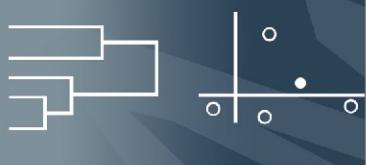


**STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION**

H.A.L.Kiers
J.-P.Rasson
P.J.F.Groenen
M.Schader
Editors

Data Analysis, Classification, and Related Methods



Springer

Studies in Classification, Data Analysis, and Knowledge Organization

Managing Editors

H.-H. Bock, Aachen
W. Gaul, Karlsruhe
M. Schader, Mannheim

Editorial Board

F. Bodendorf, Nürnberg
P. G. Bryant, Denver
F. Critchley, Birmingham
E. Diday, Paris
P. Ihm, Marburg
J. Meulmann, Leiden
S. Nishisato, Toronto
N. Ohsumi, Tokyo
O. Opitz, Augsburg
F. J. Radermacher, Ulm
R. Wille, Darmstadt

Springer

Berlin

Heidelberg

New York

Barcelona

Hong Kong

London

Milan

Paris

Singapore

Tokyo

Titles in the Series

H.-H. Bock and P. Ihm (Eds.)

Classification, Data Analysis, and Knowledge Organization. 1991
(out of print)

M. Schader (Ed.)

Analyzing and Modeling Data and Knowledge. 1992

O. Opitz, B. Lausen, and R. Klar (Eds.)

Information and Classification. 1993
(out of print)

H.-H. Bock, W. Lenski, and M. M. Richter (Eds.)

Information Systems and Data Analysis. 1994
(out of print)

E. Diday, Y. Lechevallier, M. Schader, P. Bertrand,
and B. Burtschy (Eds.)

New Approaches in Classification and Data Analysis. 1994
(out of print)

W. Gaul and D. Pfeifer (Eds.)

From Data to Knowledge. 1995

H.-H. Bock and W. Polasek (Eds.)

Data Analysis and Information Systems. 1996

E. Diday, Y. Lechevallier and O. Opitz (Eds.)

Ordinal and Symbolic Data Analysis. 1996

R. Klar and O. Opitz (Eds.)

Classification and Knowledge Organization. 1997

C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock,
and Y. Baba (Eds.)

Data Science, Classification, and Related Methods. 1998

I. Balderjahn, R. Mathar, and M. Schader (Eds.)

Classification, Data Analysis, and Data Highways. 1998

A. Rizzi, M. Vichi, and H.-H. Bock (Eds.)

Advances in Data Science and Classification. 1998

M. Vichi and O. Opitz (Eds.)

Classification and Data Analysis. 1999

W. Gaul and H. Locarek-Junge (Eds.)

Classification in the Information Age. 1999

H.-H. Bock and E. Diday

Analysis of Symbolic Data. 2000

Henk A. L. Kiers · Jean-Paul Rasson
Patrick J.F. Groenen · Martin Schader (Eds.)

Data Analysis, Classification, and Related Methods

With 96 Figures



Springer

Professor Dr. Henk A. L. Kiers
University of Groningen
Heymans Institute (PA)
Grote Kruisstraat 2/1
NL-9712 TS Groningen

Dr. Patrick J. F. Groenen
Leiden University
Data Theory Group
Department of Education
P.O. Box 9555
NL-2300 RB Leiden

Professor Dr. Jean-Paul Rasson
University of Namur
Directeur du Department
de Mathématique
Facultés Universitaires
Notre-Dame de la Paix
Rempart de la Vierge, 8
B-5000 Namur

Professor Dr. Martin Schader
University of Mannheim
Lehrstuhl
für Wirtschaftsinformatik III
Schloß
D-68131 Mannheim

*Proceedings of the 7th Conference of the
International Federation of Classification Societies (IFCS-2000)
University of Namur, Belgium
11-14 July, 2000*

Cataloguing-in-Publication Data applied for
Data analysis, classification and related methods / Henk A. L. Kiers ... (ed.). - Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Singapore; Tokyo: Springer, 2000
(Studies in classification, data analysis, and knowledge organization)

ISBN-13: 978-3-540-67521-1 e-ISBN-13: 978-3-642-59789-3

DOI: 10.1007/978-3-642-59789-3

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer-Verlag is a company in the BertelsmannSpringer publishing group.
© Springer-Verlag Berlin · Heidelberg 2000

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Softcover-Design: Erich Kirchner, Heidelberg

SPIN 10725385 43210 - Printed on acid-free paper

Preface

This volume contains a selection of papers presented at the Seventh Conference of the International Federation of Classification Societies (IFCS-2000), which was held in Namur, Belgium, July 11–14, 2000. From the originally submitted papers, a careful review process involving two reviewers per paper, led to the selection of 65 papers that were considered suitable for publication in this book.

The present book contains original research contributions, innovative applications and overview papers in various fields within data analysis, classification, and related methods. Given the fast publication process, the research results are still up-to-date and coincide with their actual presentation at the IFCS-2000 conference. The topics captured are:

- Cluster analysis
- Comparison of clusterings
- Fuzzy clustering
- Discriminant analysis
- Mixture models
- Analysis of relationships data
- Symbolic data analysis
- Regression trees
- Data mining and neural networks
- Pattern recognition
- Multivariate data analysis
- Robust data analysis
- Data science and sampling

The IFCS (International Federation of Classification Societies)

The IFCS promotes the dissemination of technical and scientific information concerning data analysis, classification, related methods, and their applications. The IFCS is a federation of the following member societies:

- British Classification Society (BCS)
- Associação Portuguesa de Classificação e Análise de Dados (CLAD)
- Classification Society of North America (CSNA)
- Gesellschaft für Klassifikation (GfKl)
- Japanese Classification Society (JCS)
- Korean Classification Society (KCS)
- Société Francophone de Classification (SFC)
- Società Italiana di Statistica (SIS)
- Sekcja Klasyfikacji i Analizy Danych PTS (SKAD)

- Vereniging voor Ordinatie en Classificatie (VOC)
- Irish Pattern Recognition and Classification Society (IPRCS)

Previous IFCS-conferences were held in Aachen (Germany, 1987), Charlottesville (USA, 1989), Edinburgh (UK, 1991), Paris (France, 1993), Kobe (Japan, 1996), and Rome (Italy, 1998).

Acknowledgements

First of all, we wish to express our gratitude towards the authors of the papers in the present volume, not only for their contributions, but also for their diligence and timely production of the final versions of their papers. Secondly, we thank the reviewers (listed at the end of this book) for their careful reviews of the originally submitted papers, and in this way, for their support in selecting the best papers in this publication.

We also thank M. Bihn, F. Holzwarth, and R. Milewski of Springer-Verlag, Heidelberg, for their support and dedication to the production of this volume.

Finally, the technical and administrative support we received from J.M. Baan, E. de Boer, K. Friesen, D. Jacquemin, B. Kip, H.J. Kreusch, and A. Verstappen-Remmers is gratefully acknowledged.

Groningen, Namur, Leiden, Mannheim
July 2000

*Henk A.L. Kiers
Jean-Paul Rasson
Patrick J.F. Groenen
Martin Schader*

Contents

Part I. Cluster Analysis

Cluster Analysis and Mixture Models

Classifier Probabilities	3
---------------------------------------	----------

J. A. Hartigan

Cluster Analysis Based on Data Depth	17
---	-----------

Richard Hoberg

An Autonomous Clustering Technique	23
---	-----------

Yoshiharu Sato

Unsupervised Non-hierarchical Entropy-based Clustering	29
---	-----------

M. Jardino

Improving the Additive Tree Representation of a Dissimilarity Matrix Using Reticulations	35
---	-----------

Vladimir Makarenkov, Pierre Legendre

Double Versus Optimal Grade Clusterings	41
--	-----------

Alicja Ciok

The Effects of Initial Values and the Covariance Structure on the Recovery of some Clustering Methods	47
--	-----------

Istvan Hajnal, Geert Loosveldt

What Clusters Are Generated by Normal Mixtures?	53
--	-----------

Christian Hennig

A Bootstrap Procedure for Mixture Models	59
---	-----------

Suzanne Winsberg, Geert deSoete

Fuzzy Clustering

A New Criterion of Classes Validity	63
--	-----------

Arnaud Devillez, Patrice Billaudel, Gérard Villermain Lecolier

Application of Fuzzy Mathematical Morphology for Unsupervised Color Pixels Classification	69
--	-----------

A. Gillet, C. Botte-Lecocq, L. Macaire and J.-G. Postaire

A Hyperbolic Fuzzy k-Means Clustering and Algorithm for Neural Networks	77
<i>Norio Watanabe, Tadashi Imaizumi, Toshiko Kikuchi</i>	
 <i>Special Purpose Classification Procedures and Applications</i>	
A Toolkit for Development of the Domain-Oriented Dictionaries for Structuring Document Flows	83
<i>Pavel P. Makagonov, Mikhail A. Alexandrov, Konstantin Sboychakov</i>	
Classification of Single Malt Whiskies	89
<i>David Wishart</i>	
Robust Approach in Hierarchical Clustering: Application to the Sectorisation of an Oil Field	95
<i>Jean-Paul Valois</i>	
A Minimax Solution for Sequential Classification Problems ...	101
<i>Hans J. Vos</i>	
 <i>Verification and Comparison of Clusterings</i>	
Comparison of Ultrametrics Obtained With Real Data, Using the P_L and VAL_{Aw} Coefficients	107
<i>Isabel Pinto Doria, Georges Le Calvé, Helena Bacelar-Nicolau</i>	
Numerical Comparisons of two Spectral Decompositions for Vertex Clustering	113
<i>P. Kuntz, F. Henaux</i>	
Measures to Evaluate Rankings of Classification Algorithms	119
<i>Carlos Soares, Pavel Brazdil, Joaquim Costa</i>	
A General Approach to Test the Pertinence of a Consensus Classification	125
<i>Guy Cucumel, François-Joseph Lapointe</i>	
 <i>Dissimilarity Measures</i>	
On a Class of Aggregation-invariant Dissimilarities Obeying the Weak Huygens' Principle	131
<i>F. Bavaud</i>	
A Short Optimal Way for Constructing Quasi-ultrametrics From Some Particular Dissimilarities	137
<i>B. Fichet</i>	

Missing Data in Cluster Analysis

Estimating Missing Values in a Tree Distance.....	143
<i>A. Guénoche, S. Grandcolas</i>	
Estimating Trees From Incomplete Distance Matrices:	
A Comparison of Two Methods.....	149
<i>Claudine Levasseur, Pierre-Alexandre Landry, François-Joseph Lapointe</i>	
Zero Replacement in Compositional Data Sets.....	155
<i>J. A. Martín-Fernández, C. Barceló-Vidal, V. Pawlowsky-Glahn</i>	
EM Algorithm for Partially Known Labels.....	161
<i>C. Ambroise, G. Govaert</i>	

Part II. Discrimination, Regression Trees, and Data Mining

Discriminant Analysis

Detection of Company Failure and Global Risk Forecasting ...	169
<i>Mireille Bardos</i>	
Discriminant Analysis by Hierarchical Coupling	
in EDDA Context	175
<i>Isabel Brito, Gilles Celeux</i>	
Discrete Discriminant Analysis: The Performance	
of Combining Models by a Hierarchical Coupling Approach... 	181
<i>Ana Sousa Ferreira, Gilles Celeux, Helena Bacelar-Nicolau</i>	

Discrimination Based on the Atypicality Index	
versus Density Function Ratio	187
<i>H. Chamlal and S. Slaoui Chah</i>	

Decision and Regression Trees

A Third Stage in Regression Tree Growing:	
Searching for Statistical Reliability	193
<i>Carmela Cappelli, Francesco Mola, Roberta Siciliano</i>	
A New Sampling Strategy for Building Decision Trees	
from Large Databases	199
<i>J.H. Chauchat, R. Rakotomalala</i>	

Generalized Additive Multi-Model for Classification and Prediction.....	205
<i>Claudio Conversano, Francesco Mola, Roberta Siciliano</i>	
Radial Basis Function Networks and Decision Trees in the Determination of a Classifier	211
<i>Rossella Miglio, Marilena Pillati</i>	
Clustered Multiple Regression	217
<i>Luis Torgo, J. Pinto da Costa</i>	
<i>Neural Networks and Data Mining</i>	
Artificial Neural Networks, Censored Survival Data, Statistical Models	223
<i>Antonio Ciampi, Yves Lechevallier</i>	
Visualisation and Classification with Artificial Life	229
<i>Alfred Ultsch</i>	
<i>Pattern Recognition and Geometrical Statistics</i>	
Exploring the Periphery of Data Scatters: Are There Outliers?	235
<i>Giovanni C. Porzio, Giancarlo Ragozini</i>	
Discriminant Analysis Tools for Non Convex Pattern Recognition	241
<i>Marcel Rémon</i>	
A Markovian Approach to Unsupervised Multidimensional Pattern Classification	247
<i>A. Sbihi, A. Moussa, B. Benmiloud, J.-G. Postaire</i>	
<hr/>	
Part III. Multivariate and Multidimensional Data Analysis	
<hr/>	
<i>Multivariate Data Analysis</i>	
An Algorithm with Projection Pursuit for Sliced Inverse Regression Model	255
<i>Masahiro Mizuta, Hiroyuki Minami</i>	
Testing Constraints and Misspecification in VAR-ARCH Models	261
<i>Wolfgang Polasek, Shuangzhe Liu</i>	

Goodness of Fit Measure based on Sample Isotone Regression of Mokken Double Monotonicity Model.....	267
<i>Teresa Rivas Moya</i>	

Multiway Data Analysis

Fuzzy Time Arrays and Dissimilarity Measures for Fuzzy Time Trajectories	273
<i>Renato Coppi, Pierpaolo D'Urso</i>	

Three-Way Partial Correlation Measures	279
<i>Donatella Vicari</i>	

Analysis of Network and Relationship Data and Multidimensional Scaling

Statistical Models for Social Networks	285
<i>Stanley Wasserman, Philippa Pattison</i>	

Application of Simulated Annealing in some Multidimensional Scaling Problems	297
<i>Javier Trejos, William Castillo, Jorge González, Mario Villalobos</i>	

Data Analysis Based on Minimal Closed Subsets	303
<i>S. Bonnevay, C. Largeron-Leteno</i>	

Robust Multivariate Methods

A Robust Method for Multivariate Regression	309
<i>Stefan Van Aelst, Katrien Van Driessen, Peter J. Rousseeuw</i>	

Robust Methods for Complex Data Structures.....	315
<i>Ursula Gather, Claudia Becker, Sonja Kuhnt</i>	

Robust Methods for Canonical Correlation Analysis	321
<i>Catherine Dehon, Peter Filzmoser, Christophe Croux</i>	

Part IV. Data Science

Data Science and Data Collection

From Data Analysis to Data Science	329
<i>Noboru Ohsumi</i>	

Evaluation of Data Quality and Data Analysis	335
<i>Chikio Hayashi</i>	

Collapsibility and Collapsing Multidimensional Contingency Tables—Perspectives and Implications	341
<i>Stefano De Cantis, Antonino M. Oliveri</i>	
 <i>Sampling and Internet Surveys</i>	
Data Collected on the Web	347
<i>Vasja Vehovar, Katja Lozar Manfreda, Zenel Batagelj</i>	
Some Experimental Surveys on the WWW Environments in Japan	353
<i>Osamu Yoshimura, Noboru Ohsumi</i>	
Bootstrap Goodness-of-fit Tests for Complex Survey Samples	359
<i>Andrea Scagni</i>	
<hr/>	
Part V. Symbolic Data Analysis	
 <i>Classification and Analysis of Symbolic Data</i>	
Regression Analysis for Interval-Valued Data	369
<i>L. Billard, E. Diday</i>	
Symbolic Approach to Classify Large Data Sets.....	375
<i>Francisco de A.T. de Carvalho, Cezar A. de F. Anselmo, Renata M.C.R. de Souza</i>	
Factorial Methods with Cohesion Constraints on Symbolic Objects	381
<i>N.C. Lauro, R. Verde, F. Palumbo</i>	
A Dynamical Clustering Algorithm for Multi-nominal Data ..	387
<i>Rosanna Verde, Francisco de A. T. de Carvalho, Yves Lechevallier</i>	
 <i>Software</i>	
DB2SO : A Software for Building Symbolic Objects from Databases	395
<i>Georges Hébrail, Yves Lechevallier</i>	
Symbolic Data Analysis and the SODAS Software in Official Statistics	401
<i>Raymond Bisdorff, Edwin Diday</i>	
Strata Decision Tree SDA Software	409
<i>M. Carmen Bravo</i>	

Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software.....	417
<i>Mireille Gettler Summa</i>	
List of Reviewers	423
Index	425

Part I

Cluster Analysis

Classifier Probabilities

J. A. Hartigan

Department of Statistics, Yale University,
P.O.Box 208290 New Haven, CT 06520-8290, USA

Abstract. In statistical clustering, we usually devise probability models that begin by specifying joint distributions of data and possible classifications and end in reporting classifications that are probable given the data. Yet the art and practice of classification is more fundamental and prior to probabilistic analysis, and so it is worthwhile to ask how one might derive probabilities from classifications, rather than derive classifications from probabilities. In this scheme, a *classifier* is assumed able to express any knowledge as a classification consisting of a number of statements of the form $\{x \in y\}$, in words, x is a member of y . We specify an *inductive* probability distribution over all such classifications. Probabilities for future outcomes are determined by the probabilities of the classifications formed by the classifier corresponding to those outcomes. Particular examples studied are coin tossing, recognition, the globular cluster Messier 5, and the next president of the United States.

1 Introduction

I do not deny but nature, in the constant production of particular beings, makes them not always new and various, but very much alike and of kin to one another: but I think it nevertheless true, that the boundaries of the species, whereby men sort them, are made by men; since the essences of the species, distinguished by different names, are, as has been proved, of man's making, and seldom adequate to the internal nature of the things they are taken from. So that we may truly say, such a manner of sorting of things is the workmanship of man.

John Locke(1689, VI,37).

There are two roles of classification in statistics; first, in any statistical inquiry, there is a *prior* classification that defines the objects of study and the ways in which they will be counted and measured; secondly, given the data, there may be *posterior* classifications that categorize the objects of study into homogeneous classes within each of which standard analyses are appropriate. These posterior classifications may be constructed in a probabilistic framework that specifies a prior probability of classifications, a conditional probability of the data given the classification, and so, by Bayes Theorem, computes the probability of the various possible classifications given the data.

On the other hand, because classification is prior to, and more fundamental than probability, classification methods may be used, without any

explicit probability calculations, in prediction. Known objects are classified into classes with similar properties; new objects, with some unknown properties, are classified into the established classes according to our knowledge about them; and the common properties of those classes are predicted for the new objects. We will be uncertain to various degrees about the classification and the predicted properties, so it remains desirable to 'probabilify' these predictions.

In reality, all arguments from experience are founded on the similarity which we discover among natural objects, and by which we are induced to expect effects similar to those which we have found to follow from such objects.... For all inferences from experience suppose, as their foundation, that the future will resemble the past, and that similar powers will be conjoined with similar sensible qualities. If there be any suspicion that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless, and can give rise to no inference or conclusion. It is impossible, therefore, that any arguments from experience can prove this resemblance of the past to the future; since all these arguments are founded on the supposition of that resemblance.

Hume(1758,IV,2)

I take the appropriate expression of Hume's similarity to be a classification of objects. Objects are similar in a certain way if they are classified into the same class. I propose a *classifier* probability in which classification is the fundamental tool of probability calculations. In this scheme, a *classifier* C is assumed able to express the knowledge to be gained in an experiment a classification c consisting of a number of *assignments* of the form $\{x \in y\}$, in words, x is a member of y . Such a family might be a partition, a hierarchy, or contain overlapping clusters. The assignments will be assumed to satisfy the *anti-transitive* constraint that $\{x \in y\}, \{y \in z\}, \{x \in z\}$ is impossible. (One possible interpretation of membership is that $\{x \in y\}$ means that x has all the properties of y . Thus we might devise some abstract higher level objects y_1, \dots, y_k from which the lower level objects x_1, \dots, x_m inherit properties.) Mathematically, the classification is a directed graph on the objects x, y, z, \dots in which $\{x \in y\}$ means the link $x \rightarrow y$.

We specify an *inductive* probability distribution over all such classifications. Let n_y be the number of x 's for which $\{x \in y\}$. Then the *inductive* probability of c is

$$p(c) = K \prod_{n_y > 0} (n_y - 1)!.$$

where K is to be computed so that the probabilities over all possible classifications for a given set of objects sums to 1. The motivation for this probability is an inductive assignment of new objects. Suppose some objects x_1, \dots, x_N are assigned to y and a new object is to be assigned to just one of the x 's, or to y ; the anti-transitive constraint makes it impossible to assign the new object to both an x and a y , although it could in general be assigned to more

than one x . The probability of assignment to x_i relative to the probability of assignment to y is n_{x_i}/n_y ; the probability of assignment to some one of the x 's compared to y is $\sum n_{x_i}/n_y$. Thus the new assignment occurs with probability equal to the empirical proportion of past assignments.

These classification probabilities are the only probabilities in this system. This model differs from the standard statistical model in classification problems which needs also a conditional probability for the data given all possible classifications. How then do we give probabilities to future outcomes appropriately determined by our knowledge of the past? Probabilities for future outcomes are determined by the probabilities of the classifications formed by the classifier corresponding to those outcomes. The classifier expresses knowledge, including possible future knowledge, as a classification, and this mapping is used to transfer probability specified for classifications to the probability of future outcomes.

I describe the classification probabilities as inductive, but they are not *frequentist*, which requires some sequence of similar experiments in which probability of an event equals the long run limiting frequency of occurrence of the event. If any such sequence were available, it should be explicitly expressed in the classification expressing available knowledge. These probabilities are intended to be *epistemic*, expressing our uncertainty due to lack of knowledge about the future. They are not proposed as *descriptive, personal* probabilities reflecting my own willingness to bet about certain events(after all, I might have all kinds of reasons to bet that you have no interest in); oh no, they are *prescriptive, public* probabilities recommended for guiding action by anyone who classifies knowledge.

Are these probabilities objective or subjective? Both. Once the classifier is determined, they are objective, since they are specified by the inductive classifier probabilities. However the human classifier will classify evidence subjectively. The subjective component is captured in the subjective classifications formed.

De Finetti(1973), one of the founders of personal probability, depicted the objectivistic position by this playful analogy " This ground is not sufficiently consistent: it is sand. Let us remove the sand, and ground the building on the void." If we use an abstract formulation of a probability space in which an unknown point takes a value in some set of possible values, and new knowledge is represented as discovery that the point takes values in a subset of the original possible set, there is indeed only a featureless void available for building probability. I propose that the sand of personal opinion be replaced by the concrete and mortar of prior classifications and probabilities: these judgements may be viewed as subjective opinions, but they are explicitly stated subjective opinions, and all who agree on the classification can agree on the derived probabilities.

2 Coin tossing

It is necessary to demonstrate that classifier probabilities can accommodate the standard games of chance that originated probability theory and frequentist conceptions of probability, Bernoulli(1713).

An Australian penny used to be marked on one side with the Head of the English Sovereign, and on the other with a Kangaroo, Heads or Tails. In the game of Heads or Tails, the coin is repeatedly tossed and money changes hands on the event of Heads or Tails turning up. It is accepted in practice that a fair bet is to receive one penny if a Head turns up and in return to pay one penny if a Tail turns up; that is, the event of Head turning up has probability $\frac{1}{2}$.

What classifications are relevant to this asserted probability?

- Tosses of the coin are performed under rules that make the tosses indistinguishable from each other. For example, each toss is performed in earth gravity from a height of one foot above a smooth granite surface, and the coin must rotate 5 times before striking the surface.
- Heads on different tosses of the same coin are similar, as are Tails.
- The coin is similar in varying degrees to other pennies from the same mint, and to pennies from different mints.
- The Head on one coin is similar to the Head on another, as are the Tails.
- The coin is similar to a homogeneous circular cylinder.

Thus the probability of $\frac{1}{2}$ for the present toss is based on knowledge expressed in five kinds of similarity judgements, the similarity between different tosses of the present coin, the similarity between the same faces of the same coin appearing on different tosses, the similarity between different pennies, the similarity between the same faces of different coins. the physical similarity of the coin to a circular cylinder. More explicitly, these similarity judgements are in the form of classes of objects, each object consisting of a coin toss with its result: the set of tosses of a coin form one class; the Head resulting tosses of a single coin and the Tail resulting tosses form two classes; the similar coins and circular cylinders are partitioned into a family of classes one member of which will contain the present coin; Heads and Tails within a coin class form two classes.

