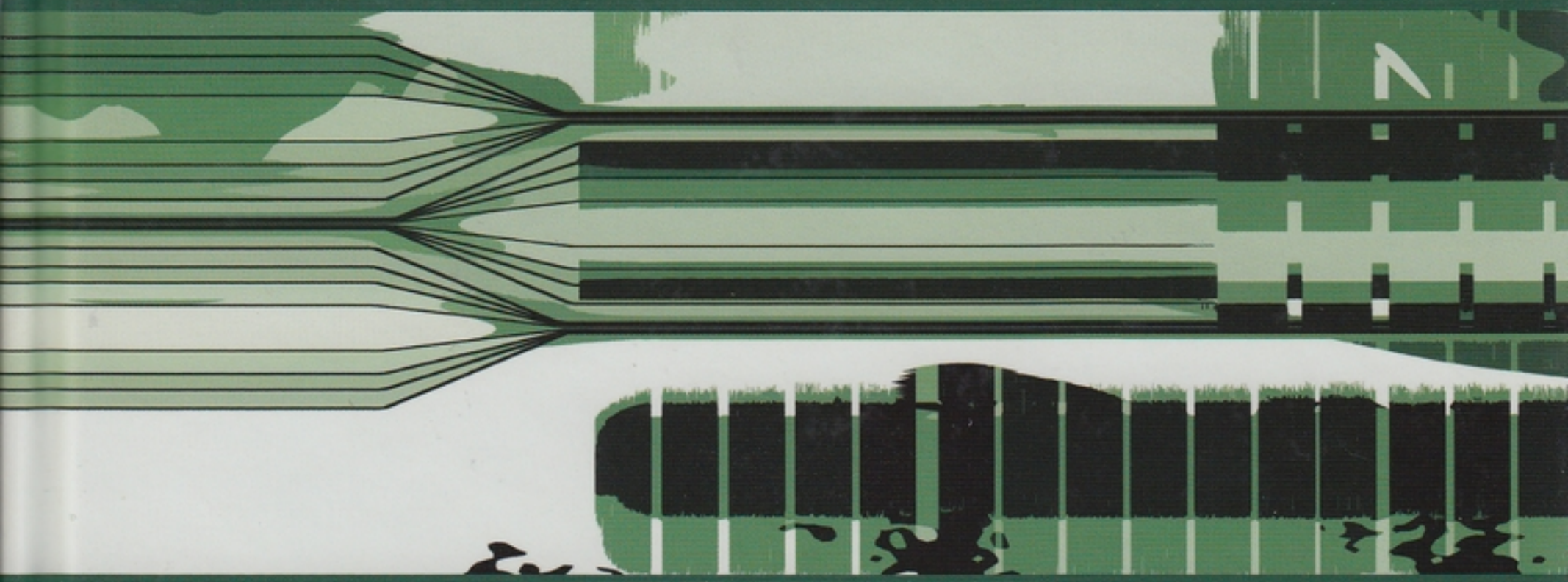
DSP

DIGITAL SIGNAL AND IMAGE PROCESSING SERIES



Data Analysis

Edited by Gérard Govaert

ISTE

WILEY

Contents

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

Traditionally, statistical analysis is divided into two stages: an exploratory stage, based on a comprehensive process of descriptive and graphic techniques; and a decision stage, based on probabilistic models. The exploratory phase, also called data analysis - and which is the object of this book - is the process of transforming data and extracting useful information. This process is normally composed of detailed and often complex exploratory methods that apply to multidimensional data and often exceed the capacity of the exploratory framework itself.

The first part of this book is devoted to methods aimed at identifying relevant dimensions of the data. The variables thus obtained provide a synthetic description which often results in a graphical representation of the data. After a general presentation of discriminating analysis, the second part of the book is devoted to a presentation of clustering methods which constitute another method (which is often complementary to the methods described in the first part of the book) to synthesize and analyze the data. The book concludes by examining the interrelations between data mining and data analysis.

Gérard Govaert is Professor at the University of Technology of Compiègne, France. He is also a member of the CNRS Laboratory Heudiasyc (Heuristics and diagnostics of complex systems). His research interests include latent structure modeling, model selection, model-based cluster analysis, block clustering and statistical pattern recognition. He is one of the authors of the MIXMOD (MIXture MODeling) software.

iSTE
www.iste.co.uk
WILEY
wiley.com
9 781848 210981