Finding Groups in Data

An Introduction to Cluster Analysis

LEONARD KAUFMAN
Vrije Universiteit Brussel,
Brussels, Belgium

PETER J. ROUSSEEUW
Universitaire Instelling Antwerpen,
Antwerp, Belgium

A JOHN WILEY & SONS, INC., PUBLICATION

This Page Intentionally Left Blank

Finding Groups in Data

This Page Intentionally Left Blank

Finding Groups in Data

An Introduction to Cluster Analysis

LEONARD KAUFMAN
Vrije Universiteit Brussel,
Brussels, Belgium

PETER J. ROUSSEEUW
Universitaire Instelling Antwerpen,
Antwerp, Belgium

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication is available.

ISBN 0-471-73578-7

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Our Parents And To Jean Haezendonck

This Page Intentionally Left Blank

Preface

Cluster analysis is a very practical subject. Some 30 years ago, biologists and social scientists began to look for systematic ways to find groups in their data. Because computers were becoming available, the resulting algorithms could actually be implemented. Nowadays clustering methods are applied in many domains, including artificial intelligence and pattern recognition, chemometrics, ecology, economics, the geosciences, marketing, medical research, political science, psychometrics, and many more. This has led to a lot of different methods, and articles on clustering have appeared not only in statistical journals but also in periodicals of all these domains. Clustering is known under a variety of names, such as numerical taxonomy and automatic data classification.

Our purpose was to write an applied book for the general user. We wanted to make cluster analysis available to people who do not necessarily have a strong mathematical or statistical background. Rather than giving an extensive survey of clustering methods, leaving the user with a bewildering multitude of methods to choose from, we preferred to select a few methods that together can deal with most applications. This selection was based on a combination of methodological aims (mainly robustness, consistency, and general applicability) and our own experience in applying clustering to a variety of disciplines.

The book grew out of several courses on cluster analysis that we taught in Brussels, Delft, and Fribourg. It was extensively tested as a textbook with students of mathematics, biology, economics, and political science. It is one of the few books on cluster analysis containing exercises. The first chapter introduces the main approaches to clustering and provides guidance to the choice between the available methods. It also discusses various types of data (including interval-scaled and binary variables, as well as similarity data) and explains how these can be transformed prior to the actual

clustering. The other six chapters each deal with a specific clustering method. These chapters all have the same structure. The first sections give a short description of the clustering method, explain how to use it, and discuss a set of examples. These are followed by two sections (marked with * because they may be skipped without loss of understanding) on the algorithm and its implementation, and on some related methods in the literature. The chapters are relatively independent (except for Chapter 3 which builds on Chapter 2), allowing instructors to cover only one chapter in a statistics course. Another advantage is that researchers can pick out the method they need for their current application, without having to read other chapters. (To achieve this structure, some things had to be repeated in the text.) Occasionally, we handed a single chapter to someone working on a particular problem.

Chapters 2, 3, and 4 are about partitioning methods, whereas Chapters 5, 6, and 7 cover hierarchical techniques. Whenever possible, we constructed methods that cannot only analyze data consisting of measurements (i.e., objects with variables) but also data consisting of dissimilarities between objects. (This excluded parametric approaches, such as those based on multivariate mixture distributions.) We also wanted the methods to be consistent for large data sets. All the selected methods are of the L_1 type, which means that they minimize sums of dissimilarities (rather than sums of *squared* dissimilarities, as in the classical nonrobust methods). Some of the methods are new, such as the approach for partitioning large data sets and the L_1 method for fuzzy clustering. Also, the clusterings are accompanied by graphical displays (called silhouettes and banners) and corresponding quality coefficients, which help the user to select the number of clusters and to see whether the method has found groups that were actually present in the data.

Current statistical software packages contain only a few clustering techniques and they are not the more modern methods. This forced us to write new programs, the use of which is described in the book. Their present version is for IBM-PC and compatible machines, but the source codes are very portable and have run on several types of mainframes without problems. The programs (together with their sources and the data sets used in the book) are available on floppy disks by writing to the authors. The programs are also being integrated in the workstation package S-PLUS of Statistical Sciences, Inc., P.O. Box 85625, Seattle, WA 98145-1625.

We are grateful to Frank Critchley, Jan de Leeuw, Gaetan Libert, Glenn Milligan, Frank Plastria, Marc Roubens, John Tukey, Bert van Zomeren, Howard Wainer, Michael Windham, and David Wishart for helpful suggestions and stimulating discussions on topics covered in this book and to

Etienne Trauwaert for contributing to Section 4 of Chapter 4. Valuable comments on the manuscript were given by Phil Hopke, Doug Martin, Marie-Paule Derde, and two reviewers. Annie De Schrijver was responsible for drawing several figures. Finally, we would like to thank our wives, Jacqueline and Lieve, for their patience and support.

LEONARD KAUFMAN
PETER J. ROUSSEEUW

Edegem, Belgium
August, 1989

This Page Intentionally Left Blank

Contents