It may well be that our classifier has available quantitative data such as the dimensions and center of gravity of this and other coins, as well as histories of various coin tosses conducted under different conditions. We take all such knowledge to be expressed in the classifier's choice c of classification of the coins, and of the subclasses of Heads and Tails within each class of coins. The classifier's probability of Heads at the next toss of our coin is obtained by considering the classification c_h that would be chosen if Heads were observed, and the classification c_t that would be chosen if Tails were observed. The relative probability of Heads to Tails is $p(c_h)/p(c_t)$.

Suppose for example that we are about to toss a new penny. We classify it with other pennies, whose history consists of H Heads and T Tails, where say $H = T = 10000$. The only change in classification after one toss of the coin will be to assign the toss to the Head class, or the Tail class; these changes will increase the size of the head class to $H + 1$, or increase the size of the Tail class to $T + 1$. So the relative probability of Heads or Tails for this next toss is $H!(T - 1)!/(H - 1)!T!$ which is 1.

Suppose now that h Heads and t Tails have been accumulated in a series of tosses for the new coin, but that the data are not sufficiently weighty after a further toss for us to alter the classification of coins except for the Head class and the Tail class. Then the probability of Heads vs Tails on the next toss is $\frac{H+h}{T+t}$; this would also be the probability in a Bayes formulation in which the tosses were independent Bernoulli variables with expectation p , and the prior density for p is the maximum likelihood prior $\frac{1}{p(1-p)}$.

The more interesting case has the classifier changing the classification of the coin with the advent of new data; for example, suppose that just one more Head in a surprising excess of Heads would convince the classifier that something is wrong with this coin, and that a new classification with this coin as a singleton is necessary. The probability of that one more Head will depend on the probability of the classification {this coin, other coins} versus {all coins}, which will be small, and on the probability of Heads given the split versus the probability of Heads given no split, which will be large. Suppose for example we assume the probability of a classification into k classes with n_1, \dots, n_k coins is proportional to $\prod n_i!$ which is analogous to Bayes uniform distribution in the binomial case. Then the probability of a split versus no split for n coins is $1/n$, and the probability of a Head producing the split versus a Tail producing no split is approximately

$$\frac{1}{n} 2^{h+t+1} (h+1)!t!/(h+t+2)!$$

which gets large when h is significantly larger than t .

In later calculations, assuming that the split is retained, the relative probability of Heads and Tails becomes $(h+1)/(t+1)$, so that now the contributions of the other coins is ignored. I suggest that this evolution of probabilities follows actual practice; we assume the coin is fair until evidence suggests otherwise, then compute the probabilities based only on the exceptional coin. This evolution is NOT a Bayesian evolution according to the joint probabilities of classifications and coin histories originally specified, which induce a conditional probability of Heads and Tails at the next toss different from the present one; the classifier intervenes with *new* knowledge, the changed classification.

3 Recognition

Historically, theories of probability began with gambling games of the coin tossing type, in which a gambling device is repeatedly operated under similar conditions to produce a sequence of results on which the participants wager. The devices are made symmetrical, and the repeated operations are under ostentatiously similar conditions so that the gamblers may confidently form expectations about the results as a base for the wagering.

And there are beautiful theorems, the laws of large numbers, the central limit theorem, the behaviour of random walks, that give asymptotic certainties to theoretically posited independent identically distributed gambling sequences. In real life, we must maintain a respectful skeptical Humean uncertainty about all matters of substance, and in particular cannot predict the course of actual gambling sequences; who knows how the equipment or its operators might change in the future.

One kind of problem that differs from the gambling scenario is that of *recognition*. When we recognize a present situation as being similar to one previously experienced, we bring our knowledge of the previous experiences to make predictions about unknown aspects of the present. Let us explore the role of classification and probability in recognition.

The recognition process is as follows.

- *The past* We remember previous experiences in the form of objects having various properties.
- *The present* We have before us a number of objects available for present study.
- *The future* We wish to infer whether or not the present objects have some property Y .

I see some objects at a distance. They move in such a way, and are of the right size, for me to recognise them as people. One or two of them are smaller, perhaps children. As they come closer, I notice by dress and facial hair that some are female and some are male; hard to tell about the child. Closer still, the objects attract my attention by waving, someone I know perhaps? Moving closer, I see it is my daughter Ros, her husband Graham and my granddaughter Bia. We meet and talk with high expectations that we will understand each other.

It is this recognition process that enables us to make useful predictions, not some hypothesized tedious accumulation of results of identically prepared experiments. In the gambling scenario, accuracy of prediction about an unknown parameter is order $\frac{1}{\sqrt{n}}$ for n repetitions; it takes many repetitions to obtain precise estimates.

In the recognition scenario described above, no such slow accumulation occurs. At the distance of 500 yards, I did not see the people at all; at a distance of 300 yards, they could have been anybody, and were of no particular

interest. By the time they reached 50 yards, I was nearly certain (probability $> .99999$, since we Humeans must always hold back for the future not resembling the past) who they were, and therefore where they lived, what languages they spoke, their ages, their behaviours, their preferences, and their interests. What is it that makes the probability of recognition jump so quickly from almost zero to almost one?

The classifier , using any present knowledge and past knowledge, is able to classify all objects and properties of objects, past and present. Consider now some property Y not yet determined about the present objects ; either Y is, or it is not. If property Y is, a classification $c(Y)$ is arrived at. If Y is not, a classification $c(\sim Y)$ is arrived at. The relative probability of Y to $\sim Y$ is defined to be the relative probability of the classification $c(Y)$ to $c(\sim Y)$.

Some improbable properties Y , such as Ros being an extraterrestrial substitute, would cause a substantial change in the new classification allowing for them, but usually the only change will be in the class of instances like Y , and the class of instances like $\sim Y$. A The inductive probability for Y versus $\sim Y$ is the number of instances of *like* Y versus the number of instances of *like* $\sim Y$.

Let us return to the example of the approaching Ros, Graham, and Bia previously given. Suppose I asked myself Y : are two of the people in this group married? At a great distance, I can see that there are three people , one considerably smaller than the other two, and classification with previous experiences of such groups would give, say, 80% probability of Y ; that is, 80% of such remembered groups had two members married to each other. As they get closer, I can tell that the two larger members are of opposite gender, and the probability might go to 95%. Then when I recognise the individuals, the probability goes to 100%, minus a tiny bit for Humean uncertainty about the future.

Each probability is based on the changing classification as data accumulates. any data. It is the sharp change in classification that causes the sharp change in probabilities.

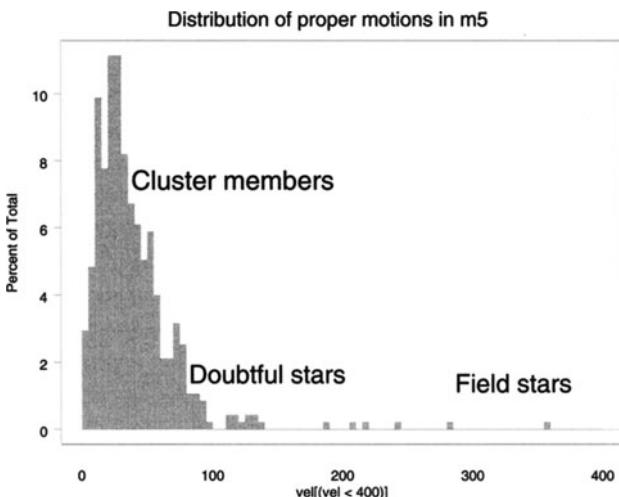
4 The probability of stars lying in the globular cluster M5

Rees(1993) derives proper motions from microdensitometer scans of 17 Yerkes refractor plates spanning an epoch range of 88 years for 515 stars of brightness exceeding that of $V = 15.6$ in the region of the globular cluster M5. Photographic photometry in B and V was obtained for these stars from these scans and scans of six Yerkes reflector plates. Membership probabilities are derived from the proper motions under the assumption that the proper motions of stars in the cluster come from a bivariate normal distribution, and the proper motions of stars in the background field come from a different bivariate nor-

mal distribution. Three stars remain with membership probabilities between 10% and 90%.

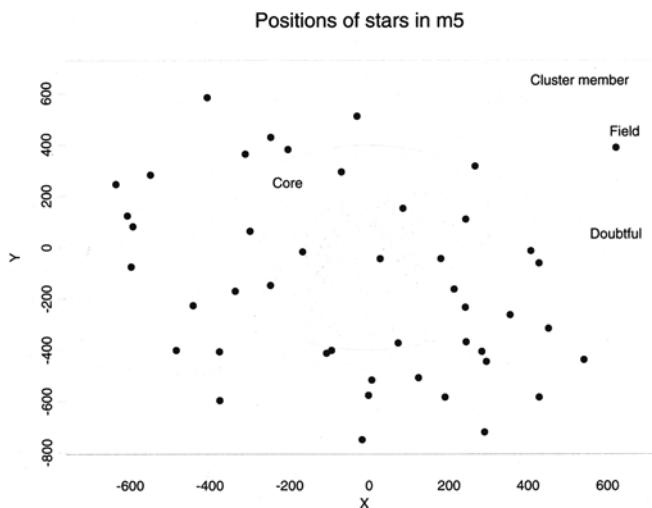
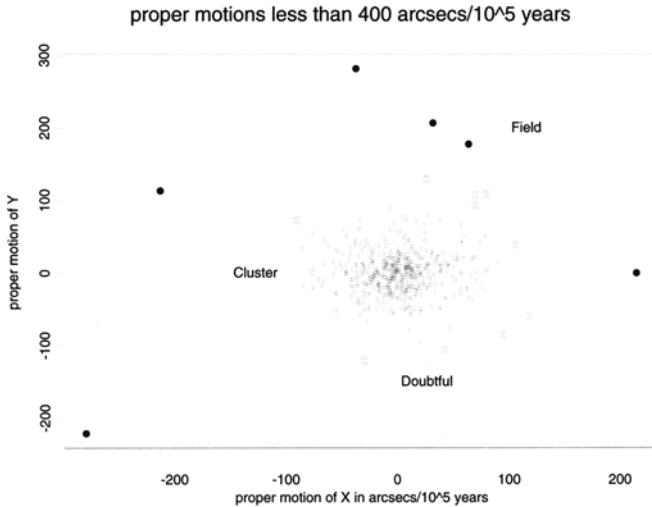
I do not wish to make assumptions about probability models. Rather, I wish to use probability to guide understanding of data, not use data to develop understanding of probabilities.

Here are the facts: the angular positions and velocities of the stars selected as candidates for members of the globular cluster M% are known, but the radial positions and velocities are not. We will place stars in the same cluster if they have nearly the same positions and velocities, but we are ignorant of 2 out of 6 of the necessary values for each star. We begin by finding all the stars, under certain magnitude constraints, that are close to the center of the cluster in angular position.(However, stars very close to the center are not included in the study, because they are so closely crowded together that their movements over time cannot be traced.) The angular positions are determined at times 88 years apart, which determines the angular velocity. Those stars with angular velocities very different from the angular velocity of the cluster center are classified as Field stars, in the background.



There is an intermediate range of stars with modest angular velocities that do not clearly belong to the Field or the Cluster. We call these Doubtful stars. Rees allows for three such stars using membership probabilities based on bivariate normal mixtures.

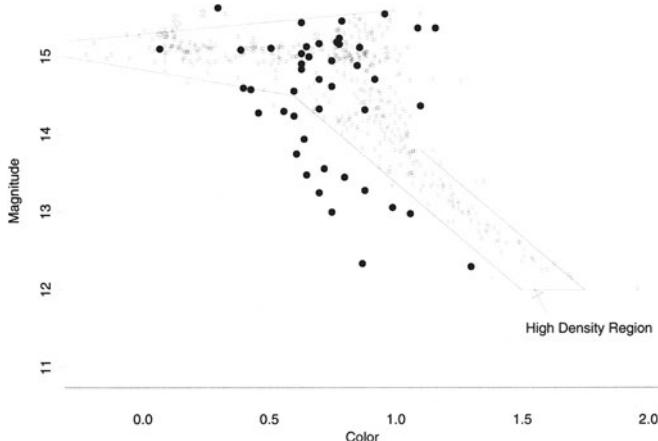
We ask what is the probability that each of these stars lies in M5? We have to take into account two pieces of evidence. First, the distribution of angular positions in M5 has a dense core, and a sparser surrounding region; the Field



stars are much more likely to be in the sparser region, and so doubtful stars are less likely to be cluster members if they are in the sparser region. The same story applies to the Magnitude Color plot of the stars. Most of the stars lie in a dense region of the plot, but the Field stars are much more likely to lie in a non-dense region. Again, doubtful stars are less likely to be cluster members if they lie in the sparser region.

For the doubtful stars, three of the ten, R29,Z209,C28 lie in the non-Core region, and two of the ten Z515, Z713 lie in the low density MCD region.

Magnitude-Color Diagram for 515 stars in M5



To compute the probabilities of cluster memberships of the various stars, we need to consider the possible classifications that the memberships imply. The classes used, with initial assignment counts, are

- 385:12 core cluster:field
- 75:33 non-core cluster:field
- 432:24 high density MCD cluster:field
- 28:21 low density MCD cluster:field

Without assignment of the doubtful stars, the probability of this classification is proportional to the product $\prod (n_i - 1)!$ over assignment counts n_i . Consider first the probable assignment of Z794 alone; this star lies in the core and in the high density MCD region. Assignment to the cluster increases the counts 385,432 by 1; assignment to the field increases the counts 12,24 by 1; thus the relative probability of cluster to field is $\frac{385}{12} \times \frac{432}{24} = 578$.

To assign all ten stars, we need to consider the 1024 possible assignments to cluster and field, and then the probability that a particular star is in the cluster is the sum over the probabilities in which the star is assigned to the cluster divided by the total probability. Since the 5 stars that are high core and high density MCD each have a high probability of being a cluster member, as was just shown, let us reduce the computation by considering the assignment of the other stars conditional on these stars being cluster members. The assignments of the remaining 5 stars are independent for the three non-Core stars and the two low density MCD stars. Considering first the non-core stars, probabilities are proportional to :

- CCC : $75 \times 76.77 \times 437.438 \times 439$

- FCC $75.76 \times 33.437 \times 438.24$
- FFC $75 \times 33.34 \times 437.24 \times 25$
- FFF $33.34 \times 35.24 \times 25 \times 26$

The probability that a particular star is assigned to the field is .022 and all three stars are assigned to the cluster with probability 0.933.

For the low density MCD stars,

- CC: $390 \times 391.28 \times 29$
- FC: 390.12×28.21
- FF: $12 \times 13.21 \times 22$

The chance of a particular star being in the field is .021, and both stars are in the cluster with probability .957. We conclude that each of the 5 doubtful stars not in both the core and the high density MCD region have probability about 1/50 of being a field star, and each of the doubtful stars in both the core and the high density MCD region have a probability about 1/600 of being a field star. The base for the probability outcome in both cases is that stars in the core and the high density MCD region are cluster members with high probability. The analysis states that more of the doubtful stars are expected to be field stars than are suggested in Rees's gaussian based analysis.

5 The next president of the United States

As of Feb 6, 2000, the person most likely to be elected President of the United States in November 2000, is the Republican Governor George W. Bush of Texas. The Democratic candidate will probably be Vice-president Al Gore. The most likely alternative Republican candidate is Senator John McCain.

Governor Bush has the backing of the Republican establishment, and inherits much political support from his father President Bush. He has a heavy advantage in fund-raising, and in nation-wide organisation over Senator McCain. Senator McCain's strategy is to concentrate money and energy on the first few primaries, in order to gather support for the later primaries. Bush won, but not by a big margin, in the Iowa caucuses, in which McCain chose not to campaign. In the New Hampshire primary on 1 February, McCain won convincingly 49% to 31%. The next primary, on 19 February, is in South Carolina. If McCain wins there, Bush will still have the long term advantage because of his national preparation and huge financial resources, but the race for the nomination will become close. Rather than attempt probabilities for president, we will consider the narrower problem of probabilities for the winner of the South Carolina primary.

Our knowledge consists of various hard facts, and other softer facts, which we need to express, together with the results of the primary, as a classification. The hard facts are the polls of the South Carolina electorate; these are

telephone polls of likely Republican primary participants in which respondents are asked who they will vote for. Typically 500-1000 respondents are used in each poll. The softer facts describe the recent history of primaries in South Carolina, and the likely affinities between the candidates and the electorate.

Polls:

Date	Organisation	McCain-Bush %
18 Nov 99	CNN-Time	15-62
27 Jan 00	Palmetto	32-54
30 Jan 00	CNN-Time	32-52
2 Feb 00	Rasmussen	40-41
2 Feb 00	Zogby	44-39
5 Feb 00	CNN-Time	44-40

South Carolina has a recent history of supporting the establishment candidate, Dole over Buchanan in 1996, Bush over Buchanan in 1992, Bush over Dole in 1988; in 1980, South Carolina did vote for the insurgent candidate Reagan over the establishment candidate Connally. South Carolina is conservative, with substantial fractions of conservative christians, and pro-military, with many veterans. Bush has more solid conservative credentials than McCain, and appeals to the religious right; McCain, a hero of the Vietnam war, appeals to the military. Bush appeals more to the Republican regulars, McCain to independents and Democrats; this is an important factor because there is no Democratic primary, so Democratic voters will be free to vote for McCain. A final fact is that Bush won the Iowa primary, McCain the New Hampshire primary.

To compute an inductive probability, we need to construct a classification representing our knowledge after McCain or Bush wins the South Carolina primary in 2000. Let us denote a McCain win by M , and a Bush win by B .

Classes:

- State primary results: \leftarrow McCain(New Hampshire), M:B=2:1
- South Carolina primary history: \leftarrow Insurgent(1980), Establishment(1988, 1992, 1996), M:B=1:3
- South Carolina constituencies: \leftarrow Religious Right(Bush), Conservatives (Bush), Independents(McCain) Military(McCain), M:B=2:2
- Polls: \leftarrow Bush(Nov, Jan, Jan, Feb), McCain(2Feb 5 Feb), M :B = 2:4
- Polls after New Hampshire primary: \leftarrow Bush(2Feb), McCain(2Feb, 5Feb), M:B= 2:1

In this classification, we treat the actual primary results as a poll. The preferences of each of the principal constituencies are also treated as poll results. We give the later polls more weight by attaching a separate class for polls after the New Hampshire primary. The Iowa caucuses have been omitted because McCain did not compete in Iowa, and the caucuses differ from regular primaries in being attended by relatively small numbers of people. The net probability for M/B is the product of all the factors associated

with individual classes, 2/3. Bush is favoured because the state history of favouring the establishment candidate overcomes the advantage of McCain in primary wins.

References

- BERNOULLI, JAMES (1713) *Ars Conjectandi*.
- DE FINETTI, B. (1973): Bayesianism: Its unifying role for both the foundations and the applications of Statistics.*Bulletin of the International Statistical Institute*, 39(4), 349–368
- HUME, DAVID(1758) *An Enquiry concerning Human Understanding*
- LOCKE, JOHN(1689) *An Essay Concerning Human Understanding*
- REES Jr. R.F.(1993) New proper motions in the globular cluster M5 *Astron. J.* 106.1524-1532

Cluster Analysis Based on Data Depth

Richard Hoberg

Seminar für Wirtschafts- und Sozialstatistik,
Universität zu Köln, D-50923 Köln, Germany
(e-mail: hoberg@snoopy.ek79.uni-koeln.de)

Abstract. A data depth $\text{depth}(y, \mathcal{X})$ measures how deep a point y lies in a set \mathcal{X} . The corresponding α -trimmed regions $D_\alpha(\mathcal{X}) = \{y : \text{depth}(y, \mathcal{X}) \leq \alpha\}$ are monotonely decreasing with α , that is $\alpha > \beta$ implies $D_\alpha \subset D_\beta$. We introduce clustering procedures based on weighted averages of volumes of α -trimmed regions. The hypervolume method turns out to be a special case of these procedures. We investigate the performance in a simulation study.

1 Introduction

The aim of cluster analysis is to divide a given set $\mathcal{X} = \{x_1, \dots, x_n\}$ of objects in k sets C_1, \dots, C_k ('clusters', 'classes') such that the clusters C_j are homogeneous. Throughout this paper, we assume $\mathcal{X} \subset \mathbb{R}^d$, the number of classes, k , to be known, and that the convex hulls $H(C_j)$ of the clusters are disjoint¹. Many clustering procedures can be described as strategies for solving a minimization problem based on a function η , which measures the inhomogeneity of the clusters. One is searching for a partition $\{C_1^*, \dots, C_k^*\}$ of \mathcal{X} which minimizes a cluster criterion $W_k(\{C_1, \dots, C_k\}) = \eta(C_1) + \dots + \eta(C_k)$ (see Bock (1974)). For example, if the data are assumed to be uniformly distributed on a union of convex sets, a maximum likelihood approach leads to the the hypervolume criterion (Hardy and Rasson (1982), Rasson and Granville (1996)):

$$W_k(\{C_1, \dots, C_k\}) := \sum_{j=1}^k \lambda^d(H(C_j)) \quad , \quad (1)$$

where λ^d denotes the d -dimensional Lebesgue measure. A disadvantage of this approach is that the best partition is strongly influenced by outliers and the corresponding clustering procedure tends to build clusters consisting of very few points. Also it is not suitable, if the underlying distribution has infinite support, e.g., in a normal mixture model, since the consistence property is lost. Note that $(\lambda^d(H(\{X_1, \dots, X_n\}))) \rightarrow \infty$ a.s. for X_1, \dots, X_n i.i.d. $\sim N(0, \sigma^2 I)$. Therefore, Ruts and Rousseeuw (1996) replaced the convex hulls $H(C_j)$ in (1) by α -trimmed regions $D_\alpha(C_j) \subset H(C_j)$. We propose here to measure the inhomogeneity of clusters by its lift zonoid volume.

¹ This is one of nine so-called admissibility conditions for clustering procedures proposed in Fisher and Van Ness (1971)

2 Data depth, trimmed regions and lift zonoids

‘Data depth’ measures the centrality of a point $y \in \mathbb{R}^d$ in a ‘data cloud’ $\mathcal{X} \subset \mathbb{R}^d$ or w.r.t. a distribution μ on \mathbb{R}^d . Various notations of data depth have been introduced by Mahalanobis (1936), Tukey (1975) and Liu et al. (1990). Here we focus on the zonoid data depth (Koshevoy and Mosler (1997b), Koshevoy and Mosler (1998), Dyckerhoff et al. (1996)).

DEFINITION Let μ be in \mathcal{M} , where \mathcal{M} is the set of measures on $(\mathbb{R}^d, \mathcal{B}^d)$ with finite expectation.

- i) For $\alpha \in (0, 1]$ define the zonoid α -trimmed region as

$$D_\alpha(\mu) := \left\{ \int xg(x)d\mu(x) : g : \mathbb{R}^d \rightarrow [0, 1/\alpha] \quad \text{and} \quad \int g(x)d\mu(x) = 1 \right\}$$

and $D_0(\mu)$ as the closure of all $D_\alpha(\mu)$, $0 < \alpha \leq 1$.

- ii) The convex compact set $\hat{Z}(\mu) \subset \mathbb{R}^{d+1}$,

$$\hat{Z}(\mu) := \left\{ \left(\int_{\mathbb{R}^d} h(x)d\mu(x), \int_{\mathbb{R}^d} xh(x)d\mu(x) \right) : h : \mathbb{R}^d \rightarrow [0, 1] \right\},$$

is called the lift zonoid of μ .

- iii) The zonoid data depth of a point $y \in \mathbb{R}^d$ w.r.t. a fixed distribution $\mu \in \mathcal{M}$ is defined as

$$\text{depth}(y|\mu) := \begin{cases} \sup\{\alpha : x \in D_\alpha(\mu)\} & \text{if } x \in D_0(\mu), \\ 0 & \text{else.} \end{cases}$$

For a set \mathcal{X} , zonoid α -trimmed regions, lift zonoid and zonoid data depth are defined as $D_\alpha(\mathcal{X}) := D_\alpha(\mu_n)$, $\hat{Z}(\mathcal{X}) := \hat{Z}(\mu_n)$ and $\text{depth}(y|\mathcal{X}) := \text{depth}(y|\mu_n)$, where μ_n is the empirical distribution of \mathcal{X} . Some basic properties of zonoid data depth, zonoid trimmed regions and lift zonoids are listed in the following proposition (see Koshevoy and Mosler (1997b), Koshevoy and Mosler (1998) for details).

PROPOSITION

- (i) A distribution $\mu \in \mathcal{M}$ is uniquely determined by its lift zonoid.
- (ii) The lift zonoid is continuous with respect to weak convergence resp. Hausdorff convergence, $\mu_n \rightarrow^w \mu$ implies $\hat{Z}(\mu_n) \rightarrow_H \hat{Z}(\mu)$.
- (iii) $D_\alpha(\mu) = \frac{1}{\alpha} \text{proj}_\alpha \hat{Z}(\mu) = \frac{1}{\alpha} \{ \zeta \in \mathbb{R}^d : (\alpha, \zeta) \in \hat{Z}(\mu) \}$.
- (iv) $D_\alpha(\mu) \subset D_\beta(\mu)$ if $\alpha > \beta$.
- (v) $D_0(\mu)$ is the convex hull of the support of μ , $D_1(\mu) = \{ \int x d\mu(x) \}$ is the singleton containing the expectation value of μ .

- (vi) $D_\alpha(N(a, \Sigma)) = \{x : (x - a)^T \Sigma^{-1} (x - a) \leq r_\alpha^2\}$.
(vii) $\text{vol}(\hat{Z}(\mathcal{X})) = \sum_{1 \leq i_1 < \dots < i_{d+1} \leq n} \left| \det \left(\frac{1}{n}(1, x_{i_1}), \dots, \frac{1}{n}(1, x_{i_{d+1}}) \right) \right|$.

The *depth* of y w.r.t. μ is therefore a real number between zero and one. Great values of $\text{depth}(y|\mu)$ indicate that y is near the expectation of μ . The corresponding α -trimmed regions, consisting of all points y with $\text{depth}(y|\mu) \geq \alpha$ are decreasing with α . Figure 1 shows the α -trimmed regions of 30 points sampled from a normal distribution ($\alpha = \frac{1}{30}, \frac{2}{30}, \dots, \frac{29}{30}$). The zonoid trimmed regions can be calculated in polynomial time (Dyckerhoff (2000)).

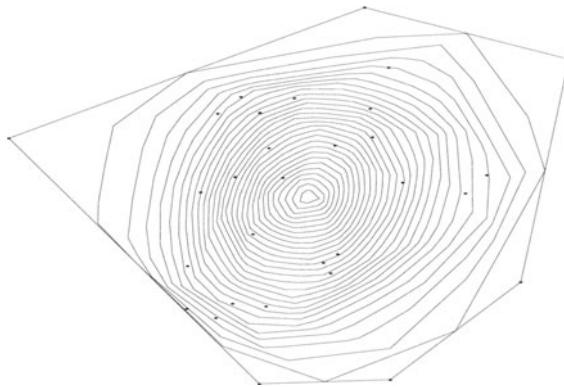


Fig. 1. α -trimmed regions for 30 points

3 The cluster criterion

Now we can define the class of inhomogeneity measures. Since dispersed data lead to larger α -trimmed regions, inhomogeneity or dispersion can be measured by the volumes of the α -trimmed regions. Our idea is to use a weighted sum of volumes of trimmed regions.

DEFINITION of (ψ, ν) -inhomogeneity: Let $\nu = \lambda^1$ or $\nu(\{\alpha_1, \dots, \alpha_l\}) = 1$ and $\psi : [0, 1] \rightarrow [0, 1]$ be strictly increasing. Then

$$\eta_\nu^\psi(\mathcal{X}) := \int_0^1 \psi(\alpha) \text{vol}(D_\alpha(\mathcal{X})) d\nu(\alpha) \quad (2)$$

is called the (ψ, ν) -inhomogeneity of \mathcal{X} .

Summing up the inhomogeneities of the clusters C_j for a given partition $\{C_1, \dots, C_k\}$, our cluster criterion turns out to be

$$W_k(\{C_1, \dots, C_k\}) := \sum_{j=1}^k \eta_\nu^\psi(C_j) \quad . \quad (3)$$

REMARK (Special cases)

- a) Choosing $\nu := \delta_0, \psi(\alpha) := 1$ one gets the hypervolume criterion,
- b) $\nu := \delta_{\alpha_0}, \psi(\alpha) := 1$ leads to the criterion of Ruts and Rousseeuw (1996) and
- c) with $\nu := \lambda^1, \psi(\alpha) := \alpha^d$ the corresponding inhomogeneity $\eta_\nu^\psi(\mu)$ is called lift zonoid inhomogeneity. In fact,

$$\begin{aligned}\eta_Z(X) := \eta_\nu^\psi(\mu) &:= \int_0^1 \psi(\alpha) \text{vol}(D_\alpha(\mu)) d\nu(\alpha) = \int_0^1 \alpha^d \text{vol}(D_\alpha(\mu)) d\alpha \\ &= \int_0^1 \lambda^d(\alpha D_\alpha(\mu)) d\alpha = \int_0^1 \lambda^d(\text{proj}_\alpha(\hat{Z}(\mu))) d\alpha = \lambda^{d+1}(\hat{Z}(\mu)) .\end{aligned}$$

The latter inhomogeneity measure was used by Koshevoy and Mosler (1997a) for measuring economic disparity. Note that for $\mu \in \mathcal{M}$ the compactness and the continuity property (see the Proposition) guarantee consistency. Figure 2

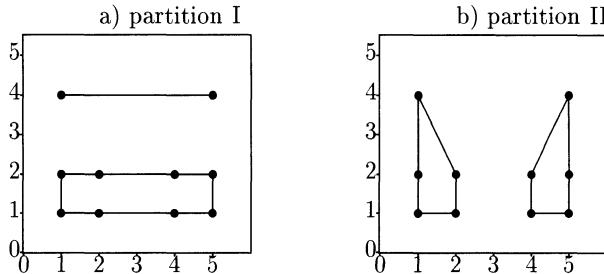


Fig. 2. Two possible convex partitions of 10 points

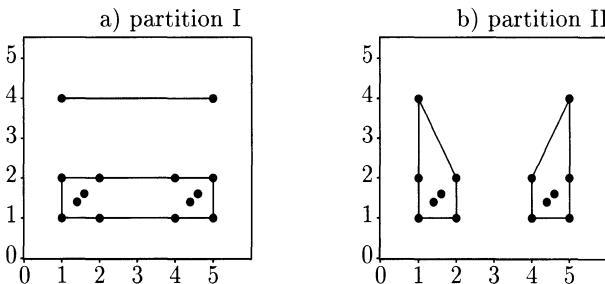


Fig. 3. Fig. 2 with four additional interior points

and 3 demonstrate a fundamental difference between the hypervolume and

the lift zonoid criterion: if four interior points are added (figure 3), the hypervolume criterion remains the same (value=5 in each situation), while the lift zonoid criterion in figure 2 prefers partition I ($W_k = 0.219$ vs. $W_k = 0.24$), and in figure 3 partition II ($W_k = 0.185$ vs. $W_k = 0.180$).

4 A small simulation study

In order to compare the different criteria of section 3, we conducted some simulations. Each sample (x_1, \dots, x_n) was generated in two steps: 1. choose randomly points y_1, \dots, y_n according to a uniform distribution on two disjoint triangles, 2. add to these points y_i some noise $\epsilon_i \sim N(0, \sigma^2 I)$. $x_i := y_i + \epsilon_i$. By the first step, a ‘correct’ 2-partition $\{\hat{C}_1, \hat{C}_2\}$ of \mathcal{X} is defined (depending on in which triangle y_i lies) (see figure 4). For five clustering procedures and

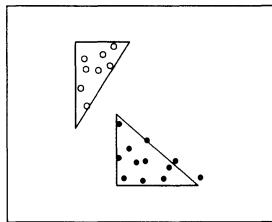


Fig. 4. 20 points in two triangles + noise

different ‘noise parameters’ σ we compared the proportion of ‘correctly’ classified points. The results are reported in table 1 (in each situation, 20 samples consisting of 20 points were generated). We determined the best partition for the depth-based methods by testing each convex 2-partition. In this simulation the k-means algorithms was the best for all values of σ . Generally the hypervolume-method tends to form very small groups consisting of ‘outliers’. This is avoided by the lift zonoid criterion and the criterion of Ruts and Rousseeuw (1996), because there the volume of the convex hull of the sample has not so much influence.

5 Conclusion

The lift zonoid criterion seems to be a good alternative for clustering in the case of slightly noisy data, especially if the original data points are assumed to be uniformly distributed on disjoint compact sets. Of course, in a future simulation study sample size and number of simulations have to be increased. But searching for the best k -partition is very time consuming. Although

there exists a polynomial time bound (the number of convex k -partitions is polynomial in n), one has to use local search strategies like genetic algorithms, simulated annealing etc. for larger data sets.

Inhomogeneity	$\sigma_{0.00}$	$\sigma_{0.04}$	$\sigma_{0.08}$	$\sigma_{0.18}$	$\sigma_{0.28}$
$\eta(\mathcal{X}) = \text{vol}(H(\mathcal{X}))$	97.25	96	94.25	81	69
$\eta(\mathcal{X}) = \text{vol}(\hat{Z}(\mathcal{X}))$	98	97.25	95	87	77.65
$\eta(\mathcal{X}) = \text{vol}(D_{0.1}(\mathcal{X}))$	97	96.5	94.25	79.5	73
$\eta(\mathcal{X}) = \int_0^1 \text{vol}(D_\alpha(\mu)) d\nu(\alpha)$	97.75	97	96	83.75	72
k-means	98.25	97.55	95.75	88.75	82

Table 1. Proportion of correctly classified points

References

- BOCK, H.-H. (1974): *Automatische Klassifikation*. Vandenhoeck & Ruprecht, Göttingen.
- DYCKERHOFF, R., KOSHEVOY, G. and MOSLER, K. (1996): Zonoid Data Depth: Theory and Computation. In A. Pratt (Ed.): *Proceedings in Computational Statistics*, Physica, Heidelberg, 235–240.
- DYCKERHOFF, R. (2000): Computing Zonoid Trimmed Regions of Bivariate Data Sets, *COMPSTAT 2000 – Proceedings in Computational Statistics (to appear)*.
- FISHER, L. and VAN NESS, J.W. (1971): Admissible Clustering Procedures. *Biometrika*, 58, 91–104.
- HARDY, A. and RASSON, J.-P. (1982): Une Nouvelle Approche des Problèmes de Classification Automatique. *Statistique et Analyse des Données*, 7, 41–56.
- KOSHEVOY, G. and MOSLER, K. (1997a): Multivariate Gini Indices. *Journal of Multivariate Analysis*, 60, 252–276.
- KOSHEVOY, G. and MOSLER, K. (1997b): Lift Zonoid Trimming for Multivariate Distributions. *Annals of Statistics*, 25, 1998–2017.
- KOSHEVOY, G. and MOSLER, K. (1998): Lift Zonoids, Random Convex Hulls and the Variability of Random Vectors. *Bernoulli*, 4, 377–399.
- LIU, R.Y., PARELIUS, J.M., and SINGH, K. (1990): On a Notion of Data Depth Based on Random Simplices. *Annals of Statistics*, 18, 405–414.
- MAHALANOBIS, P.C. (1936): On the Generalized Distance in Statistics, *Proceedings of National Academy India*, 12, 49–55.
- RASSON, J.-P. and GRANVILLE, V. (1996): Geometrical Tools in Classification, *Computational Statistics and Data Analysis*, 23, 105–123.
- RUTS, I. and ROUSSEEUW, P.J. (1996): Computing Depth Contours of Bivariate Point Clouds, *Computational Statistics and Data Analysis*, 23, 153–168.
- TUKEY, J.W. (1975): Mathematics and Picturing of Data, In: R.D. James (Ed.): *The Proceedings of the International Congress of Mathematicians Vancouver*, 523–531.

An Autonomous Clustering Technique

Yoshiharu Sato

Division of Systems and Information, Hokkaido University,
Kita 13, Nishi 8, Kita-ku, Sapporo, 060-8628, Japan
(e-mail: ysato@main.eng.hokudai.ac.jp)

Abstract. The basic idea of this paper is that it will be possible to construct clusters by moving each pair of objects closer or farther according to their relative similarity to all of the objects. For this purpose, regarding a set of objects as a set of autonomous agents, each agent decides its action to the other agents by taking account of the similarity between its self and others. And consequently, we get the clusters autonomously.

1 Introduction

In cluster analysis, the following three methods have been frequently appearing in practical applications. The first is a hierarchical method, which represents the clustering process by a dendrogram (Lance, et al., 1967). This method has been most frequently used for it is intuitively easy to understand. The second one is the k-means method (Hartigan, 1975). This is a fundamental algorithm for dividing a region into a given number of subregions in such a way that it minimizes the sum of within variances. The third is a mixture model in which the clusters are assumed to be probabilistic distributions, usually normal distributions (Bock, 1996). From the observed data, each distribution is estimated by the parametric or non-parametric method.

In this paper, we propose an algorithm for autonomously constructing clusters. We regard the objects as autonomous agents, who change their similarity based on the relative similarity relation. The process of the change of the similarity is repeated until the similarity converges 0 or 1. This process is considered to be an autonomous clustering.

2 Action rule for autonomous agents

We suppose that the observed similarity between n objects is given by

$$S = (s_{ij}), \quad 0 \leq s_{ij} \leq 1, \quad s_{ij} = s_{ji}, \quad s_{ii} = 1. \quad (1)$$

We assume that the initial state of the agents is given by the points in a configuration space generated by the observed similarity. The dimension of the configuration space will be less than n . We also assume that the agents can move to any directions. Since it is impossible to construct the expected

clusters without any restriction of the behavior of the agents, we introduce an action rule for the agents.

The actions of the agents are determined as follows: Looking over the configuration from each agent, when If two agents have the relatively similar positions, they move closer. Otherwise they go away from each other. The relative positions of two objects, o_i and o_j , are represented by the two column vectors, \mathbf{s}_i and \mathbf{s}_j , of the similarity matrix S ,

$$\begin{aligned}\mathbf{s}'_i &: (s_{1i}, s_{2i}, \dots, 1, \dots, s_{ji}, \dots, s_{ni}) \\ \mathbf{s}'_j &: (s_{1j}, s_{2j}, \dots, s_{ij}, \dots, 1, \dots, s_{nj})\end{aligned}$$

If these two vectors are similar each other, then two agents, o_i and o_j , move closer. This moving is implemented by increasing the similarity between o_i and o_j . Repeating this action, we can get the clusters.

Formally, this action is denoted as follows: Suppose $S^{(0)} = (s_{ij}^{(0)})$ and $S^{(t)} = (s_{ij}^{(t)})$ denote the observed similarity and the similarity at t -step, respectively, the action from t -step to $(t+1)$ -step is defined by

$$s_{ij}^{(t+1)} = \sum_{k=1}^n (s_{ik}^{(t)})^\alpha (s_{jk}^{(t)})^\alpha / \left\{ \left(\sum_{\ell=1}^n (s_{i\ell}^{(t)})^{2\alpha} \right) \left(\sum_{m=1}^n (s_{jm}^{(t)})^{2\alpha} \right) \right\}^{\frac{1}{2}}, \quad (2)$$

where the parameter α is assumed to be greater than 1, and it play the important role to get non-trivial clusters.

Using matrix notations, (3) is expressed as follows: Putting

$$S^{(t)} = [\mathbf{s}_1^{(t)}, \mathbf{s}_2^{(t)}, \dots, \mathbf{s}_n^{(t)}], \quad S^{(t)\alpha} = [\mathbf{s}_1^{(t)\alpha}, \mathbf{s}_2^{(t)\alpha}, \dots, \mathbf{s}_n^{(t)\alpha}], \text{ and}$$

$$D^{(t)\alpha} = \text{diag}[\mathbf{s}_1^{(t)\alpha} \mathbf{s}_1^{(t)\alpha}, \mathbf{s}_2^{(t)\alpha} \mathbf{s}_2^{(t)\alpha}, \dots, \mathbf{s}_n^{(t)\alpha} \mathbf{s}_n^{(t)\alpha}] \quad (3)$$

Then (3) is denoted as

$$\begin{aligned}S^{(1)} &= D^{(0)\alpha-\frac{1}{2}} S^{(0)\alpha} S^{(0)\alpha} D^{(0)\alpha-\frac{1}{2}} \\ S^{(2)} &= D^{(1)\alpha-\frac{1}{2}} S^{(1)\alpha} S^{(1)\alpha} D^{(1)\alpha-\frac{1}{2}} \\ &\vdots \\ S^{(t+1)} &= D^{(t)\alpha-\frac{1}{2}} S^{(t)\alpha} S^{(t)\alpha} D^{(t)\alpha-\frac{1}{2}} \\ &\vdots\end{aligned} \quad (4)$$

We shall show the convergence of the above sequence (4).

Proposition 2.1. Let $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$ be the column vectors of S . A region C^α , defined by

$$C^\alpha = \{v | v = \lambda_1 \mathbf{s}_1^\alpha + \lambda_2 \mathbf{s}_2^\alpha + \dots + \lambda_n \mathbf{s}_n^\alpha, \lambda_i \geq 0, \alpha \geq 1\}$$

is convex cone in n -dimensional space E^n .

Proposition 2.2. If $|S^\alpha| \neq 0$, C^α contains a vector $\mathbf{1}' = (1, 1, \dots, 1)$.

Proof. We assume that each column vector \mathbf{s}_i is denoted by

$$\mathbf{s}_i^\alpha = s_{1i}^\alpha \mathbf{e}_1 + s_{2i}^\alpha \mathbf{e}_2 + \dots + s_{ni}^\alpha \mathbf{e}_n,$$

where $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is an orthogonal base in E^n . Let T be such a orthogonal matrix that

$$T' = [\frac{1}{\sqrt{n}} \mathbf{1}', \boldsymbol{\ell}'_2, \dots, \boldsymbol{\ell}'_n], \quad \boldsymbol{\ell}'_k \mathbf{1} = 0.$$

Using T , one of the bases $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ is transformed to the direction of $\mathbf{1}$ and

$$T'[\mathbf{s}_1^\alpha, \mathbf{s}_2^\alpha, \dots, \mathbf{s}_n^\alpha] = \begin{bmatrix} \frac{1}{\sqrt{n}} \mathbf{1}' \mathbf{s}_1^\alpha & \frac{1}{\sqrt{n}} \mathbf{1}' \mathbf{s}_2^\alpha & \cdots & \frac{1}{\sqrt{n}} \mathbf{1}' \mathbf{s}_n^\alpha \\ \boldsymbol{\ell}'_2 \mathbf{s}_1^\alpha & \boldsymbol{\ell}'_2 \mathbf{s}_2^\alpha & \cdots & \boldsymbol{\ell}'_2 \mathbf{s}_n^\alpha \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\ell}'_n \mathbf{s}_1^\alpha & \boldsymbol{\ell}'_n \mathbf{s}_2^\alpha & \cdots & \boldsymbol{\ell}'_n \mathbf{s}_n^\alpha \end{bmatrix}. \quad (5)$$

If C^α does not contain $\mathbf{1}$, then there exists at least one direction $\boldsymbol{\ell}_n$ such that

$$\boldsymbol{\ell}'_n \mathbf{s}_1^\alpha = \boldsymbol{\ell}'_n \mathbf{s}_2^\alpha = \cdots = \boldsymbol{\ell}'_n \mathbf{s}_n^\alpha = 0. \quad (6)$$

Using $\mathbf{1}' \boldsymbol{\ell}_k = 0$, (6) is written by

$$\begin{bmatrix} 1 - s_{21}^\alpha & s_{23}^\alpha - s_{21}^\alpha & \cdots & s_{2n}^\alpha - s_{21}^\alpha \\ s_{32}^\alpha - s_{31}^\alpha & 1 - s_{31}^\alpha & \cdots & s_{3n}^\alpha - s_{31}^\alpha \\ \vdots & \vdots & \ddots & \vdots \\ s_{n2}^\alpha - s_{n1}^\alpha & s_{n3}^\alpha - s_{n1}^\alpha & \cdots & 1 - s_{n1}^\alpha \end{bmatrix} \begin{bmatrix} \ell_{n2} \\ \ell_{n3} \\ \vdots \\ \ell_{nn} \end{bmatrix} = \Psi^\alpha \boldsymbol{\ell}_n = 0$$

Hence, if $|S^\alpha| \neq 0$, then $|\Psi^\alpha| \neq 0$. Therefore, there exists no vector except $\boldsymbol{\ell}_n = 0$. If $|S^\alpha| = 0$, we may consider the same in a subspace spanned by the column vectors of S^α .

Proposition 2.3. The product $S^{\alpha'} S^\alpha$ in (4) can be regarded as a linear transformation of S^α by $S^{\alpha'}$. If we denote

$$[\tilde{\mathbf{s}}_1^\alpha, \tilde{\mathbf{s}}_2^\alpha, \dots, \tilde{\mathbf{s}}_n^\alpha] = S^{\alpha'}[\mathbf{s}_1^\alpha, \mathbf{s}_2^\alpha, \dots, \mathbf{s}_n^\alpha]$$

and

$$\tilde{\mathcal{C}}^\alpha = \{\tilde{\mathbf{v}} | \tilde{\mathbf{v}} = \lambda_1 \tilde{\mathbf{s}}_1^\alpha + \lambda_2 \tilde{\mathbf{s}}_2^\alpha + \cdots + \lambda_n \tilde{\mathbf{s}}_n^\alpha, \lambda_i \geq 0, \alpha \geq 1\},$$

then

$$\tilde{\mathcal{C}}^\alpha \subseteq \mathcal{C}^\alpha$$

Proof. If we denote

$$\mathbf{s}_i^\alpha = s_{1i}^\alpha \mathbf{e}_1 + s_{2i}^\alpha \mathbf{e}_2 + \cdots + s_{ni}^\alpha \mathbf{e}_n, \quad \mathbf{e}'_k = (0, \cdots, 1, \cdots, 0),$$

then

$$S^{\alpha'} \mathbf{s}_i^\alpha = s_{1i}^\alpha (S^{\alpha'} \mathbf{e}_1) + s_{2i}^\alpha (S^{\alpha'} \mathbf{e}_2) + \cdots + s_{ni}^\alpha (S^{\alpha'} \mathbf{e}_n).$$

Since $S^{\alpha'}$ is a symmetric matrix and $S^{\alpha'} \mathbf{e}_i = \mathbf{s}_i^\alpha$, we get

$$S^{\alpha'} \mathbf{s}_i^\alpha = s_{1i}^\alpha \mathbf{s}_1^\alpha + s_{2i}^\alpha \mathbf{s}_2^\alpha + \cdots + s_{ni}^\alpha \mathbf{s}_n^\alpha.$$

that is, $S^{\alpha'} s_i^\alpha$ is a linear combination of $\{s_1^\alpha, s_2^\alpha, \dots, s_n^\alpha\}$ with positive coefficients. Hence, each s_i^α is transformed into \mathcal{C}^α .

Proposition 2.4. If S^α is partitioned as

$$S^\alpha = \begin{bmatrix} S_{11}^\alpha & 0 \\ 0 & S_{22}^\alpha \end{bmatrix},$$

then the product $S^{\alpha'} S^\alpha$ has the same partition.

Theorem 2.1. If $\alpha = 1$ and S is not partitioned as Proposition 2.4, then the sequence in (4) converges to $\mathbf{1}\mathbf{1}'$, i.e.

$$S^{(1)}, S^{(2)}, \dots \dots \Rightarrow \mathbf{1}\mathbf{1}'.$$

Proof. Using the diagonal matrix in (3), if we denote

$$S^{(t)} D^{(t)-\frac{1}{2}} = [\xi_1^{(t)}, \xi_2^{(t)}, \dots, \xi_n^{(t)}],$$

then $\|\xi_k^{(t)}\| = 1$ ($k = 1, 2, \dots, n$). From the Proposition 2.3, if we put

$$\begin{aligned} S^{(t)'} \{S^{(t)} D^{(t)-\frac{1}{2}}\} &= [\zeta_1^{(t)}, \zeta_2^{(t)}, \dots, \zeta_n^{(t)}], \\ \mathcal{R}^{(t)} &= \{\nu | \nu = \lambda_1 \zeta_1^{(t)} + \lambda_2 \zeta_2^{(t)} + \dots + \lambda_n \zeta_n^{(t)}, \lambda_i \geq 0\}, \end{aligned}$$

then

$$\mathcal{R}^{(1)} \supseteq \mathcal{R}^{(2)} \supseteq \dots \supseteq \mathcal{R}^{(t)} \supseteq \dots$$

And also, since the linear transformation by $D^{(t)-\frac{1}{2}}$ is an affine transformation, the ratio is invariant. Hence, if we write

$$D^{(t)-\frac{1}{2}} \{\mathcal{R}^{(t)}\} = \{u | u = D^{(t)-\frac{1}{2}} v, \forall v \in \mathcal{R}^{(t)}\} = \bar{\mathcal{R}}^{(t)},$$

then

$$\bar{\mathcal{R}}^{(1)} \supseteq \bar{\mathcal{R}}^{(2)} \supseteq \dots \supseteq \bar{\mathcal{R}}^{(t)} \supseteq \dots$$

And from Proposition 2.2, $\bigcap_{t \rightarrow \infty} \bar{\mathcal{R}}^{(t)} = \mathbf{1}$. Therefore, $S^{(t)}$ converges to $\mathbf{1}\mathbf{1}'$. Note: if S is partitioned as Proposition 2.4, we can apply Theorem 2.1 to each submatrix S_{ii} .

Proposition 2.5. If we put

$$\begin{aligned} \mathcal{C} &= \{v | v = \lambda_1 s_1 + \lambda_2 s_2 + \dots + \lambda_n s_n, \lambda_i \geq 0, \}, \\ \mathcal{C}^\alpha &= \{w | w = \lambda_1 s_1^\alpha + \lambda_2 s_2^\alpha + \dots + \lambda_n s_n^\alpha, \lambda_i \geq 0, \alpha \geq 1\}, \end{aligned}$$

then $\mathcal{C}^\alpha \supseteq \mathcal{C}$.

Proof. If s_i^α is a linear combination of $\{s_1, s_2, \dots, s_n\}$ with non-negative coefficients, i.e.

$$s_i^\alpha = \lambda_1 s_1 + \lambda_2 s_2 + \dots + \lambda_i s_i + \dots + \lambda_n s_n, (\lambda_k \geq 0),$$

then

$$\begin{aligned}
\lambda_1 s_{11} + \lambda_2 s_{12} + \cdots + \lambda_i s_{1i} (1 - s_{1i}) + \cdots + \lambda_n s_{1n} &= 0 \\
\lambda_1 s_{21} + \lambda_2 s_{22} + \cdots + \lambda_i s_{2i} (1 - s_{2i}) + \cdots + \lambda_n s_{2n} &= 0 \\
&\dots \\
\lambda_1 s_{n1} + \lambda_2 s_{n2} + \cdots + \lambda_i s_{ni} (1 - s_{ni}) + \cdots + \lambda_n s_{nn} &= 0
\end{aligned}$$

Hence, because of $0 \leq s_{ij} \leq 1$, all of the λ_k ($k = 1, 2, \dots, n$) could not be positive. Therefore, $\mathcal{C}^\alpha \supseteq \mathcal{C}$.

Proposition 2.6. Suppose $S = [s_1, s_2, \dots, s_n]$ and $S^{(t)} = [s_1^{(t)}, s_2^{(t)}, \dots, s_n^{(t)}]$. If s_1 and s_2 are given by

$$s'_1 = [1, a, 0, \dots, 0], \quad s'_2 = [a, 1, 0, \dots, 0], \quad s_k \neq \mathbf{1}_n, (k = 3, \dots, n),$$

then $s_1^{(t)}$ converges as follows:

$$\begin{aligned}
\alpha = 2 : s_1^{(t)'} &\rightarrow (1, 1, 0, \dots, 0) \quad \text{for } 0 \leq a \leq 1 \\
\alpha = 3 : s_1^{(t)'} &\rightarrow (1, 0, 0, \dots, 0) \quad \text{for } 0 \leq a < \sqrt{(-1 + \sqrt{5})/2} \\
s_1^{(t)'} &\rightarrow (1, 1, 0, \dots, 0) \quad \text{for } \sqrt{(-1 + \sqrt{5})/2} \leq a \leq 1 \\
\alpha = 4 : s_1^{(t)'} &\rightarrow (1, 0, 0, \dots, 0) \quad \text{for } 0 \leq a < 0.87 \text{ (approximate)} \\
s_1^{(t)'} &\rightarrow (1, 1, 0, \dots, 0) \quad \text{for } 0.88 < a \leq 1 \text{ (approximate)}
\end{aligned}$$

Theorem 2.2. For $\alpha > 1$ the sequence (4),

$$S^{(1)}, S^{(2)}, \dots, S^{(t)}, \dots$$

converges to $\mathbf{1}\mathbf{1}'$ or the matrix of the type in Proposition 2.4. i.e.

$$\begin{bmatrix} U_{11} & 0 & \cdots & 0 \\ 0 & U_{22} & \cdots & 0 \\ 0 & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & U_{kk} \end{bmatrix}, \quad (k \leq n), \quad U_{ii} = \mathbf{1}_{m_i} \mathbf{1}'_{m_i}, \quad (m_1 + \cdots + m_k = n) \quad (7)$$

Proof. From Proposition 2.5 and 2.6, the components of the initial vector s_i tend to 0 or 1 according to the value of α . And using Proposition 2.4, $S^{(t)}$ converges to the matrix in (7)

The following is a typical example of the clustering using the above action rule. The data, in Fig. 1, are generated from three normal distributions. The similarity matrix $S = (s_{ij})$ is calculated from a Euclidean distance matrix $D = (d_{ij})$ as follows:

$$s_{ij} = 1 - d_{ij} / \max_{k,\ell} \{d_{k\ell}\}.$$

Figure 2 shows the process of autonomous clustering and the convergence to 3 clusters. The value of alpha is chosen as 3, in the example. The action is terminated by using the following matrix norm,

$$\|S^{(t+1)} - S^{(t)}\| \leq \varepsilon,$$

where $\varepsilon > 0$ is sufficiently small, for instance, $\varepsilon = 1.0 \times 10^{-12}$.

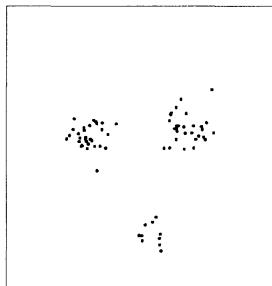


Fig.1. Three normal clusters

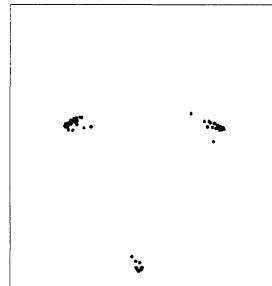


Fig.2-1. 1st step of clustering for Fig.1

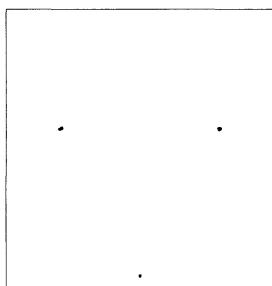


Fig. 2-2. 2nd step of clustering for Fig.1

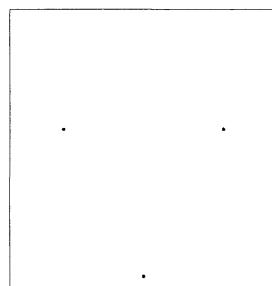


Fig. 2-3. 3rd step of clustering for Fig.1

3 Concluding remarks

In the autonomous clustering, the merits seem to be that the number of clusters and any initial partitions are not required. But we must select a suitable α . When $\alpha_1 > \alpha_2 > 1$, the obtained number of clusters for α_1 is greater than for α_2 . But the result for α_1 is not a refinement of the result for α_2 , in general.

The example in this paper is illustrated in two dimensional configuration space. However, two dimension are not essential, because, the method requires only similarity data, whose dimension of the configuration space, determined by a multidimensional scaling, is arbitrary.

References

- BOCK, H.H.(1996): Probability Models and Hypotheses Testing in Partitioning Cluster Analysis. In: P. Arabie, L.J. Hubert and G. De Soete (Eds.): *Clustering and Classification*. World Scientific Publ., 377-453.
- HARTIGAN, J.A.(1975): *Clustering Algorithms*. John Wiley & Sons, New York.
- LANCE, G.N. and WILLIAMS, W.T.(1967): A General Theory of Classificatory Sorting Strategies I, Hierarchical Systems. *Computer Journal*, 9, 373-380.

Unsupervised Non-hierarchical Entropy-based Clustering

M. Jardino

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur,
BP 133 - F91403 ORSAY Cedex(France),
(e-mail: jardino@limsi.fr)

Abstract. We present an unsupervised non-hierarchical clustering which realizes a partition of unlabelled objects in K non-overlapping clusters. The interest of this method rests on the convexity of the entropy-based clustering criterion which is demonstrated here. This criterion permits to reach an optimal partition independently of the initial conditions, with a step by step iterative Monte-Carlo process. Several data sets serve to illustrate the main properties of this clustering.

1 Theory

The automatic clustering described here is a partitioning method for a group of objects in K clusters. It permits to get clusters with not well defined borders, unlike the hierarchical partitioning (Duda(1973)).

Main features of the present clustering method. It is a non-hierarchical clustering similar to K-means (Celeux(1989)), in the following sense: each object is compared to the dynamic representation of the K clusters along the optimization process, in terms of a within-group distance. Each cluster is represented by its center of gravity defined as in Lerman and Tallur (1980). It differs from K-means in two ways. First, it is a step by step process which moves at once, only one object from a cluster to another one, so that the cluster representation can be updated before the comparison. Secondly, this comparison is performed with a non-symmetrical entropy-based distance, whereas symmetrical distances like euclidean ones are more generally employed.

A very similar method has been used for the first time to group words (Kneser(1993)), The algorithm presented here is a faster version which uses the convexity feature of the entropy-based optimization criterion, demonstrated further for the first time. Before describing the mathematical features of our approach, we present a brief description of our partitional clustering.

Clustering algorithm.

- 1- define a priori, K, the cluster number.
- 2- initialize: put all objects in the same cluster, calculate the entropy (equation 6 below).
- 3- do the random selection of one object and of another cluster for this object.
- 4- put the object in this new cluster, calculate the new entropy.

- 5- If the new entropy is lower, let the object in its new cluster, otherwise put back the object in its initial cluster.
 6- repeat until there is no more change.

Object representation. Objects are represented by a set of J vectors, y_j , which can be the results of polls or experiments over a set of I variables, x_i , each vector element will be named f_{ij} . We assume that the elements f_{ij} belong to the R space. Each vector element can be normalized by the marginal frequency. Doing that, we obtain a discrete distribution representation of the vector y_j which can be viewed as the discrete distribution of the conditional probabilities to have observed the variables x_i knowing y_j . In order to simplify the writing, we will name $q(i|j)$ these probabilities instead of $q(x_i|y_j)$, so that :

$$q(i|j) = f_{ij} / \sum_i f_{ij} \quad (1)$$

with the probability sum rule, $\sum_i q(i|j) = 1$.

Cluster representation. A partition of the J objects in a fixed number K of clusters ($K \leq J$), is represented by the vector set c_k , with each vector element named f_{ik} . Each vector c_k is obtained by merging vectors y_j and summing their components, so that :

$$f_{ik} = \sum_{j \in k} f_{ij} \quad (2)$$

A new distribution p is obtained which is defined by :

$$p(i|j \in c_k) = p(i|k) = f_{ik} / \sum_{i,k} f_{ik} \quad (3)$$

Entropy-based clustering criteria. The overall distribution q of the J objects is measured in the information theory framework (Cover(1991)) by its conditional entropy, $H(q)$, defined by:

$$H(q) = - \sum_{i,j} q(i,j) * \log[q(i|j)] \quad (4)$$

where $q(i,j)$ is the joint probability:

$$q(i,j) = f_{ij} / \sum_{i,j} f_{ij} \quad (5)$$

The great interest of entropy is that it is a bounded variable : its lower value is 0 (assuming $0 * \log(0) = 0$), its upper value is $\log I$, this value corresponds to a uniform distribution over the variables x_i . A discriminative distribution will have an entropy value closer to 0 than to $\log I$.

Another point of view is to consider the exponential of entropy. It varies from

1 to I, and can be seen as the average number of variables which are necessary to describe the observations, the exponential of the entropy is usually named perplexity in language modelling (Jelinek(1998)).

In this framework, the conditional entropy $H(p)$ of the K clusters is :

$$H(p) = - \sum_{i,k} p(i,k) * \log[p(i|k)] \quad (6)$$

It can be shown by the log-sum rule (Cover(1991)) that the entropy $H(p)$ is always higher or equal to $H(q)$. This rule gives (7):

$$p(i,k) * \log[p(i|k)] = \frac{\sum_{j \in k} f_{ij}}{\sum_{i,j} f_{ij}} * \log\left(\frac{\sum_{j \in k} f_{ij}}{\sum_{i,j \in k} f_{ij}}\right) \leq \frac{1}{\sum_{i,j} f_{ij}} \sum_{j \in k} f_{ij} * \log\left(\frac{f_{ij}}{\sum_i f_{ij}}\right)$$

Summing over i and k , this becomes $H(p) \geq H(q)$. So any clustering process increases the initial entropy, the gap between the distributions q and p , $\delta(H) = H(p) - H(q)$ is the Kullback-Leibler divergence between the distributions p and q . It is always positive and if we want to get the distribution p which is the closest to the distribution q we have to minimize $H(p)$.

Convexity of $H(p)$. In order to demonstrate the convexity of our criteria, we can observe that:

$$H(p) = - \sum_{i,k,j \in k} q(i,j) * \log[p(i|k)] \quad (7)$$

because $p(i|k)$ is the same for all $j \in k$ and $\sum_{j \in k} q(i,j) = p(i,k)$. This equation shows that $H(p)$ is convex when varying the clustering because $q(i,j)$ is an invariant in this process, so $H(p)$ evolves like a negative logarithm which is a convex function. An important feature of the convexity of $H(p)$ is that a global minimum can be reached when minimizing its value, regardless of the initial conditions.

Search of the optimal partition. The number of possible partitions of the set of J objects in K clusters is the Stirling coefficient of order 2. The first variations of this coefficient when K increases from 1, are exponential. So the search of the optimal partition is a crucial point in the clustering process. Starting with an initial partition, a greedy algorithm is to systematically allot each object to each of the K clusters and to choose this one which insures the minimum $H(p)$ value and to iterate until convergence (Kneser(1993)). This algorithm can be very tedious and it is generally improved with an appropriate initialization.

We have found that a random search performs faster than this algorithm especially with all vectors gathered in the same cluster at the beginning of the clustering. In these conditions the initial entropy, $H(p)_{init}$ is :

$$H(p)_{init} = - \sum_i p(i) * \log[p(i)] = - \sum_i q(i) * \log[q(i)] \quad (8)$$

which is the entropy of the variables, regardless of the vectors. This value is the highest entropy value given by the data. As mentionned above, non-discriminant data would have an entropy equal to $\log I$.

Remark. In this paper, we improperly name a normalized value, a probability. We use this notation because it is easier to manipulate and as we only use the mathematical features of the entropy to establish our clustering criteria, this does not affect its properties.

The clusters obtained with this method are clearly representative of the analysed data and only of these data. Any a priori knowledge like a functional distribution is used to perform the clustering. This is a method which can be applied to any data array. A special case is when the collected data can be considered as an exhaustive statistical set. In this case, the concept of probability is fully correct, the probabilities are the maximum likelihoods of the underlying law which describes the data.

2 Experiments and validation

We initially used this entropy-based clustering to group words of a text according to their neighbours (Jardino(1997), Gauvain(1999)) then we enlarge it to the clustering of topics and documents according to the frequencies of the words of which they are made up. Our algorithm withstands large data sets, ten millions elements are currently used.

In order to illustrate some characteristics described above, we have applied this algorithm to smaller data sets, the first set, S1, is a matrix 20×20 which represents responses to perceptive tests made in our laboratory and the second set, S2, is a matrix 500×256 representing the frequencies of 256 characters appearing in 500 texts of the English Brown corpus.

Gradient descent. At the beginning of the iterative process, all objects are gathered in the same cluster. The other clusters are empty. These objects will expand in a fixed number of clusters, during an iterative process. At each step of this process, one object experiments a new cluster. We have compared two types of descent: a local one which systematically searches the best solution at each step and a Monte-Carlo descent which accepts any better solution.

Figure 1 shows these two processes. It reports only the entropy descent. The Monte-Carlo process is clearly faster. The equal number of trials which are needed to decrease the entropy at the beginning of the local gradient descent can be viewed as the effect of the systematic search. The small discrepancy between the two entropies at the end of the process is due to the discrete and finite values taken by the entropy variation when moving an object from a cluster to another one.

Clustering according to the number of classes. The second set has been used to observe the minimum entropy evolution varying the number of classes. This is particularly interesting to find those partitions which insure

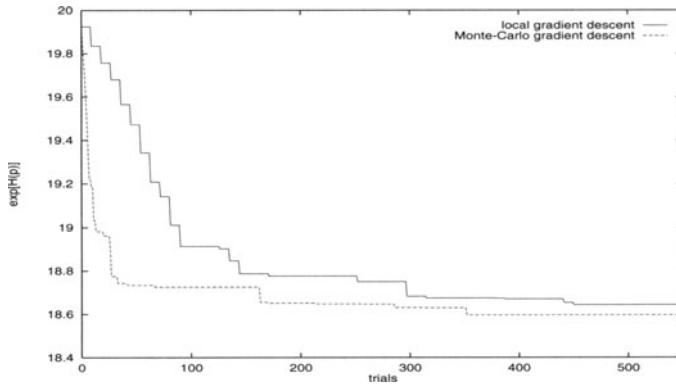


Fig. 1. Entropy descent during the clustering optimization. Comparison between two gradient descents.

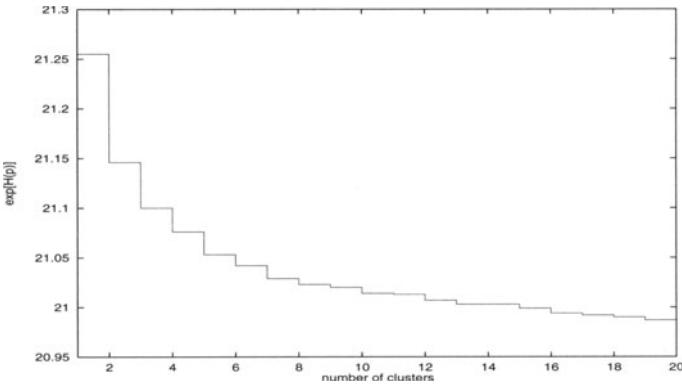


Fig. 2. Optimized entropy according to the number of clusters

the maximum gaps between them. For example, in Figure 2 the partitions 10 and 11 are almost identical, while the first partition in two sets leads to the greatest descent.

Relevance of the clusters Detailed class contents have always been described for the clustering of words according to their contexts (Jardino(1996), Jardino(1997)). Here we only present the splitting in two clusters, C1 and C2, of the data set S2 described above. These clusters are automatically built according to the number of characters of which they are made up. In this data set, texts are labelled according to their style, either "Informative prose" or "Imaginative prose". Table 1 gives the number of texts which have these tags and which are respectively in C1 and C2. As can be seen in column C2, there is a strong correlation between the character distribution and the informative prose. This feature is meaningful for the linguist who studies the text style. This is an example of how information can be extracted from data with this method.

	C1	C2	total
Informative prose	53	321	374
Imaginative prose	125	1	126
total	178	322	500

Table 1. Number of texts in labelled versus induced style classes

3 Conclusion

We have shown the interest to use an entropy-based criterion to automatically map unlabelled objects in a predefined number of clusters. We have demonstrated that this criterion is a convex function of the clustering, so that an optimal mapping can be reached, whatever the initial conditions. We have shown with several data sets, several properties of the algorithm. A lot of informations can be extracted from the clusters, they depend on the aim of the analysis: either to build robust language models, or to find correlations between data, or to separate informative prose from imaginative prose... The great advantage of this non-hierarchical clustering is that data are used in an unsupervised way, the only parameter to define is the number of classes. The choice of this number depends on the level of precision wished by the user.

References

- CELEUX G., DIDAY E. et al (1989): *Classification automatique des données*. Ed. Dunod.
- COVER T. and THOMAS J. (1991): *Elements of Information Theory*. Ed. Wiley & sons.
- DUDA R.O. and HART P.E. (1973): *Pattern Classification and Scene Analysis*. Ed. Wiley & sons.
- GAUVAIN J.-L., ADDA G. and JARDINO M. (1999): Language modeling for broadcast news transcription. In *Proceedings of the European Conference on Speech Technology*, EuroSpeech, Budapest, 1759–1762.
- JARDINO M. (1996): Multilingual stochastic n-gram class language models. In *Proceedings of the IEEE-ICASSP*, Atlanta.
- JARDINO M. and BEAUJARD C. (1997): Rôle du Contexte dans les Modèles de Langage n-classes, Application et Evaluation sur MASK et RAILTEL. In *Actes des Journées Scientifiques et Techniques*, 71–74.
- JELINEK F. (1998): *Statistical Methods for Speech Recognition*. Ed MIT Press.
- KNESER R. and NEY H. (1993): Improved Clustering Techniques for Class-Based Statistical Language Modelling. In *Proceedings of the European Conference on Speech Technology*, EuroSpeech, Berlin, 973–976.
- LERMAN I.C. and TALLUR B. (1980): Classification des éléments constitutifs d'une juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, n°28, 3, Paris.

Improving the Additive Tree Representation of a Dissimilarity Matrix Using Reticulations

Vladimir Makarenkov¹ and Pierre Legendre²

¹ Département de sciences biologiques, Université de Montréal,
C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada
and Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia

² Département de sciences biologiques, Université de Montréal,
C.P. 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada

Abstract. This paper addresses the problem of approximating a dissimilarity matrix by means of a reticulogram. A reticulogram represents an evolutionary structure in which the objects may be related in a non-unique way to a common ancestor. Dendograms and additive (phylogenetic) trees are particular cases of reticulograms. The reticulogram is obtained by adding edges (reticulations) to an additive tree, gradually improving the approximation of the dissimilarity matrix. We constructed a reticulogram representing the evolution of 12 primates. The reticulogram not only improved the data approximation provided by the phylogenetic tree, but also depicted the homoplasy contained in the data, which cannot be expressed by a tree topology. The algorithm for reconstructing reticulograms is part of the T-Rex software package, available at URL <<http://www.fas.umontreal.ca/BIOL/legendre>>.

1 Introduction

Several algorithms have been proposed for the representation of empirical dissimilarity data using a general network where the objects are represented by the nodes of a valued graph whose minimum path-length distances are associated with the dissimilarities (Feger and Bien 1982; Orth 1989; Klauer and Carroll 1989). An expanding tree structure based on weak clusters has also been proposed by Bandelt and Dress (1989) leading to a weak hierarchy for an empirical similarity matrix. Bandelt and Dress (1992) and Bandelt (1995) resumed investigation of weak clusters and proposed the method of split decomposition.

We outline the main features of a reticulogram reconstruction algorithm offering another way of modelling a dissimilarity matrix by means of a network. Our representation uses a topology called a *reticulogram* which includes the vertices associated with the objects in a set X as well as the intermediate nodes. A reticulogram can represent relationships among objects that may be related in a non-unique way to a common ancestor: such a structure cannot be represented by a tree. In a reticulogram, the distance between i and j is the *minimum-path-length distance* over the set of all paths linking i and j .

Infering an additive tree from a dissimilarity matrix is a very well-studied issue in the literature. We launch the reticulogram reconstruction algorithm

from an additive tree topology providing an initial fit for the dissimilarity matrix. The algorithm adds new edges or *reticulations* to a growing reticulogram, minimising the least-squares loss function computed as the sum of the quadratic differences between the original dissimilarities and the associated reticulogram estimates.

Reticulate patterns are found in nature in some phylogenetic problems. (1) In bacterial evolution, lateral gene transfer (LGT) produces reticulate evolution; LGT represents the mechanisms by which bacteria can exchange genes across “species” through a variety of mechanisms (Sonea & Panisset 1976, Margulis 1981). (2) Reticulate evolution also occurs in plants where allopolyploidy may lead to the instantaneous appearance of a new species possessing the chromosome complement of its two parent species. (3) It is also found in within-species micro-evolution in sexually reproducing eukaryotes. Reticulate patterns may also occur in non-phylogenetic problems such as host-parasite relationships involving host transfer and in the field of ecological biogeography.

2 Algorithm for constructing reticulograms

This section describes the most important features of our reticulogram reconstruction algorithm. A *reticulogram* or *tree network* R is a triplet (E, V, l) where V is a set of vertices, E is a set of edges and l is a *function* of edge lengths assigning real non-negative numbers to the edges. Each vertex i is either an object belonging to a set X or a node belonging to $V - X$. In this study we considered only connected and undirected reticulograms. The algorithm uses as input a dissimilarity matrix \mathbf{D} on the set of n objects and an additive tree T inferred from \mathbf{D} using one of the classical reconstruction algorithms. At each step, the algorithm adds to the additive tree T a new edge (reticulation) of optimal length ensuring the minimisation of the following least-squares loss function:

$$Q = \sum_{i \in X} \sum_{j \in X} (\delta(i, j) - d(i, j))^2 \rightarrow \min \quad (1)$$

where $d(i, j)$ is a dissimilarity value between objects i and j , and $\delta(i, j)$ is the corresponding value of reticulogram distance defined as a *minimum-path-length distance* between vertices i and j in R .

Makarenkov & Legendre (1999) introduced a statistical criterion Q_1 which measures the gain in fit when a new reticulation is added. The minimum of this criterion provides a stopping rule for addition of reticulations. This function takes into account the least-squares loss function as well as the number of degrees of freedom of the reticulogram under construction:

$$Q_1 = \frac{1}{(n(n-1)/2 - N)} \sqrt{\sum_{i \in X} \sum_{j \in X} (\delta(i, j) - d(i, j))^2} = \frac{\sqrt{Q}}{(n(n-1)/2 - N)} \quad (2)$$

N is the number of edges in the reticulogram. N is equal to $2n - 3$ in a binary additive tree with n leaves corresponding to the objects in X and $n - 2$ internal nodes. Thus, in this study, the reticulogram will always contain $2n - 2$ internal nodes, n of which correspond to the observed objects.

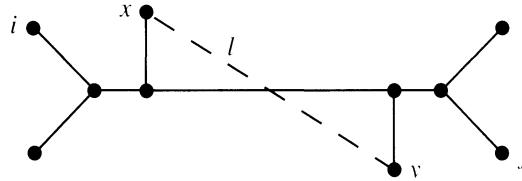


Fig. 1. A new edge of length l can be added to tree T between vertices x and y .

Consider now a binary additive tree T inferred from a dissimilarity d by means of an appropriate fitting method and a pair of vertices x and y in T not linked by an edge (Fig. 1). Using the least-squares loss function, we have to determine an optimal value l for a new edge xy that may be added to the tree T . Let us consider the set $A(xy)$ of all pairs of objects ij of X such that:

$$\text{Min } \{\delta(ix) + \delta(jy); \delta(jx) + \delta(iy)\} < \delta(ij) \quad (3)$$

The set $A(xy)$ represents the distances between pairs of objects that are susceptible of changing if a new reticulation xy is added. Actually, the set $A(xy)$ can be subdivided into the m subsets A_1, A_2, \dots, A_m such that $A(xy) = \{A_1 \cup A_2 \cup \dots \cup A_m\}$. They are defined in the following way:

$A_1 = \{ij\}$ such that:

$$\begin{aligned} \delta(i,j) - \text{Min}\{\delta(i,x) + \delta(j,y); \delta(j,x) + \delta(i,y)\} = \\ \text{Min}_{\{ij \in A(xy)\}} \{\delta(i,j) - \text{Min}\{\delta(i,x) + \delta(j,y); \delta(j,x) + \delta(i,y)\}\} \\ = l_1 \end{aligned}$$

...

$A_k = \{ij\}$ such that:

$$\delta(i,j) - \text{Min}\{\delta(i,x) + \delta(j,y); \delta(j,x) + \delta(i,y)\} = l_k > l_{k-1}$$

...

$A_m = \{ij\}$ such that:

$$\begin{aligned} \delta(i,j) - \text{Min}\{\delta(i,x) + \delta(j,y); \delta(j,x) + \delta(i,y)\} = \\ \text{Max}_{\{ij \in A(xy)\}} \{\delta(i,j) - \text{Min}\{\delta(i,x) + \delta(j,y); \delta(j,x) + \delta(i,y)\}\} \\ = l_m = \delta(x,y) > l_{m-1} \end{aligned}$$

This subdivision is performed because each different subset A_i can be associated with an interval of possible edge lengths l for which a particular optimisation problem may be formulated. Let us compose a special quadratic function to be minimised for a fixed interval of edge length values. To obtain its optimal solution, suppose that $l_k \leq l \leq l_{k+1}$, where $k = 0 \dots m - 1$. This constraint means that if a new edge xy of length l is added to T , only the

set of distances $\delta(ij)$ such that $ij \in \{A_m \cup A_{m-1} \cup \dots \cup A_{k+1}\}$ will change lengths. Thus, the function to minimise to compute the optimal length value of l is as follows:

$$Q^*(xy, k) = \sum_{p=k+1}^m \sum_{ij \in A_p} (\text{Min}\{\delta(i, x) + \delta(j, y); \delta(j, x) + \delta(i, y)\} + l - d(i, j))^2 \rightarrow \min \quad (4)$$

subject to the constraint $l_k \leq l \leq l_{k+1}$. $Q^*(xy, k)$ comprises the quadratic sum of differences between the dissimilarities d and the associated reticulogram distances δ , considering only the distances that may change in the reticulogram. The non-trivial solution $l^*(xy, k)$ is (Makarenkov and Legendre, submitted):

$$\sum_{p=k+1}^m \sum_{ij \in A_p} (d(i, j) - \text{Min}\{\delta(i, x) + \delta(j, y); \delta(j, x) + \delta(i, y)\}) / \sum_{p=k+1}^m |A_p| \quad (5)$$

This calculation is repeated over all intervals of edge lengths $l_k \leq l \leq l_{k+1}$, for $k = 0 \dots m - 1$, for the given pair of vertices xy . The global optimum for criterion Q found for every particular solution, as well as the global optimum of the edge length l over the set of defined intervals, are recursively obtained. To obtain the optimum value for Q over the set of all possible new edges, the computations are repeated for all pairs of tree (reticulogram) vertices not linked by an edge.

3 Application

In a recent study, Makarenkov & Legendre (1999) considered two applications of reticulogram reconstruction. The first one concerned the postglacial dispersal of freshwater fishes in the Québec Peninsula. The second example depicted the morphological differentiation of muskrats in a river valley in Belgium. We will now examine how the method can be applied to represent homoplasy in the phylogenetic tree of primates. Homoplasy is the portion of phylogenetic similarity resulting from convergence. The data, from Hayasaka et al. (1988), consisted of a portion of the protein-coding mitochondrial DNA (898 bases) over 12 species of primates. The dissimilarity matrix (Table 1) was obtained by computing the Hamming distance among the species. First, a phylogenetic tree was inferred from the dissimilarity matrix using the neighbor-joining method (Saitou & Nei 1987). The tree is represented by full lines in Fig. 2.

The phylogeny separated four basic groups of primates. The values of criteria Q and Q_1 after approximation of the edge lengths (about this technique, see Makarenkov & Leclerc 1999) were 0.002479 and 0.001106, respectively. Five new edges (reticulations, dashed lines in Fig. 2) were added to the tree by the algorithm. The minimum of Q_1 was reached at the fifth step of the algorithm, which allowed to decrease Q_1 to 0.001041, whereas Q dropped to 0.001733 (gaining about 30%).

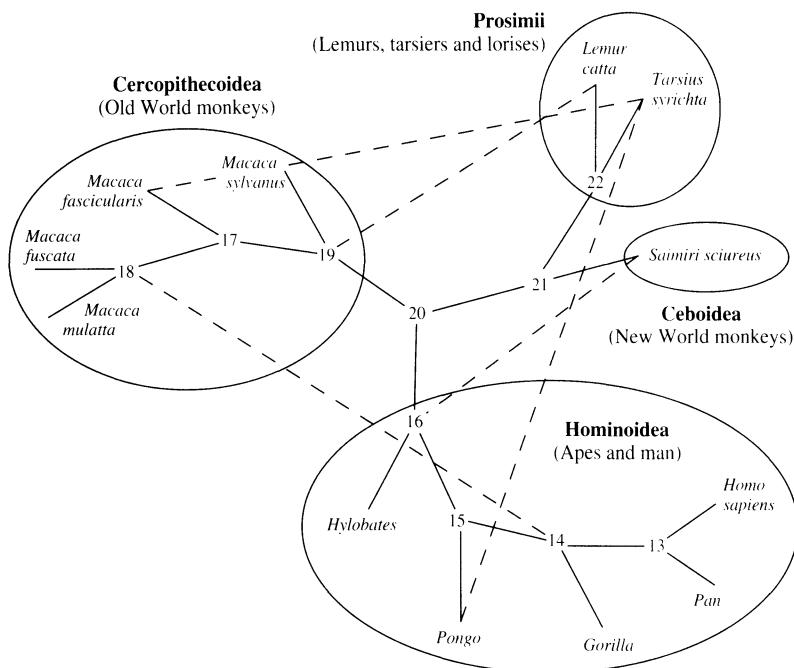


Fig. 2. Reticulogram representing the phylogeny of the primates from Table 1. Full lines: edges of the additive tree. Dashed: reticulations added by algorithm.

How can we interpret reticulations? From the mathematical point of view, each reticulation improves the representation of matrix \mathbf{D} by the classical additive tree, allowing an optimal gain in fit. From the biological point of view, the lengths of the reticulations are of great importance. If the length of a reticulation is small with respect to the other edge lengths, it may represent a mutation event that occurred during evolution. In the example, the reticulations are long and they occur between distant groups, so that they represent homoplasy (i.e., information representing convergent evolution: parallel evolution and reversals) in the data, which the phylogenetic tree was unable to correctly represent. For instance, the distance between *Homo sapiens* and *Macaca fuscata* is 0.23215 in Table 1. The distance between these species is 0.24133 along the tree, whereas the minimum-path-length reticulogram distance, which includes the reticulation linking the Cercopithecoidea and Hominoidea, is 0.23549. This value is a better representation of the original dissimilarity than the tree path-length distance is.

References

- BANDELT, H.-J. (1995): Combination of Data in Phylogenetic Analysis. *Plant Systematics and Evolution Supplementum*, 9, 355–361.

	1	2	3	4	5	6	7	8	9	10	11
2 Pan	0.089										
3 Gorilla	0.104	0.106									
4 Pongo	0.161	0.171	0.166								
5 Hylobates	0.182	0.189	0.189	0.188							
6 Macaca fus.	0.232	0.243	0.237	0.244	0.247						
7 M. mulatta	0.233	0.251	0.235	0.247	0.239	0.036					
8 M. fascicul.	0.249	0.268	0.262	0.262	0.257	0.084	0.093				
9 M. sylvan.	0.256	0.249	0.244	0.241	0.242	0.124	0.120	0.123			
10 Saimiri sc.	0.273	0.284	0.271	0.284	0.269	0.289	0.293	0.287	0.287		
11 Tarsius sy.	0.322	0.321	0.314	0.303	0.309	0.314	0.316	0.311	0.319	0.320	
12 Lemur ca.	0.308	0.309	0.293	0.293	0.296	0.282	0.289	0.298	0.287	0.285	0.252

Table 1. Dissimilarity matrix among primates; species 1 is *Homo sapiens*.

- BANDELT, H.-J. and DRESS A.W.M. (1989): Weak Hierarchies Associated with Similarity Measures - An Additive Clustering Technique. *Bulletin of Mathematical Biology*, 51, 133–166.
- BANDELT, H.-J. and DRESS A.W.M. (1992): Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data. *Molecular Phylogenetics and Evolution*, 1, 242–252.
- HAYASAKA, K., GOJOBORI, T., and HORAI, S. (1988): Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, 5, 626–644.
- KLAUER, K.C. and CARROLL, J.D. (1989): A Mathematical Programming Approach for Fitting General Graphs. *Journal of Classification*, 6, 247–270.
- MAKARENKO, V. and LECLERC, B. (1999): An Algorithm for the Fitting of a Tree Metric According to a Weighted Least-Squares Criterion. *Journal of Classification*, 16, 3–27.
- MAKARENKO, V. and LEGENDRE, P. (2000): General Network Representation of a Dissimilarity Matrix: Adding Reticulations to an Additive Tree. *Journal of Classification* (submitted).
- MARGULIS, L. (1981): *Symbiosis in Cell Evolution*, San Francisco, CA: W. H. Freeman.
- ORTH, B. (1988): Representing Similarities by Distance Graphs: Monotonic Network Analysis (MONA). In: H. H. Bock (Ed.), *Classification and related methods of data analysis*, Amsterdam: North-Holland, 489–494.
- SAITOU, N. and NEI, M. (1987): The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, 4, 406–425.
- SONEA, S. and PANISSET, M. (1976): Pour une nouvelle bactériologie. *Revue Canadienne de Biologie*, 35, 103–167.

Double Versus Optimal Grade Clusterings

Alicja Ciok

Institute of Computer Science PAS, 01-237 Warsaw, Ordona 21, Poland

Abstract. Two clustering methods based on grade correspondence analysis will be compared on a real data example. Special attention will be paid to the interpretation aspects versus the formal inference based on clustering quality measures. The discussed example shows that formally similar solutions may differ significantly from the interpretation point of view.

1 Double grade clustering and optimal grade clustering

Grade correspondence analysis and univariate and bivariate clustering methods based on it were presented during the IFCS'98 conference (Ciok (1998)). Here we recall only a few general ideas underlying these methods.

- Input data must have the *form of a bivariate contingency table*. Due to that, data structures may be expressed in terms of stochastic dependence (even when the data table contains values of variables instead of frequencies). Let X denote the row variable and Y the column variable and let $\rho^*(X, Y)$ denote the Spearman's ρ , known also as the grade correlation coefficient (equal in the continuous case to $\text{cor}(F_X(X), F_Y(Y))$).
- Grade correspondence analysis (GCA) maximizes $\rho^*(X, Y)$ in the set of all pairs of permutations of categories of X (rows) and Y (columns). It arranges similar rows (columns) close to each other and consequently defines new variable scales (cf. Sec. 4 of Ciok et al. (1998)).
- Double grade clustering (DGC) is based directly on optimal permutations provided by GCA. Given the numbers of clusters after aggregation, we maximize separately $\rho^*(\text{aggr}X, Y)$ and $\rho^*(X, \text{aggr}Y)$, where $\text{aggr}X$ denotes aggregated X : in each case we aggregate *adjacent* categories according to the GCA permutations.
- Optimal grade clustering (OGC) maximizes $\rho^*(\text{aggr}X, \text{aggr}Y)$. In this case, the clusters may not consist of categories adjacent according to the GCA arrangements. These arrangements serve just as good *starting points* for the optimization procedure.

We note that DGC and OGC are *not* hierarchical, contrary to the clustering methods based on correspondence analysis and introduced by Greenacre (1988).

During the last year we analysed many data tables (real as well as the artificial) and a lot of attention was paid to comparing results of OGC and DGC. One pattern repeats persistently. Aggregated tables after OGC and

after DGC have very similar values of coefficient ρ^* (the difference never exceeds 0.01); in other words, both results are almost equally good with respect to the optimization criterion. As OGC computations may be much more costly than DGC (although the procedure converges very quickly), the question arises: is it worth this cost? Obviously, there exist data tables for which both procedures provide the same results but what about the others?

2 Grade correspondence analysis of the time allocation data

Let consider an exemplary data table published by Mooijaart et al. (1999) which characterises time allocation in the Netherlands cross-classified by gender, age and year. Every cell of the table contains the average number of minutes in a day spent in 1975, 1980 or 1985 by people of a certain gender and age on some activity. There are 30 compositions of gender, age and year corresponding to rows of the table, and 18 activities, corresponding to columns. Tab. 1, Tab. 2 and Fig. 1 show the results of GCA applied to this data table, previously suitably normalized. Let $T = (t_{ij})$ denotes this normalized and optimally permuted GCA table. These permutations of rows and columns are given in Tab. 1 and Tab. 2. The respective maximal value of ρ^* is equal to 0.2582.

Seq. no	Gender	Age	Year	Cluster	Seq. no	Gender	Age	Year	Cluster
1	M	12-24	75	1	16	M	>64	85	3
2	M	12-24	85	1	17	M	>64	80	3
3	M	12-24	80	1	18	M	>64	75	3
4	F	12-24	85	1	19	F	25-34	85	3
5	M	25-34	75	1	20	F	>64	85	4
6	M	35-49	85	1	21	F	25-34	80	4
7	M	25-34	85	1	22	F	50-64	75	4
8	M	25-34	80	2	23	F	25-34	85	4
9	F	12-24	80	2	24	F	>64	80	4
10	M	35-49	75	2	25	F	35-49	75	4
11	M	35-49	80	2	26	F	35-49	85	4
12	F	12-24	75	2	27	F	35-49	80	4
13	M	50-64	75	2	28	F	>64	75	4
14	M	50-64	80	2	29	F	50-64	85	4
15	M	50-64	85	2	30	F	50-64	80	4

Table 1. The optimal GCA permutations and the optimal clusters for rows.

The visualization of table T is given in Fig. 1. It is the so called overrepresentation map, where (i, j) -th rectangle corresponds to i -th category of row variable X and j -th category of Y . Widths of rows (columns) reflect

Number in permutation	Activity no number	Activity	Cluster
1	8	education	1
2	1	paid work	1
3	11	going out	1
4	14	recreation outside	2
5	15	tv, radio, audio	2
6	7	sleeping, resting	2
7	18	other	2
8	17	relaxing	2
9	6	eating, drinking	3
10	5	personal needs	3
11	13	gardening, pets	3
12	12	sports, hobbies	3
13	9	volunteer work	3
14	16	reading	3
15	10	social contacts	3
16	4	shopping	4
17	3	caring for members household	4
18	2	domestic work	4

Table 2. The optimal GCA permutations and the optimal clusters for columns.

respective marginal sums of T . Values of $h_{i,j} = \frac{t_{i,j}}{t_{i\bullet} t_{\bullet j}}$, where $t_{i\bullet} = \sum_j t_{i,j}$, $t_{\bullet j} = \sum_i t_{i,j}$, measure over-representation with respect to *independence* of variable X and Y . These values, discretized into 5 categories, are represented by various shades of grey. Black corresponds to the highest values (the highest over-representation), white - to the lowest (the highest under-representation).

A visible pattern emerges in Fig. 1. Two opposite ends define the new column scale: education and paid work on the left, domestic work on the right. These activities best differentiate the analysed population. For rows, the opposite ends are formed by: men and very young women, which belong to the upper part of scale; women (with exception of the very young), which belong to the lower part. Of course, such description of the revealed pattern is too rough: clustering methods are needed for identifying types of behaviour.

3 DGC applied to the time allocation data

In our example, four was chosen arbitrarily as the number of clusters for rows as well as for columns. The results provided by DGC are given in the last columns of Tab. 1 and Tab. 2. Fig. 2 shows the over-representation map for the table aggregated according to these clusterings.

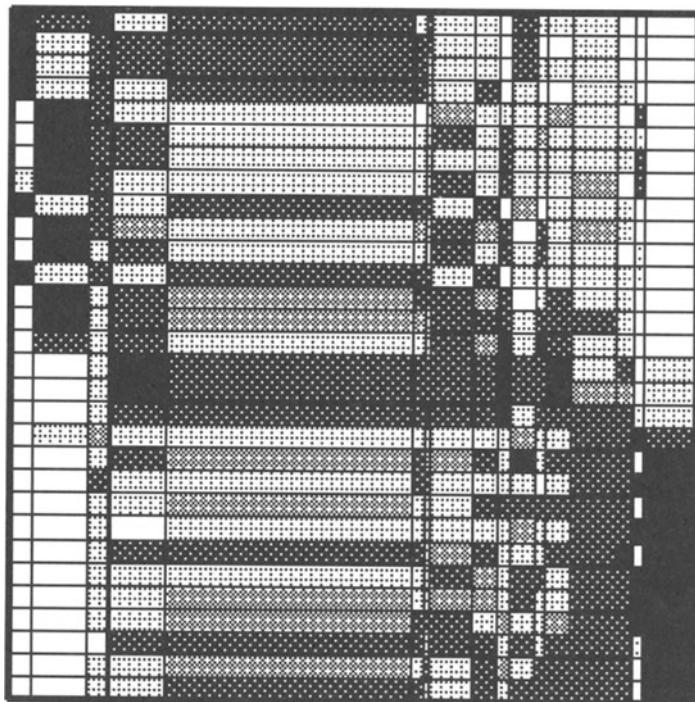


Fig. 1. Over-representation map after GCA.

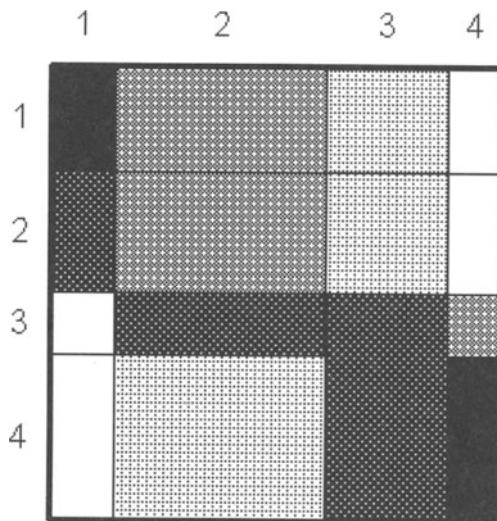


Fig. 2. Over-representation map after DGC.

4 OGC versus DGC

Tab. 3, Tab. 4 and Fig. 3 present results obtained by OGC. Value of ρ^* for the table aggregated according to DGC is equal to 0.2369, the analogical value corresponding to OGC is equal to 0.2374, so the difference between them is almost negligible. We also compare the values of ρ^* *before aggregation* for two orderings of rows and columns: once according to GCA ($\rho^* = 0.2582$) and once according to OGC ($\rho^* = 0.2574$; the clusters consist of adjacent rows (columns) in these permutations). The difference is equal to 0.0008.

Cluster	Gender	Age	Year
1	M	12-49	75, 80, 85
2	F	12-24	75, 80, 85
	M	50-64	75, 80, 85
3	M	>64	75, 80, 85
	F	25-34	85
4	F	>24	75, 80, 85

Table 3. The optimal OGC clusters for rows.

Cluster	Activity numbers	Activities
1	8, 1, 11	education; paid work; going out
2	14, 15, 13, 7, 6	recreation outside; tv, radio, audio; gardening, pets; sleeping, resting; eating, drinking
3	9, 18, 17, 5, 16, 12, 10	volunteer work; other: relaxing; personal needs; reading; sports, hobbies; social contacts
4	4, 3, 2	shopping; caring for members household; domestic work

Table 4. The optimal OGC clusters for columns.

Hence, one may say that formally the results of OGC and DGC are almost equally good, but it is not so. The difference between obtained clusters (for rows or columns) is decidedly *not* negligible although the general trend implied by GCA is present in both cases. OGC results have some significant interpretative advantages. One of them is that the generated clusters are more differentiated. It is particularly well visible in the case of row clusters: No 1 and 2 (compare Fig. 2 and Fig. 3). Moreover, row OGC clusters are more homogeneous with respect to gender and age, which implies a better characterization of the analysed subpopulations. In particular, the group of very young women is much more scattered among the clusters provided by DGC than among those provided by OGC. In the former case, such a dispersion would be interpreted as the proof that the young ladies group has no

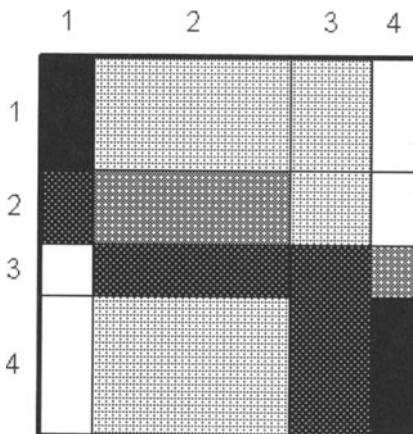


Fig. 3. Over-representation map after OGC.

specific characteristics. In the latter, it is part of a single cluster (jointly with the men of age 50-64). Moreover, the group occupies an intermediate position between two groups: the younger men, and the men older than 64 years old. This implies that this group is specific, although it combines characteristics from both ends of the GCA activity scale.

It should be mentioned that the results of OGC and DGC become identical for sufficiently regular data tables (but considerations on regularity are beyond the scope of this paper). GCA transforms a data table into one which can be treated as the best possible approximation of a regularly monotonely dependent table, but in practice this approximation is rarely regular enough. Therefore it seems useful to consider and interpret OGC as well as DGC solutions. The more they differ, the better recognition of latent patterns is possible, and the better sources of irregularity can be identified.

References

- CIOK, A. (1998): Discretization as a tool in cluster analysis. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer, Heidelberg, 349–354.
- CIOK, A., KOWALCZYK, T., PLESZCZYNSKA, E., SZCZESNY, W. (1998): How a new statistical infrastructure induced a new computing trend in data analysis. In: L. Polkowski, A. Skowron (Eds.): *Rough sets and current trends in computing*. Lecture Notes in Artificial Intelligence 1424, Springer, 75–82.
- GREENACRE, M.J. (1988): Clustering the rows and columns of a contingency table. *Journal of Classification*, 5, 39–51.
- MOOIJJAART, A., VAN DER HEIJDEN, P.G.M., and VAN DER ARK, L.A. (1999): A least squares algorithm for a mixture model for compositional data. *Computational Statistics and Data Analysis*, 30, 359–379.

The Effects of Initial Values and the Covariance Structure on the Recovery of some Clustering Methods

Istvan Hajnal and Geert Loosveldt

Department of Sociology, University of Leuven,
Edward Van Evenstraat 2B, 3000 Leuven, Belgium
(e-mail: istvan.hajnal@soc.kuleuven.ac.be)

Abstract. Some clustering methods are compared in a simulation study. The data used in the analysis are generated in a mixture modeling framework. The methods included are some hierarchical methods, k -means as implemented in the FASTCLUS procedure of SAS and cluster analysis by means of normal mixtures with the NORMIX program. We demonstrate that the poor recovery found in some studies for normal mixture type of clustering is partly due to the use of bad initial values, and partly due to the specification of covariance structure within the cluster. We further find that an important factor in the relative success of FASTCLUS lies in the initial seed selection.

1 Introduction

Recently there has been a growing interest in mixture models. In this paper we focus on multivariate normal mixtures as a method for cluster analysis (see McLachlan and Basford (1988) for an overview on this subject). The number of validation studies that have included normal mixtures is surprisingly low (see Milligan (1996) for an overview of validation studies in the field of cluster analysis). We took a closer look at the validation studies that included the NORMIX (Wolfe, 1970) program. Two studies were not based on simulated data, but rather evaluated the methods by applying them to an existing data set and comparing the results with those of other methods. One of the studies was in favor of NORMIX (Wolfe, 1978) and one was negative for this method (Mezzich, 1978). We could only find 3 studies in the literature that used NORMIX in a simulation study with artificial data. The first one was a rather small scale study by Everitt (1974) that was positive for this method. The second study, by Bayne et al. (1980), was rather negative for NORMIX. The results of this study were challenged by Price (1993), by insisting that recovery measure should take account of NORMIX's fuzzy classification. One of the possible explanation for NORMIX's weak performance in the study of Bayne et al. (1980) could be the use of weak starting values (Price, 1993). Another potential explanation might be that the other methods are particularly successful in situation without correlation within the clusters, but fail when correlation is introduced (See Price (1993), Donoghue

(1995) found a positive effect for positive correlations and a negative effect for negative correlations). In NORMIX, however, the elements of the covariance matrices are estimated and thus the method should be more flexible in handling situations with correlation. Since NORMIX uses maximum likelihood to estimate its parameters it is often stressed in the literature that the sample size should be high enough to estimate that many parameters. All this leads to the hypothesis that the recovery of NORMIX will be higher when appropriate initial values are used, that introducing correlation in the clusters will not negatively affect the recovery of NORMIX and finally, that the higher the sample size the higher the recovery will be.

Additionally, we will take this opportunity to evaluate the seed selection mechanism in FASTCLUS, since the general conclusion of validation studies that included the k -means method seems to be that k -means methods perform quite well, provided that ‘good’ initial seeds are used (see Milligan (1980), Milligan (1981) amongst others). Basically, in FASTCLUS the initial seeds are selected by a very simple algorithm. The algorithm always selects the first observation without missing values as the first seed. The next seeds are simply the next observations without missing values that are separated from all previous seeds by at least the ‘radius’, which is a user defined (minimum) distance (SAS,1989).

2 Design

The data were generated with a multivariate normal mixture approach (with equal cluster sizes, the specification of the mean vectors and the covariance matrices will be discussed below). Several factors are known to affect cluster recovery. The factors that we varied in this simulation study were based on the hypothesis mentioned in the first paragraph: T , the total number of entities ($N = 99, 201, 300$); D , the distance between the cluster centroids’ ($d = 1, 2, 5, 10$); and C , the covariance structure ($\rho = .0, .4, .8$). Other factors were kept constant, such as the number of clusters (3) and the number of variables (4), the variances of the variables (1). The cluster centroids were placed in an equidistant triangle. The mean vectors then become:

$$\mu_1 = \begin{bmatrix} -\frac{d}{2\sqrt{3}} \\ -\frac{d}{2} \\ 0 \\ 0 \end{bmatrix}, \mu_2 = \begin{bmatrix} -\frac{d}{2\sqrt{3}} \\ \frac{d}{2} \\ 0 \\ 0 \end{bmatrix}, \mu_3 = \begin{bmatrix} \frac{d}{\sqrt{3}} \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad (1)$$

Price (1993) argued that the results of Everitt (1974) and Bayne et al. (1980), that were until then viewed as conflicting (Milligan, 1981), could be seen as complementary in the sense that in Everitt’s (1974) study the distance between the clusters was high, whereas in the study of Bayne et al. (1980) the clusters were much closer to each other. That’s why we selected 2 of the 4 values of d such that 2 extreme cases would occur. When $d = 1$, the

cluster overlap would make the clustering task very difficult, whereas $d = 10$, the distance between the clusters is so high that we would expect that any clustering method would find the true clustering solution. The middle two distances $d = 2$ and $d = 5$ would correspond to situations where clustering is realistic (not too easy and not too difficult). The covariance matrices within the clusters are equal and were specified as:

$$\boldsymbol{\Sigma}_g = \begin{bmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix}, \text{ for } g = 1, \dots, 3 \quad (2)$$

with $\rho = 0$, for the local independence case, $\rho = .4$, in the case of medium correlation, $\rho = .8$, in the case of high correlation. The combination of the 3 factors resulted in $3 \times 4 \times 3 = 36$ setups. Within each setup we used 5 replications, which resulted in a total of $36 \times 5 = 180$ data sets.

Amongst the 12 clustering methods that were studied, 6 are some of the hierarchical clustering methods included in the CLUSTER procedure of SAS (Ward's method (s_1), single linkage (s_2), complete linkage (s_3), centroid method (s_4), median method (s_5) and average linkage (s_6)). The other methods are some non hierarchical clustering methods. s_{10} is the NORMIX program with the default options, which means that the initial values are computed by the program itself. In the design of this simulation study the covariances of the clusters are equal ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3$). Since this situation coincides with the equal covariances option in NORMIX (sometimes called NORMAP, (s_{11})), this method was also included. In NORMAP all clusters are supposed to have the same covariance matrix, so the number of parameters to be estimated will be lower than in NORMIX. We also implemented the normal mixture model in a SAS macro which uses starting values derived from FASTCLUS (s_{12}). s_9 is the k -means method with the FASTCLUS seed selection, whereas s_8 is k -means with random seeds. Finally s_7 is nothing but the assignment of the entities to the closest FASTCLUS seed.

The 180 data sets were analyzed by the 12 clustering methods. The total number of cluster analyses performed for this study was 2160. The extent in which the cluster solution corresponds to the generated data (the true clustering) is often called the cluster *recovery*. The cluster recovery was measured by using the Hubert and Arabie (1985) adjusted Rand statistic (r_{HAR}).

3 Results

The results were analyzed in a full factorial analysis of variance as in Milligan (1980), with the recovery measure r_{HAR} as the dependent variable and the design factors as the independent variables. To assess the importance of different factors we used $\eta^2 = SS_{\text{effect}}/SS_{\text{tot}}$ as a measure of effect size (see also Donoghue (1995)). From Table 1 it is clear that by giving the marginal

table of the effect $D \times C \times S$ (including the effects within this marginal table) we capture the most important results for this simulation (see Table 2). As

Table 1. Simulation results. Df, SS, F, prob, η^2 and $\sum \eta^2$ for each effect on r_{HAR}

Effect	Df	SS	F	Prob	η^2	$\sum \eta^2$
D	3	365.892832	17123.365981	0.000000	0.843792	0.843792
$D \times S$	33	16.617068	70.696344	0.000000	0.038321	0.882113
$D \times C \times S$	66	14.692080	31.253297	0.000000	0.033882	0.915995
S	11	12.784314	163.170352	0.000000	0.029482	0.945477
$C \times S$	22	8.004832	51.084137	0.000000	0.018460	0.963937
$D \times C$	6	1.004304	23.500156	0.000000	0.002316	0.966253
$T \times D \times C \times S$	132	0.835285	0.888417	0.808909	0.001926	0.968180
$T \times D \times S$	66	0.610672	1.299034	0.055923	0.001408	0.969588
$T \times C \times S$	44	0.256196	0.817477	0.798072	0.000591	0.970179
$T \times S$	22	0.202948	1.295147	0.162052	0.000468	0.970647
$T \times D$	6	0.161965	3.789887	0.000935	0.000374	0.971020
$T \times D \times C$	12	0.149827	1.752932	0.050842	0.000346	0.971366
$T \times C$	4	0.056042	1.967015	0.097047	0.000129	0.971495
T	2	0.036099	2.534118	0.079626	0.000083	0.971578
C	2	0.016502	1.158441	0.314219	0.000038	0.971616

expected, all methods increased their recovery values with increasing values of d , generally going from $r_{har} \approx 0$ when $d = 1$ to $r_{har} \approx 1$ when $d = 10$. Notice that when $d = 10$ k -means with random seeds (s_8) sometimes failed to find the true clustering. As is often observed in this type of simulations, the single linkage (s_2) method performs the worst and Ward's method (s_1) performs the best amongst the hierarchical methods. The results for the non-hierarchical methods are clearly superior. Even when only the starting values of FASTCLUS (s_7) are used (with 1 k -means iteration) this method is generally better than the hierarchical methods. In the no correlation case k -means (s_9) is the best method. When correlation is added almost all methods, except the normal mixture methods (s_{10}, s_{11} en s_{12}) and single linkage (s_2) decrease their recovery values. In the no correlation case, when the FASTCLUS results are used as the starting values for the normal mixture method (s_{12}), the recovery is higher than NORMIX (s_{10}) and even NORMAP (s_{11}), but lower than FASTCLUS (s_9) itself. Since the covariances were equal in this study, it comes as no surprise that NORMAP (s_{11}) outperforms NORMIX (s_{10}). Overall the highest recovery values were found for the NORMAP (s_{11}) in the presence of high correlation. The effect of the increasing sample size on the recovery of the normal mixture methods (s_{10}, s_{11} en s_{12}) was not as strong as one might have expected (not shown in table).

4 Discussion

It is clear that the design of this study is too limited to make final conclusions. First, the choice of the most appropriate algorithm depends not only on recovery, but criteria such as robustness, CPU time and memory storage

Table 2: Simulation results. Mean r_{HAR} for effect $D \times C \times S$

S	\mathcal{S}											
	C^D			D			C^D			S		
Ward's method s_1	$c = .0$ $\rho = .4$ $\rho = .8$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .021702$ $\rho = .016219$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .349346$ $\rho = .342882$ $\rho = .531006$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .041033$ $\rho = .028767$ $\rho = .005237$	$d=2$ $d=5$	$d=10$
single linkage s_2	$c = .0$ $\rho = .4$ $\rho = .8$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .000043$ $\rho = .000039$ $\rho = .000030$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .036586$ $\rho = .0854207$ $\rho = .0905906$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .065510$ $\rho = .041058$ $\rho = .08 = .8$	$d=2$ $d=5$	$d=10$
complete linkage s_3	$c = .0$ $\rho = .4$ $\rho = .8$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .039755$ $\rho = .019052$ $\rho = .008937$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .185491$ $\rho = .127847$ $\rho = .048323$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .055796$ $\rho = .019605$ $\rho = .019497$	$d=2$ $d=5$	$d=10$
centroid method s_4	$c = .0$ $\rho = .4$ $\rho = .8$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .000207$ $\rho = .001020$ $\rho = .007382$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .000229$ $\rho = .004871$ $\rho = .019070$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .482855$ $\rho = .149169$ $\rho = .0440366$	$d=2$ $d=5$	$d=10$
median method s_5	$c = .0$ $\rho = .4$ $\rho = .8$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .003070$ $\rho = .011410$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .008057$ $\rho = .070253$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .075659$ $\rho = .010973$	$d=2$ $d=5$	$d=10$
average linkage s_6	$c = .0$ $\rho = .4$ $\rho = .8$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .001072$ $\rho = .010752$ $\rho = .008825$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .031619$ $\rho = .063639$ $\rho = .057985$	$d=1$ $d=2$ $d=5$	$d=10$	$c = .0$ $\rho = .0396708$ $\rho = .0960382$ $\rho = .0834908$	$d=2$ $d=5$	$d=10$

The mean r_{HAR} values in the body of this table are based on 15 cluster analyses.

should also be considered in practical situations. Second, this study does not include some of the newer methods. Third, the data are generated in such a way that it should lead to good results for the mixture methods. Nevertheless, it shows that the negative results in some previous studies, and particularly in Bayne et al. (1980) should be reexamined. The good results for normal mixtures in the high correlation case might be due to an artefact of the design. A more profound study if the effect of the correlational structure, such as in Donoghue (1995) is needed. In future validation studies, normal mixture methods should be included, with proper initial values and with the appropriate options selected, only then a fair comparison is possible.

Another result from this study is that the initial seed selection used in FASTCLUS seems to be very successful, especially in the no correlation situation.

References

- BAYNE, C.K., BEAUCHAMP, J.J., BEGOVICH, C.L. and KANE, V.E. (1980): Monte carlo comparisons of selected clustering procedures. *Pattern Recognition*, 12, 51–6.
- DONOGHUE, J.R. (1995): The effects of within-group covariance structure on recovery in cluster analysis. I. The bivariate case. *Multivariate Behavioral Research*, 30(2):227–254.
- EVERITT, B.S. (1974): *Cluster Analysis*. Heinemann Educational Books, London, UK.
- HUBERT, L. and ARABIE, P. (1985): Comparing partitions. *Journal of Classification*, 2, 193–218.
- MCLACHLAN, G.J. and BASFORD, K.E. (1988): *Mixture Models. Inference and applications to Clustering*. Marcel Dekker, New York.
- MEZZICH, J. E. (1978): Evaluating clustering methods for psychiatric diagnosis. *Biological Psychiatry*, 13(2), 265–281.
- MILLIGAN, G.W. (1980): An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3), 325–342.
- MILLIGAN, G.W. (1981): A review of monte carlo tests of cluster analysis. *Multivariate Behavioral Research*, 16, 379–407.
- MILLIGAN, G.W. (1996): Clustering validation: Results and implications for applied analysis. In: G. De Soete, P. Arabie and L.J. Hubert (Eds.): *Clustering and Classification*. World Scientific Publ., River Edge, NJ, 341–375.
- PRICE, L.J. (1993): Identifying cluster overlap with normix population membership probabilities. *Multivariate Behavioral Research*, 28(2). 235–262.
- SAS Institute Inc. (1989): *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1, ANOVA-FREQ*. SAS Institute, Cary, NC.
- WOLFE, J.H. (1970): Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, 5, 329–350.
- WOLFE, J.H. (1978): Comparative cluster analysis of patterns of vocational interest. *Multivariate Behavioral Research*, 13, 33–44.

What Clusters Are Generated by Normal Mixtures?

Christian Hennig

Institut für Mathematische Stochastik,
Bundesstr. 55, D-20146 Hamburg
(e-mail: hennig@math.uni-hamburg.de)

Abstract. Model based cluster analysis is often carried out by estimation of the parameters of a normal mixture. But mixture components do not necessarily reflect the idea of a “cluster”. I discuss how to formalize the concept of “clusters” w.r.t. probability distributions on the real line by means of fixed point clusters, i.e., sets that do not contain any outlier and with respect to which the rest of the real line consists of outliers. The concept is applied to some normal mixtures.

1 Theoretical clusters of distributions

There are lots of different goals of cluster analysis (CA) and therefore there are lots of different meanings and formal definitions of the term “cluster”. Most of them have in common that a cluster should be a group of entities which is “homogeneous” in some manner, i.e., the entities are similar to each other, and which is in some manner “separated” from the rest.

In CA based on probability models, it is assumed that the entities (usually from \mathbb{R}^p ; here I consider only \mathbb{R}) are generated by some probability distribution. The clusters of the data are determined by using estimators of the features of the distribution (i.e., parameters or regions of high density). That is, the found clusters of data can be viewed as estimators of certain properties of the distribution as well, which I call “theoretical clusters”. In order to explain the very meaning of the data clusters, it is necessary to think about the theoretical clusters generated by the considered model.

Often a CA is based on a normal mixture model of the form

$$P = \sum_{i=1}^s \epsilon_i \mathcal{N}(b_i, \sigma_i^2), \quad \sum_{i=1}^s \epsilon_i = 1, \quad \epsilon_i > 0, \quad i = 1, \dots, s. \quad (1)$$

where $\mathcal{N}(b, \sigma^2)$ denotes the normal distribution with mean b and variance σ^2 , the (b_i, σ_i^2) , $i = 1, \dots, s$, being pairwise distinct. Most of the following considerations apply also to normal fixed partition models. See Bock (1996) for literature concerning the use of these kinds of models in CA.

How to define the theoretical clusters of such normal mixtures? To my knowledge there are three approaches up to now used in the literature more or less implicitly (see Bock (1996) for references):

- 1. Mixture components.** When data points are assigned to the clusters on the basis of the a posteriori probabilities of their generation by the mixture components, the component memberships define the theoretical clusters.
- 2. Mode clusters.** Normal mixtures with many components may be unimodal. This illustrates that these components do not necessarily form reasonable clusters. Theoretical clusters may be defined alternatively by assigning each point to a “nearest” mode (to be defined) of the density.
- 3. Areas of high density.** Since not every mode may be meaningful, all connected subsets of \mathbb{R} where the density exceeds some threshold t can be defined as theoretical clusters.

It is reasonable to demand of a formal definition of a theoretical cluster, that it should not lead to a very different clustering if the distribution is only weakly modified in the sense that the modification can only be observed in case of a very large number of data points. This may be formalized by the Kolmogorov metric. The definitions above have certain drawbacks:

- In arbitrary small Kolmogorov-neighborhoods of a normal mixture there are mixtures with arbitrarily many components. The description of clusters as mixture components loses any sense if the underlying distribution is not precisely a normal mixture.
- In the neighborhoods of a distribution there are distributions with arbitrarily many modes. Moreover, the concept of a probability density is discontinuous. In arbitrary small neighborhoods of a distribution there are other ones with a completely different, or even without, Lebesgue density.
- Areas of density “above t ” are not scale equivariant.

In the next section, fixed point clusters (FPCs) are defined. They provide an alternative definition of a “cluster” of a probability distribution. As mentioned above, there may be lots of ideas about a “cluster” and I do not claim that the approach presented here is more “correct” than others. But it has some reasonable properties that are not shared by the concepts above:

- It does not assume the existence of a Lebesgue density.
- It is a local concept in the sense that it is not based on a model for the whole dataset but only on a model for a single cluster and the relation to its neighborhood. This means in particular that an FPC of a distribution does not lose this property if the distribution is changed in areas far from its location.
- “Homogeneity” and “separation” are formalized directly in a way that varying ideas of a cluster may be implemented by choice of constants defining the degree of similarity to a homogeneous reference distributions and the degree of separation required for a theoretical cluster.

The practical applicability of the approach is discussed briefly in the conclusion.

2 Definition of fixed point clusters

The idea of FPCs is that a set of points is homogeneous if it does not contain any outlier, and separated from the rest if all other points are outliers with respect to it. Davies and Gather (1993) define outliers as follows: First they choose a reference model for the “good” part of the data, say, the family of $\mathcal{N}(b, \sigma^2)$ -distributions, $b \in \mathbb{R}$, $\sigma^2 > 0$. Then they specify an “outlier region” with probability smaller than α , reflecting the idea of “outlyingness”:

$$A(\mathcal{N}(b, \sigma^2), \alpha) = \{(x - b)^2 > c(\alpha)\sigma^2\}, \quad (2)$$

where $c(\alpha)$ is the $(1 - \alpha)$ -quantile of the χ_1^2 -distribution. All points of $A(\mathcal{N}(b, \sigma^2), \alpha)$ are called “ α -outliers” with respect to $\mathcal{N}(b, \sigma^2)$. This definition may be extended to other distributions P on the real line:

$$A(P, \alpha) = \begin{cases} \emptyset & \text{if } E \text{ or } Var \text{ do not exist,} \\ \{(x - EP)^2 > c(\alpha)VarP\} & \text{else.} \end{cases} \quad (3)$$

That is, P is treated as if it were a normal distribution, the reference for a single cluster, and expectation and variance should be meaningful to P . Replacement of EP and $VarP$ by the more robust functionals Median and 1.483 MAD leads to similar results in section 3.

Definition 1. Given some $0 < \alpha < 1$, a subset B of the real line is called a (theoretical) **fixed point cluster** w.r.t. the distribution P , if

$$B = A(P_B, \alpha)^c \text{ and } P(B) > 0, \text{ where } P_B(C) = P(C|B) \forall C.$$

That is, B has to be exactly the set of non-outliers w.r.t. P_B , which is the restriction of P to B itself. The given definition may be adapted to other definitions of outlier regions A , i.e., to other classes of reference distributions (Hennig (1998)) defining other concepts of “homogeneity”. The constant α defines the degree of required separatedness: The smaller α , the more points are non-outliers, and the more separation is needed for a set to be an FPC.

The equation (2) forces all FPCs to be closed intervals. Here is an equivalent representation: The interval $I_{a,s^2} = \{x : (x - a)^2 \leq c(\alpha)s^2\}$ with $s^2 > 0$ is an FPC w.r.t. P iff (a, s^2) is a fixed point of f_P , where

$$f_P(a, s^2) = (EP_{I_{a,s^2}}, VarP_{I_{a,s^2}}). \quad (4)$$

E and Var always exist for distributions of the form $P_{I_{a,s^2}}$, $|a|, s^2 < \infty$.

The following result is a trivial consequence of the definition. It shows that the FPC property of some set B does not depend on untypical areas of \mathbb{R} , “untypical” meaning that they consist only of outliers w.r.t. P_B :

Corollary 1. Let P and Q be distributions on \mathbb{R} and B a set with $P(B) > 0$, $Q(B) > 0$, $P_B = Q_B$. Then B is an FPC w.r.t. P iff it is an FPC w.r.t. Q .

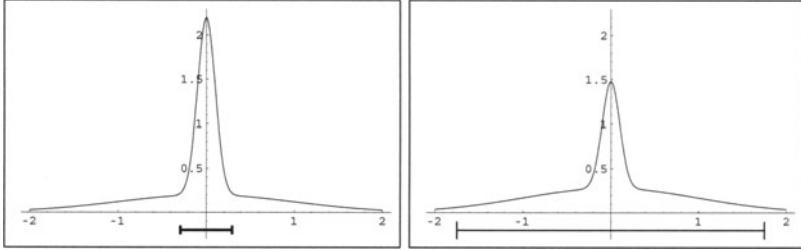


Fig. 1. Densities of $P_{2,0.5}$ and $P_{2,0.7}$ with FPCs.

3 Fixed point clusters of normal mixtures

If P is a normal mixture, f_P may be calculated explicitly. In the following the FPCs of such distributions are discussed. This requires the choice of α which will be discussed below. By default, $\alpha = 0.01$.

Example 1. For $P_1 = \mathcal{N}(0, 1)$ there exists a unique FPC at $(a_1, s_1^2) = (0, 0.9001)$. If $\alpha \leq 0.1$, there always exists a unique FPC with $a = 0$ and $s^2(\alpha) \nearrow 1$ with $\alpha \searrow 0$. This is reasonable: A homogeneous $\mathcal{N}(b, \sigma^2)$ -distribution leads to a single cluster, the interval $[b - 2.44\sigma, b + 2.44\sigma]$ for $\alpha = 0.01$ (FPCs are scale and location equivariant). Corollary 1 says that this interval remains FPC for all distributions of the form $(1 - \epsilon)\mathcal{N}(b, \sigma^2) + \epsilon P^*$, where P^* puts mass 0 on $[b - 2.44\sigma, b + 2.44\sigma]$. If α gets smaller, this interval gets larger.

Example 2. $P_{2,p} = p\mathcal{N}(0, 1) + (1 - p)\mathcal{N}(0, 0.01)$, see Figure 1. This is a mixture with two components, but unimodal. For $p = 0.5$ there exists a unique FPC with $(a_{21}, s_{21}^2) = (0, 0.013)$. The FPC corresponds obviously to the mixture component with variance 0.01. This component is separated well from the other one in the sense that there is only a very small mass from $\mathcal{N}(0, 1)$ in the area of its non-outliers. On the other hand, the area of non-outliers of $\mathcal{N}(0, 1)$ contains almost the whole mass of $\mathcal{N}(0, 0.01)$ so that it does not form a cluster in the sense of fixed point clustering. The situation changes when increasing p : For $p = 0.7$ (Figure 1, right side), the first component puts so much mass in the area of the second one that the second component is no longer clearly separated from the first. A kind of compromise, namely $(a_{22}, s_{22}^2) = (0, 0.461)$, results as the unique FPC. If $p \nearrow 1$, the FPC converges to that of $\mathcal{N}(0, 1)$.

Example 3. $P_{3,d} = 0.5\mathcal{N}(0, 1) + 0.5\mathcal{N}(d, 0.25)$, see Figure 2. For $d = 5$ there are five FPCs defined by $(a_{31}, s_{31}^2) = (0, 0.9)$, $(a_{32}, s_{32}^2) = (5, 0.23)$, $(a_{33}, s_{33}^2) = (0.4, 2.39)$, $(a_{34}, s_{34}^2) = (4.45, 1.89)$ and $(a_{35}, s_{35}^2) = (2.5, 6.87)$. The first two FPCs (marked fat in Figure 2) correspond clearly to the well separated mixture components. One may not expect further theoretical clus-

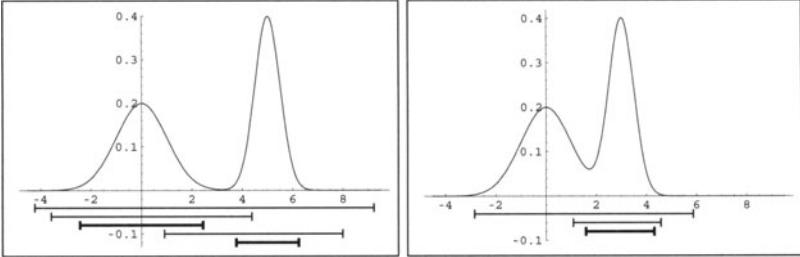


Fig. 2. Densities of $P_{3,5}$ and $P_{3,3}$ with FPCs.

(a, s^2, P)	$D(a, s^2, P)$	(a, s^2, P)	$D(a, s^2, P)$
$(a_{21}, s_{21}^2, P_{2,0.5})$	0.010	$(a_{31}, s_{31}^2, P_{3,5})$	$6 * 10^{-5}$
$(a_{22}, s_{22}^2, P_{2,0.7})$	0.100	$(a_{32}, s_{32}^2, P_{3,5})$	0.003
$(a_{36}, s_{36}^2, P_{3,3})$	0.015	$(a_{33}, s_{33}^2, P_{3,5})$	0.114
$(a_{37}, s_{37}^2, P_{3,3})$	0.068	$(a_{34}, s_{34}^2, P_{3,5})$	0.236
$(a_{38}, s_{38}^2, P_{3,3})$	0.141	$(a_{35}, s_{35}^2, P_{3,5})$	0.207

Table 1. Kolmogorov distances of mixture FPCs to an FPC from a homogeneous normal distribution. Fat numbers correspond to fat marked FPCs in the figures.

ters for this density, but there are three of them. The reason is that the classification of outliers defined in (3) is based on the normal distribution and works well if P is at least similar to a normal. But f_P may have fixed points at intervals I where P_I deviates strongly from a normal distribution.

In section 2 it was discussed that the meaning of “homogeneity” is defined by the class of reference distributions, i.e. the normal family. The homogeneous FPCs in this sense can be distinguished from the non-homogeneous by the requirement that they should be similar to FPCs from a single normal distribution by means of the Kolmogorov metric.

Definition 2. A closed interval $I = [a - \sqrt{c(\alpha)s^2}, a + \sqrt{c(\alpha)s^2}]$ is called a **normal fixed point cluster** w.r.t. P , if it is a fixed point cluster and

$$D(a, s^2, P) = \sup_{x \in I} \left| F_I(x) - \Phi_{a, \frac{s^2}{s^2(\alpha)}, I}(x) \right| < q.$$

where F_I denotes the c.d.f. of P_I , $\Phi_{b, \sigma^2, I}$ denotes the c.d.f. of $\mathcal{N}(b, \sigma^2)$ restricted to I , q is a constant that specifies how close to a normal distribution the theoretical clusters should lie, say, $q = 0.05$ or $q = 0.1$. $s^2(\alpha)$ is defined as in Example 1 and denotes the variance of the FPC of an $\mathcal{N}(0, 1)$ -distribution. Table 1 shows that the Kolmogorov distance serves well to distinguish optically reasonable FPCs from others. The FPCs defined by (a_{22}, s_{22}^2) and

(a_{37}, s_{37}^2) (see below) illustrate the degree of normal similarity required by $q = 0.05$, $q = 0.1$ respectively.

Example 3. Continued. For $d = 3$ there are three FPCs at $(a_{36}, s_{36}^2) = (2.95, 0.28)$, $(a_{37}, s_{37}^2) = (2.83, 0.46)$, $(a_{38}, s_{38}^2) = (1.51, 2.86)$, of which the first one (fat in Figure 2) is the only normal FPC for $q = 0.05$. If d decreases, the mixture components stop to be well enough separated to produce FPCs. Here, the second mixture component with small variance is still sufficiently “cluster-shaped” as opposed to the first one, which disappeared at about $d = 3.7$. At about $d = 2.8$, the FPC with the small variance vanishes as well. The amount of separation necessary for an FPC decreases with increasing α : For $\alpha = 0.025$, the FPC of the second component vanishes at about $d = 2.4$, and for $\alpha = 0.05$, even the first component keeps its FPC down to $d = 2.6$.

4 Conclusion

A proposal is made to define the theoretical clusters for distributions by direct formalization of “homogeneity” and “separation”. While the definition of “homogeneity” is based on the normal reference distributions here, the FPCs may be calculated for arbitrary distributions. It is clear that the meaning of a cluster has to depend on individual aims and judgements. The approach of fixed point clustering shows where such judgements may enter: The required separation can be varied by choice of α , the homogeneity by the choice of the definition of the outlier region A and by the cutoff q for the Kolmogorov distance. There are two ways to apply the FPC concept to concrete data:

- The computation of theoretical FPCs of an estimated distribution can help to decide about the occurrence of well separated, homogeneous clusters, if a density or the parameters of a normal mixture have been estimated.
- The definition of FPCs can be transferred directly to datasets. The resulting CA method is sketched for various definitions of outlier regions in Hennig (1998) and will be analyzed elsewhere in greater detail. Software for this is available from

<http://www.math.uni-hamburg.de/home/hennig/>

References

- BOCK, H.H.(1996): Probability Models and Hypotheses Testing in Partitioning Cluster Analysis. In: P. Arabie, L. J. Hubert, G. De Soete (Eds.): *Clustering and Classification*. World Scientific Publishers, New Jersey, 377–453.
- DAVIES, P.L. and GATHER, U.(1993): The identification of multiple outliers. *Journal of the American Statistical Association*, 88, 782–801.
- HENNIG, C.(1998): Clustering and Outlier Identification: Fixed Point Cluster Analysis. In: A. Rizzi, M. Vichi, H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer, Berlin, 37–42.

A Bootstrap Procedure for Mixture Models

Suzanne Winsberg¹ and Geert deSoete²

- ¹ IRCAM,
1 Place Igor Stravinsky, Paris 75004, France
(e-mail: winsberg@ircam.fr)
- ² ARC, University of Ghent,
Krijslaan 281, B9000, Ghent, Belgium

Abstract. A bootstrap procedure useful in latent class models has been developed to determine the sufficient number of latent classes required to account for systematic group differences in the data. The procedure is illustrated in the context of a multidimensional scaling latent class model, CLASCAL. Real and artificial data are presented. The bootstrap procedure for selecting a sufficient number of classes seems to correctly select the correct number of latent classes at both low and high error levels. At higher error levels it outperforms Hope's (1968) procedure.

1 Introduction

Latent class formulations, or more general mixture distribution approaches, have been explored in many contexts. In all of these applications, latent class modeling has been shown to be a useful technique for capturing systematic group differences in a parsimonious way. A problem encountered in the latent class approach is to determine the appropriate number of latent classes. A Monte Carlo procedure suggested by Hope (1968) has been used to address this problem. We propose another approach using the bootstrap. The bootstrap, introduced by Efron (1979), is a resampling technique in which B random samples are drawn with replacement from the observed values. Our technique can be used in lieu of or in addition to the Hope test. This bootstrap approach may be used in any latent class context. DeSoete and DeSarbo (1991) used a resampling technique for a pick any/ N latent class model. Although our technique applies in any latent class context, we illustrate the approach in the multidimensional scaling context using CLASCAL, a latent class MDS model, (Winsberg and De Soete, 1993). Multidimensional scaling, MDS, may be defined broadly as a set of multivariate models and methods for representing objects as points in a multidimensional space of low dimension, given pairwise dissimilarity measures for the objects.

2 Bootstrapping the appropriate number of classes

The bootstrap algorithm begins by generating a large number of independent bootstrap samples $x^{*1}, x^{*2}, \dots, x^{*B}$, each of size N , where N is the sample size

of the original data (see Efron & Tibshirani, 1993). Each individual bootstrap sample is generated by sampling with replacement from the data N times.

The problem before us is to determine how many latent classes are required to account for group differences in a data set. Let us consider, for example, a data set consisting of pairwise dissimilarity judgments for J objects or stimuli, that is $J(J - 1)/2$ stimuli pairs, obtained from N sources, or subjects. The researcher may employ an MDS CLASCAL analysis, to locate the objects in a low dimensional weighted Euclidean space. To test the postulate that T classes are sufficient to account for group differences in the data, we propose to conduct three bootstrap analyses, one bootstrap analysis with T classes, a second bootstrap analysis with $(T - 1)$ classes, and the third bootstrap analysis with $(T + 1)$ classes.

For each of these analyses draw B bootstrap samples of size N ; that is, draw B samples with replacement of size N . Consider the analysis for T classes. Construct the $N \times N$ matrix in which the ik th element is $1 - (\text{the number of times the } i\text{ th subject is a member of the same class as the } k\text{ th subject divided by the number of times both the } i\text{ th and } k\text{ th subject appear in any of the bootstrap samples})$. This matrix, denoted D_T , will have elements which vary between zero and one. The resultant matrix, D_T , containing a measure of dissimilarity for each pair of *subjects*, should then be analyzed with an average-link-hierarchical-clustering algorithm. We have so far restricted our attention to this type of clustering algorithm. If indeed there are T classes, the clustering analysis will yield a tree that shows T distinct groups. Three clustering analyses, are performed: on D_{T-1}, D_T, D_{T+1} . In one of these analyses the number of clusters will be consonant with the number of latent classes; that value is the appropriate number of latent classes. Naturally, one may extend the range of this testing procedure, if necessary, or desirable, by creating $D_{T-q}, D_{T-q+1}, \dots, D_{T+q}$, where q is any integer, and performing the subsequent clustering analyses, until one finds a number of clusters equal to the number of latent classes. Moreover, the series need not be symmetric about T . That is, one could equally well test from $D_{T-q}, D_{T-q+1}, \dots, D_T, \dots, D_{T+r}$, where q and r are any integers. In general, however, one may have a good candidate for T ; then, $q = r = 1$ is adequate.

3 An artificial and a real data example

We have analyzed two sets of artificial data, comprising pairwise dissimilarity judgements from 16 subjects on 20 rectangles, differing as to shape and area. These artificial data are based on real data we have previously analyzed, (Winsberg and DeSoete, 1997). For the artificial data sets, the distances were normally distributed around the model values obtained from the analysis of the real data set, with the error variance equal to 15 percent of the standard deviation of the model distance values for set 1, and 45 percent of the standard deviation of the model distance values for set 2.

Using our bootstrap approach we were correctly able to determine the number of latent classes for both artificial sets. However, for these examples Hope's technique recovered the correct number of latent classes at the lower error level but failed to do so, at the higher error level.

Since the analysis of the artificial rectangle data set revealed a discrepancy between our bootstrap procedure and Hope's procedure for selecting the appropriate number of latent classes, with only our bootstrap procedure pointing to the true number of latent classes in both cases, we applied our bootstrap procedure to the real rectangle data. We used the bootstrap procedure to verify the hypothesis that $T = 3$ latent classes are sufficient, the number of latent classes selected by Hope's procedure for the real data. So bootstrap analyses were performed for 2, 3, and 4, ($T - 1, T, T + 1$) latent classes respectively. The bootstrap procedure selects two latent classes.

The first recovered dimension was size, that is area, the second and third dimensions relate to shape factors. The second dimension, we denoted as up-down, and related to whether the height-width ratio was greater or less than one. This dimension was essentially categorical with three categories: $h/w < 1$, $h = w$, $h/w > 1$. The third dimension we denoted as squareness, and was related to h/w or w/h , whichever was less than one. For class one, consisting of 7 subjects, the weights for the three dimensions were: 1.05, 0.38, and 0.68 respectively, for class two, consisting of 8 subjects, the weights were: 0.95, 1.62, and 1.31 respectively. One subject was poorly classified. Class one weights area most, and up-down little; class two weights up-down most, next squareness, and area least. Clearly, shape, corresponding to the two dimensions up-down and squareness, is more important for class 2, and size, corresponding to area, is more important for class 1. This two-class solution, reflects more clearly than did the three-class solution, the predilection for a subject to emphasize either size (by approximately a two-to-one ratio), or shape factors (by a three-to-two ratio), when making dissimilarity judgments on rectangles. In the three class solution the classes were primarily distinguished from one another by how strongly they emphasised top-down as compared to the other two dimensions size and squareness.

In conclusion this bootstrap technique is useful for deciding how many latent classes should be retained.

References

- DE SOETE, G. and DE SARBO, W. (1991). A latent class Probit model for analyzing pick any/N data. *Journal of Classification*, 8, 45-63.
- EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- EFRON, B. and TIBSHIRANI, R. (1993). *An Introduction to the Bootstrap* Chapman and Hall, New York.
- HOPE, A.C. (1968). A simplified Monte Carlo test procedure. *Journal of the Royal Statistical Society, Series B* 30, 582-598.

- WINSBERG, S. and DE SOETE, G. (1993). A latent class approach to fitting the weighted Euclidean model, CLASCAL. *Psychometrika*, 58, 315–330.
- WINSBERG, S. and DE SOETE, G. (1997). Multidimensional scaling with constrained dimensions: CONSCAL. *British Journal of Mathematical and Statistical Psychology*, 50, 55–72.

A New Criterion for Obtaining a Fuzzy Partition from a Hybrid Fuzzy/Hierarchical Clustering Method

Arnaud Devillez, Patrice Billaudel and Gérard Villermain Lecolier

Laboratoire d'Automatique et de Microélectronique,
Faculté des Sciences - Moulin de la Housse - B.P. 1039 - 51687 Reims Cedex 2
e-mail : arnaud.devillez@univ-reims.fr

Abstract. Classical fuzzy clustering methods are not able to compute a partition into a set of points, when classes have non-convex shape. Furthermore, we know that in this case, the usual criteria of class validity, such as fuzzy hypervolume or compactness - separability, do not allow one to find the optimal partition.

The purpose of our paper is to provide a criterion allowing one to find the optimal fuzzy partition in a set of points including classes of any shape. To that effect we shall use the Fuzzy C Means algorithm to divide the set of points into an overspecified number of subclasses. A fuzzy relation is established between them in order to extract the structure of the set of points. The subclasses are merged according to this relation and the criterion that we propose allows one to find the optimal regrouping.

1 Introduction

Fuzzy clustering methods give good results when the shape of classes is elliptic and when these classes are separable by a hyperplane. In any other cases it is absolutely necessary to use a classic hierarchical method which has several deficiencies. Indeed what we obtain as a result is a hard partition of the set of points, the notion of graduated membership does not exist. Moreover the addition of a point changes possibly in a major way, the structure of the hierarchy what may entail some important calculations. To make up for these inconveniences we have proposed a fuzzy hierarchical clustering method called UFGC (Billaudel et al. (1999)). The purpose of our paper is to introduce a new criterion allowing one to find the optimal cut which must be realised in the fuzzy relation or in the hierarchy to find the real classes.

2 Division of the set of points into subclasses

The first stage of our algorithm consists in using the Fuzzy C-Means algorithm (Bezdek (1981)) to divide the set into c' subclasses. The result of this algorithm is a fuzzy membership matrix U' . The element u'_{kj} represents the

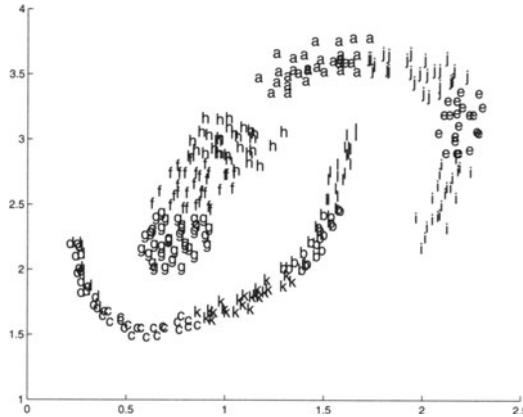


Fig. 1. Division of the set into 12 subclasses. All points marked by a similar letter belong to the same subclass.

membership degree of sample x_j to subclass SC_k . The number of subclasses must be much greater than the number of real classes (Billaudel et al. (1998)). In our example c' must be higher than 3. We have obtained the partition into the 12 subclasses shown on Figure 1. Each subclass belongs to one real class only. Points are affected to the class for which they have their maximum membership degree.

3 Similarity degrees between subclasses

The neighbourhood or proximity between two subclasses can be quantified. The coefficient we are using is the similarity degree defined by Krishnapuram (Frigui and Krishnapuram (1996)) :

$$n_{kl} = 1 - \frac{\sum_{x_j \in SC_k \text{ or } x_j \in SC_l} |u'_{kj} - u'_{lj}|}{\sum_{x_j \in SC_k} u'_{kj} + \sum_{x_j \in SC_l} u'_{lj}}.$$

We obtain a neighbourhood matrix N , the size of which is $c' \times c'$. The element n_{kl} corresponds to the proximity value between subclass SC_k and subclass SC_l . The nearer the value n_{kl} is to 1, the closer the subclasses are to one another. Terms on the diagonal are all equal to 1 because each subclass is a neighbour of itself.

4 Construction of the fuzzy proximity graph

The neighbourhood matrix N defines a fuzzy proximity graph in which the vertices represent the subclasses and the arcs represent the links. The arcs

are graduated by the proximity degrees. The graph corresponding to our example includes 12 vertices and 66 arcs, knowing that the arcs linking a subclass to itself are not represented and that matrix N is symmetrical. For a comfortable representation we use a reduced graph representing the highest $c' - 1$ proximity degrees linking the c' vertices presented on Figure 2. We note that this graph reflects structures existing in the set of points. Calculations are based on matrix N , consequently they take account of the complete graph.

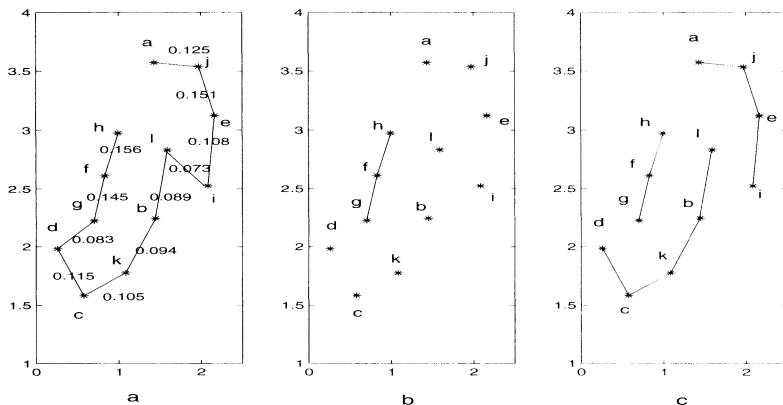


Fig. 2. Reduced fuzzy proximity graph (a) and graphs obtained by relations of different levels (b) and (c)

5 Defuzzification of the proximity graph

The neighbourhood matrix N defines a fuzzy relation (Bouchon-Meunier (1995)) which expresses the idea of neighbourhood between subclasses. It can be decomposed into a series of relations of level $\alpha : N^\alpha$ defined by :

$$n_{kl}^\alpha = 1 \text{ if } n_{kl} \geq \alpha \text{ and } n_{kl}^\alpha = 0 \text{ if } n_{kl} < \alpha.$$

This decomposition entails the defuzzification of the proximity graph into a hard graph which reveals the structure of the set of points. Two subclasses are then considered as neighbouring and consequently merged if an arc exists between the two vertices which correspond to them. For the two levels of cut $\alpha_1 = 0.12$ and $\alpha_2 = 0.085$, we have obtained the two hard neighbourhood graphs shown on Figure 2.

6 Fusion of subclasses

It is necessary to achieve a fusion of subclasses existing in each connected component of the hard proximity graph. This fusion is made by a sum in

order to obtain a membership degree for a point x to each real class. When a number s_i of subclasses SC_k exists within a class C_i , the membership function of a point x_j to this class is defined by :

$$u_{ij}(x_j) = \sum_{k=1}^{s_i} u'_{kj}(x_j).$$

The new membership matrix is calculated from matrix U' . We obtain a membership matrix U the size of which is $c \times n$ with c being the number of connected components and n the number of points.

7 Research of the optimal level of the cut

Few criteria can help us to determine the optimal partition of a set of points. Those proposed in (Milligan and Cooper (1985), Bezdek (1981)) give good results with classes of spherical or elliptical shapes. With classes of complex shapes there exists no adequate criterion to find the optimal partition. The criterion we propose, uses the principle of the compactness criterion defined in (Xie and Beni (1991)). The global and average compactness of a partition are respectively defined by:

$$co_{gl} = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n (u_{ij} \cdot d_{ji})^2 \text{ and } co_{av} = \frac{1}{c} \sum_{i=1}^c \frac{\sum_{j=1}^n (u_{ij} \cdot d_{ji})^2}{\sum_{j=1}^n u_{ij}},$$

where d_{ji} represents the distance from the point x_j to the nearest centroid of class C_i computed during the division of the set of points into subclasses by Fuzzy C Means algorithm. The definition of our criterion is based on the equivalence between average and global compactness when the optimal partition into c^* classes is attained. Thus the minimization of the following criterion allows one to determine the number of classes existing in the set of points :

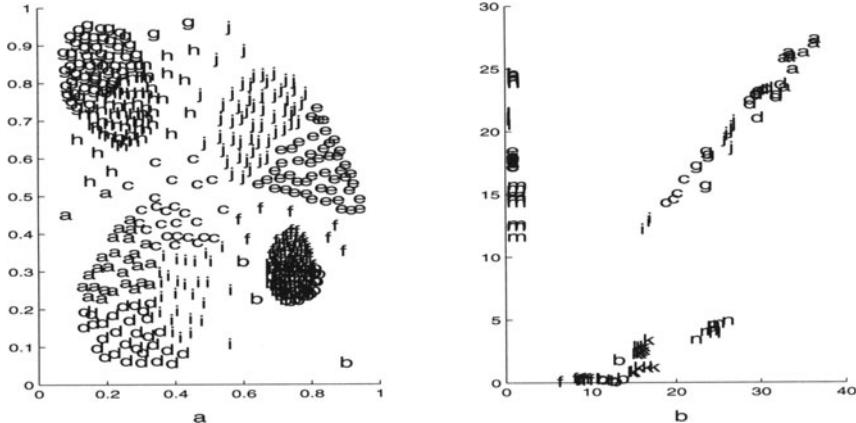
$$K_c = \left| 1 - \frac{co_{gl}}{co_{av}} \right|.$$

Table 1 illustrates the results we obtained for our example. We can note that the minimum is attained for three classes. The partition corresponds to the real classes.

8 Applications

We have applied our criterion to two examples in order to show its ability to determine the number of classes existing into a set of points. These examples include difficulties currently met into the classification field.

$\alpha \cdot 10^{-2}$	8	8.5	9	10	10.5	11	12	14	15	15.5	16
$K_c \cdot 10^{-2}$	0.5	0.25	6.48	9.46	12.2	8.6	11	6.47	3.22	1.37	1
Number of classes	2	3	4	5	6	7	8	9	10	11	12

Table 1. Values of the criterion as function of the number of classes**Fig. 3.** Sets of points of examples 1 (a) and 2 (b). All the points marked by the same letter belong to the same subclass.

8.1 Example 1

The set of points is presented in Figure 3a. It comprises 441 points divided into 4 classes of varying shapes and densities. We have added some ambiguous points between classes. This set is divided into 10 subclasses as is shown in Figure 3a. Table 2 shows the values of criterion K_c as a function of the number of classes. We can note that the minimum is attained for a value of 4 classes, which are similar to the real ones. Our criterion succeeds for this set of points.

$\alpha \cdot 10^{-2}$	8.5	9	16	16.2	17	20	22	24	26
$K_c \cdot 10^{-2}$	0.75	1.66	0.27	1.41	2.57	5.67	2.79	2.01	2.94
Number of classes	2	3	4	5	6	7	8	9	10

Table 2. Values of the criterion as function of the number of classes

8.2 Example 2

To finish we have applied the method to the set of points coming from a study concerning the classification of plastic bottles and presented in Figure 3b. It comprises 3 classes of elongated shape and was divided into 14 subclasses as

it is shown in Figure 3b. The number of classes found by criterion K_c is 3 as we can see in Table 3.

$\alpha \cdot 10^{-2}$	0.5	0.7	1.3	1.4	4.3	5.1	6.8	7.4	8.5	9	10
$K_c \cdot 10^{-2}$	14.4	0.07	24.6	24.6	36.8	20.8	23.4	21.3	14.5	10.6	3.91
Number of classes	2	3	4	5	6	7	8	9	10	12	15

Table 3. Values of the criterion as function of the number of classes

9 Conclusion

This study shows the ability of our criterion to determine the number of classes in a set of points with our fuzzy hierarchical classification method. Our methodology combines the fuzzy c-means and single link methods of analysis. It shows several positive aspects : the number of classes may be unknown and classes can have complex shapes, something which was impossible to detect accurately by using solely the fuzzy c-means algorithm.

The criterion determines the optimal level of the cut to be made in the order relation and it performs better than the usual criteria of class validity. This cut provides the subclasses to merge in order to build the real classes. Thus with such a method, the fuzzy partition of the set of points is conserved contrary to the usual hierarchical methods.

References

- BEZDEK, J. C. (1981) : *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New-York.
- BILLAUDEL, P., DEVILLEZ, A. and VILLERMAIN LECOLIER , G. (1998) : An unsupervised fuzzy classification algorithm for the non elliptic classes, *EURISCON'98*, June 22 - 25, Athens, Greece.
- BILLAUDEL, P., DEVILLEZ, A. and VILLERMAIN LECOLIER , G. (1999) : Unsupervised fuzzy classification method based on a proximity graph and a graduated hierarchy, *FUZZ-IEEE'99*, 1054-1058 August 22 - 25, Seoul, Korea.
- BOUCHON-MEUNIER, B.(1995) : *La logique floue et ses applications*, Addison Wesley, Paris.
- FRIGUI, H. and KRISHNAPURAM, R. (1996) : A robust algorithm for automatic extraction of an unknown number of clusters from noisy data, *Pattern Recognition Letters*, 17, 1223-1232.
- MILLIGAN, G. W. and COOPER, M. C.(1985) : An examination of procedures for determining the number of clusters in data set, *Psychometrika*, 50, 159-179.
- TRAUWERT, E., KAUFMAN, L. and ROUSSEEUW, P.(1991) : Fuzzy clustering algorithms based on the maximum likelihood principle, *Fuzzy Sets and Systems*, 42, 213-227.
- XIE, X. L. and BENI, G.(1991) : A validity measure for fuzzy clustering, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13, 841-847.

Application of Fuzzy Mathematical Morphology for Unsupervised Color Pixels Classification

A. Gillet, C. Botte-Lecocq, L. Macaire and J.-G. Postaire

Laboratoire d'Automatique I³D

Université des Sciences et Technologies de Lille - Cité Scientifique - Bâtiment P2

59655 Villeneuve d'Ascq, FRANCE

(e-mail: ag@cal.univ-lille1.fr)

Abstract. In this paper, we present a new color image segmentation algorithm which is based on fuzzy mathematical morphology. After a color pixel projection into an attribute space, segmentation consists of detecting the different modes associated with homogeneous regions. In order to detect these modes, we show how a color image can be viewed as a fuzzy set with its associated membership function corresponding to a mode which is defined by a color cooccurrence matrix and by mode concavity properties. A new developed fuzzy morphological transformation is then applied to this membership function in order to identify the modes. The performance of our proposed fuzzy morphological approach is then presented using a test color image, and is then compared to the competitive learning algorithm.

1 Introduction

Color segmentation partitions a color image into disjoint regions which are sets of connected pixels containing uniform color feature characteristics. The approaches to the segmentation of color images can be classified into four groups, namely, histogram-based techniques, neighbourhood-based segmentation, physically-based segmentation techniques, and multidimensionnal data classification methods. For this last group of methods, color image segmentation is achieved by pixel classification according to color features. Pixels can be associated with observations whose coordinates are their color features in a color attribute space. Parts of this attribute space with high local concentration of observations can be considered as modes separated by valleys, which are parts of the attribute space with low local concentration of observations. In the field of multidimensional data analysis, several mathematical morphological operators have been developed to extract the different modes which are associated to each cluster of the analysed set of observations (Park et al. (1998)).

We present in this paper an original method based on adaptive fuzzy morphological tools for unsupervised color pixel classification. First, we show how we extract, from a color image, a fuzzy set with its associated membership function. Then, we present a fuzzy morphological transformation of this

membership function in order to detect the different modes. Finally, our proposed classification procedure is applied to the segmentation of a color image which contains several different colored balls. We compare the segmentation results of our method with those of a well-known segmentation method, based on competitive learning (Uchiyama and Arbib (1994)).

2 Construction of a fuzzy set associated with a color image

Many clustering procedures, described in the literature, are based on the detection of the modes of the probability density function (*pdf*) underlying the distribution of the observations in the data space. So, we propose to first estimate the underlying *pdf* by a color texture measure, the color cooccurrence matrix. Then, in order to detect the modes of the *pdf*, we construct from this cooccurrence matrix a fuzzy set X which, together with its associated membership function yields the degree of membership of each element of X to a "pdf mode", by the analysis of the local convexity of the *pdf*.

2.1 Estimation of the *pdf* by a color cooccurrence matrix

The color of a pixel $P(i, j)$, where i and j are the spatial coordinates in the color image, can be represented by three levels (c_1, c_2, c_3) according to the considered color representation system (C_1, C_2, C_3) . We consider that this pixel is associated to an observation $X(c_1, c_2)$, whose coordinates are (c_1, c_2) in the 2-D attribute space. The estimation of the *pdf* can be obtained by the evaluation of a color cooccurrence matrix which takes into account the spatial and colorimetric interactions between pixels (Tréneau et al. (1996)).

A Color Cooccurrence Matrix, denoted CCM , is defined as a table whose inputs are two color features C_1 and C_2 in order to reduce the computational complexity. Furthermore, let X be the set of N^2 observations, or points, where N is the number of values of each color feature. Let $CCM(X(c_1, c_2))$ be the number of times that a pixel, with a level equal to c_2 of the color feature C_2 is in the 24-neighbourhood of a pixel with a level equal to c_1 of the color feature C_1 . This CCM is nothing else than a discrete *pdf*. If this evaluated *pdf* is normalized between 0 to 1, it can be considered as a membership function, in a fuzzy set context, in the restricted case where the different clusters are equiprobable. In the other cases, the membership function must be derived from a fuzzification phase.

2.2 Evaluation of a membership function to a mode

The membership function corresponding to a mode can be evaluated by considering each mode as a part of the attribute space where the *pdf* is concave (Postaire and Vasseur (1980)). This approach is based on the local convexity

of the *pdf* estimated by the values of the *CCM* at an observation $X(c_1, c_2)$, determined by analysing the variation of that *pdf* estimator when the observation domain grows around $X(c_1, c_2)$. We first determinate an estimator $\hat{p}_1(X(c_1, c_2))$ of the *pdf* using an observation domain $D_1(X(c_1, c_2))$, and then, we evaluate another estimator $\hat{p}_2(X(c_1, c_2))$ achieved with an observation domain $D_2(X(c_1, c_2))$, slightly larger than $D_1(X(c_1, c_2))$, such as:

$$\begin{aligned}\hat{p}_1(X(c_1, c_2)) &= \sum_{i=-n/2}^{n/2} \sum_{j=-n/2}^{-n/2} CCM(X(c_1 + i, c_2 + j)) / (n+1)^2 \\ \hat{p}_2(X(c_1, c_2)) &= \sum_{i=-m/2}^{m/2} \sum_{j=-m/2}^{-m/2} CCM(X(c_1 + i, c_2 + j)) / (m+1)^2\end{aligned}$$

where $n+1$ is the size of D_1 and $m+1$ is the size of D_2 , with $m > n$. In these conditions, if $\hat{p}_2(X(c_1, c_2)) > \hat{p}_1(X(c_1, c_2))$, the estimated *pdf* is considered as locally convex, so that the observation $X(c_1, c_2)$ will get a "local convex" label. Otherwise, $X(c_1, c_2)$ will get a "local concave" label.

Since we define a mode as the part of the attribute space where the *pdf* is considered as concave, we introduce a Pi-fuzzification function where $f(X)$ associated to $X(c_1, c_2)$ represents the likelihood degree of the event " $X(c_1, c_2)$ is a modal observation". This depends on k_1 , the number of "concave" observations which are included in $D_1(X(c_1, c_2))$. The fuzzification function used in this paper is defined by :

$$\begin{aligned}\text{if } f(X) \leq a, \quad \text{then } \mu(X(c_1, c_2)) = 0 \\ \text{if } a < f(X) < b, \quad \text{then } \mu(X(c_1, c_2)) = \frac{1}{2} + \frac{1}{2} \sin \frac{\pi}{b-a} (f(X) - \frac{a+b}{2}) \\ \text{if } f(X) \geq b, \quad \text{then } \mu(X(c_1, c_2)) = 1\end{aligned}$$

Henceforth, the set X of observations is considered as a fuzzy set with its associated mode membership function μ_X . To illustrate this method, we show in Figure 1 a color image, which contains five kinds of homogeneous regions : four kinds of colored balls and the background. Figure 2 shows the mode membership function extracted from this color image. Here, we used a 5x5 neighborhood window for the domain $D_1(X(c_1, c_2))$ and a 7x7 neighborhood window for $D_2(X(c_1, c_2))$. In this figure, we can see that the five clusters that we have to detect are surrounded by a considerable amount of noise.

3 Mode detection by fuzzy morphological operators

In classical morphology, the basic operations of erosion enlarges the valleys, by eliminating the irregularities of the distribution, and it also tends to shrink the modes. On the other hand, dilation enlarges the modes but it also tends to enhance the valleys (Bloch and Maitre (1995)). It would therefore be interesting to erode and to dilate the membership function, to a degree which correspond to the distance of each observation from a mode in the attribute space (Turpin-Dhilly and Botte-Lecocq (1998)).

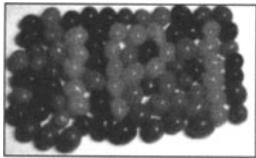


Fig. 1. Original color test image

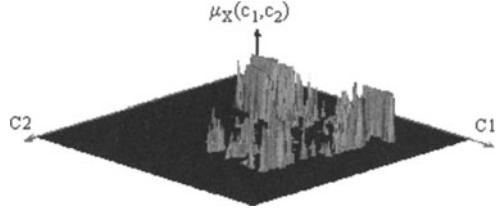


Fig. 2. Membership function $\mu_X(c_1, c_2)$

3.1 Fuzzy erosion and dilation

Several definitions of fuzzy erosion and dilation can be found in the literature. In this paper, fuzzy erosion and fuzzy dilation are defined by :

$$E_\nu[\mu_X(X(c_1, c_2))] = \inf_{Y \in S_{X(c_1, c_2)}^\nu} \max [\mu_X(Y), 1 - \nu(Y - X(c_1, c_2))] \\ D_\nu[\mu_X(X(c_1, c_2))] = \sup_{Y \in S_{X(c_1, c_2)}^\nu} \min [\mu_X(Y), \nu(Y - X(c_1, c_2))]$$

where $S_{X(c_1, c_2)}^\nu$ is the observation domain of size $n + 1$ centered at the observation $X(c_1, c_2)$ and limited by the structuring element S^ν . The definition of the structuring function ν associated to S^ν depends on the problem to solve.

3.2 Definition of structuring functions

We want the degree of erosion and dilation to depend on whether an observation is located at a valley or at a mode. This can be approximated by the observation's mode membership degree. Indeed, if $\mu_X(X(c_1, c_2))$ is low (resp. high), the observation $X(c_1, c_2)$ is supposed to be located at a valley (resp. a mode). This localisation of each observation in the attribute space can be complemented by the gradient of $\mu_X(X(c_1, c_2))$. Indeed, modes and valleys are separated by border parts where the membership function level is highly varying. By means of the analysis of its membership degree and of its gradient value, we are able to decide if an observation is located at a mode, a valley or at a border. Note that both dilation and erosion definitions use a structuring function ν . We propose, in this paper, an original structuring function ν_E for the erosion operation and a specific structuring function ν_D for the dilation operation. These structuring functions are, respectively, defined by :

$$\nu_E(X(c_1, c_2)) = [1 - \mu_X(X(c_1, c_2))(1 - g_{\mu_X}(X(c_1, c_2)))] = [1 - A] \\ \nu_D(X(c_1, c_2)) = [\mu_X(X(c_1, c_2))(1 - g_{\mu_X}(X(c_1, c_2)))] = [A]$$

where $g_{\mu_X}(X(c_1, c_2))$ is the gradient of μ_X evaluated at the observation $X(c_1, c_2)$. The higher (resp. lower) the quantity A , the more the observation $X(c_1, c_2)$ can be considered to be modal (resp. located at a valley or at a border part). Consequently, the effect of fuzzy dilation on this observation is higher (resp. very low), and the effect of fuzzy erosion is lower (resp. higher).

3.3 Mode detection by fuzzy morphological transformation

In order to extract the different modes associated with each cluster, we define a new fuzzy morphological transformation combining fuzzy erosion of the fuzzy set X using the structuring function ν_E , followed by fuzzy dilation of the resulting fuzzy set using the structuring function ν_D . This transformation, denoted t , is applied to the fuzzy set X and yields the resulting fuzzy set X^t , where the new mode membership function μ_X^t is defined by :

$$\mu_X^t = t(\mu_X) = D_{\nu_D}[E_{\nu_E}(\mu_X)]$$

We iterate this transformation until the stabilisation of the resulting mode membership function X^T is reached : $\mu_X^T = (\mu_X^t)^\infty$ with $(\mu_X^t)^j = t[(\mu_X^t)^{j-1}] = D_{\nu_D}[E_{\nu_E}[(\mu_X^t)^{j-1}]]$, $(\mu_X^t)^1 = \mu_X$.

The kernel of this fuzzy set X^T , defined by : $\mu_X^T(X(c_1, c_2)) = 1$, yields the "modal observations" $X(c_1, c_2)$ which certainly belong to the different modes. Figure 3 presents the result of this transformation where we can see that the five different modes are well detected.

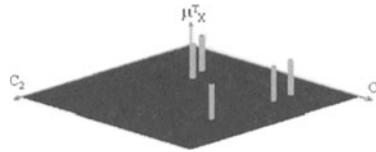


Fig. 3. Resulting mode membership function μ_X^T

4 Experimental results

The different modes which are extracted from the fuzzy set X^T , are associated with each class of pixels of the analysed color image. In this section, we present the pixel labeling schemes of the different constructed classes and then, we compare the performance of the proposed fuzzy morphological approach, with the competitive learning method.

4.1 Pixel labeling

The labeling procedure consists of assigning each pixel $P(i, j)$ whose color features are (c_1, c_2) to the class whose Euclidean distance between its corresponding mode center and $X(c_1, c_2)$ is lowest. The color features of a labelled pixel are equal to the coordinates of the associated mode center (Figure 4).

4.2 Comparison with competitive learning segmentation

We now apply the segmentation scheme which uses the competitive learning technique (Uchiyama and Arbib (1994)). Figure 5 presents the results of this segmentation approach. As it is difficult to compute a pixel confusion matrix

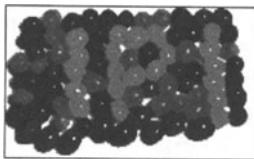


Fig. 4. Segmentation by fuzzy morphological approach.

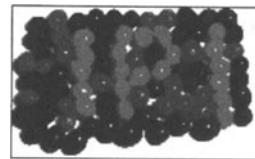


Fig. 5. Segmentation by competitive learning method.

from this image, we compute a table (Table 1) which indicates the number of the different-color balls which are reconstructed by the two methods. The criterion to decide whether a ball is well reconstructed is only a visual criterion.

	<i>Test Image</i>	<i>Competitive learning</i>	<i>Fuzzy segmentation</i>
<i>Black</i>	24	34	25
<i>Green</i>	17	7	16
<i>Red</i>	20	20,5	20
<i>Orange</i>	21	20,5	21

Table 1. Constructed balls

5 Conclusion

We have presented in this paper, an original color image segmentation algorithm which is based on a fuzzy morphological transformation. Each homogeneous region are associated to a mode by projecting each pixel of the image in an 2-D space. The *pdf* of each observation is evaluated thanks to a color cooccurrence matrix. A fuzzy morphological transformation, based on iterative fuzzy erosions and dilations, is applied to the membership function extracted from the color image in order to detect the different modes. Each pixel of the original color image is finally assigned to the class associated with the appropriate detected mode. A comparison between this segmentation method and a more classical color segmentation technique has shown the efficiency of our proposed procedure.

References

- BLOCH I. and MAITRE H. (1995): Fuzzy mathematical morphologies : A comparative study. *Pattern Recognition*, 9(28):1341–1387.
 PARK S.H., YUN I.D. and LEE S.U. (1998): Color image segmentation based on 3-D clustering : morphological approach. *Pattern Recognition*, 31(8):1061–1076.

- POSTAIRE J.-G. and VASSEUR C. (1980): A convexity testing method for cluster analysis. *I.E.E.E. Transactions on Systems and Man Cybernetics*, 10:145–149.
- TREMEAU A., COLANTONI P. and LAGET B. (1996): On color segmentation guided by the cooccurrence matrix. *OSA Annual conference on Optics and Imaging in the Information Age*, 30–38.
- TURPIN-DHILLY S. and BOTTE-LECOCQ C. (1998): Application of fuzzy mathematical morphology for pattern classification. *Advances in Data Science and Classification, Proceeding of the 6th Conference of IFCS'98*, 125–130.
- UCHIYAMA T. and ARBIB M. A. (1994): Color image segmentation using competitive learning. *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence*, 16(12):1197–1206.

A Hyperbolic Fuzzy k -Means Clustering and Algorithm for Neural Networks

Norio Watanabe¹, Tadashi Imaizumi², and Toshiko Kikuchi³

¹ Department of Industrial and Systems Engineering, Chuo University,
Kasuga 1-13-27, Bunkyo-ku, Tokyo 112-8551, Japan
(e-mail: watanabe@indsys.chuo-u.ac.jp)

² Tama University,
Hijirigaoka, Tama-shi, Tokyo 206-0022, Japan

³ Faculty of Economics, Tohoku Gakuin University,
Sendai-shi, Miyagi 980-8511, Japan

Abstract. A new fuzzy k -means clustering algorithm is proposed by introducing crisp regions of clusters. Boundaries of the regions are determined by hyperbolas and membership values are given by one or zero in each region. The area between crisp regions is a fuzzy region, where membership values are proportional to distances to crisp regions. Though the traditional hard k -means is a limit of the usual fuzzy k -means, results of the latter are fuzzy and then are not the same as results of the former. On the other hand a new method can produce the same results as those by the hard k -means. An algorithm for neural networks is given and a numerical example is illustrated.

1 Introduction

In this paper we try to extend the usual k -means clustering, which we call the hard k -means, to fuzzy clustering. A new method permits the existence of crisp regions of clusters, in which membership values are given by one or zero. Since boundaries of regions are determined by hyperbolas, we call our method the hyperbolic fuzzy k -means.

One of the well known fuzzy methods is the fuzzy c -means algorithm by Bezdek (1981). We call it the fuzzy k -means in this paper. Results of the usual fuzzy k -means are fuzzy. Only the data on the center can have the membership value one. Though the hard k -means is the limit of the usual fuzzy k -means when a parameter tends to some value, the former is not a special case of the latter. Thus there is a gap between the hard k -means and the fuzzy k -means. On the other hand, by choosing a special value of a parameter, our method is identical to the hard k -means. Moreover, even if parameters are not special, our fuzzy method can produce crisp results, when the data set can be divided clearly.

In Section 2 we state the principle of the hyperbolic fuzzy k -means and the classification function. In Section 3 we provide the classification algorithm, which can be achieved by neural networks. A numerical example is demonstrated in Section 4.

2 Classification function

Let $\{x_1, x_2, \dots, x_N\}$ be a data set, where x_n is p -dimensional. Assume that the number of clusters k is given. Let u_{jn} and v_j denote the grade of x_n belonging to the j -th cluster and the center vector of the j -th cluster respectively. Since our aim is to propose a fuzzy clustering, we assume that $0 \leq u_{jn} \leq 1$, $\sum_{j=1}^k u_{jn} = 1$ and $0 < \sum_{n=1}^N u_{jn} < N$. We define the objective function J by

$$J(U, V) = \sum_{j=1}^k \sum_{n=1}^N u_{jn}^m \|x_n - v_j\|^2, \quad (1)$$

where $U = \{u_{jn} | 1 \leq j \leq k, 1 \leq n \leq N\}$, $V = \{v_j | 1 \leq j \leq k\}$ and m is a given number larger than one. The minimization of $J(U, V)$ is the usual fuzzy k -means. For introducing the hyperbolic fuzzy k -means (HFKM), we reformulate the usual fuzzy k -means as the minimization of $J(U, V)$ under the condition that (U, V) satisfies the equations:

$$u_{jn} = \frac{1/\|x_n - v_j\|^{2/(m-1)}}{\sum_{l=1}^k 1/\|x_n - v_l\|^{2/(m-1)}}, \quad (2)$$

$$v_j = \sum_{n=1}^N u_{jn}^m x_n / \sum_{n=1}^N u_{jn}^m. \quad (3)$$

This minimization problem is the same as the usual fuzzy k -means because the equations (2) and (3) are necessary conditions for that (U, V) attains the minimum of J . The equation (2) can be regarded as the classification function. Our approach is to replace (2) with another classification function in order to extend the hard k -means directly. We formulate the HFKM as follows:

Definition 1. The prototype of HFKM is the minimization of $J(U, V)$ with respect to (U, V) under the condition that (U, V) satisfies the equations (3) and

$$(u_{1n}, \dots, u_{kn}) = U_r(V, x_n), \quad (4)$$

where U_r is a classification function given by Definition 2 and r is a parameter such that $0 \leq r < 2$.

For the definition of U_r we use the notation:

$$\begin{aligned} d_j &= \|x - v_j\|, \quad d_L = \min_{1 \leq j \leq k} d_j, \quad C_{Lj} = \|v_L - v_j\|, \\ A &= \{j \mid d_L > d_j - \frac{r}{2} C_{Lj}, j \neq L\}, \end{aligned} \quad (5)$$

for p -dimensional vectors x, v_1, \dots, v_k . For given x , the set A consists of indices j for which the membership values u_j will be nonzero, except the index $j = L$ for which u_L is always nonzero. Now we define the classification function for x and $V = \{v_j | 1 \leq j \leq k\}$.

Definition 2. $(u_1, \dots, u_k) = U_r(V, x)$ is the k -valued function defined as follows:

$$u_j = \begin{cases} 1 & \text{if } \Lambda = \emptyset \text{ and } j = L \\ 0 & \text{if } \Lambda = \emptyset \text{ and } j \neq L \\ \delta_j^{2/(m-1)} / \sum_{i=1}^k \delta_i^{2/(m-1)} & \text{if } \Lambda \neq \emptyset, \end{cases} \quad (6)$$

where

$$\delta_j = \begin{cases} 1 & \text{if } j \notin \Lambda \text{ and } j = L \\ 0 & \text{if } j \notin \Lambda \text{ and } j \neq L \\ \frac{\sqrt{\frac{r^2}{4-r^2} \left(\beta_j^2 + \left(1 - \frac{r^2}{4}\right) \frac{C_{Lj}^2}{4} \right)} + \alpha_j}{\sqrt{\frac{r^2}{4-r^2} \left(\beta_j^2 + \left(1 - \frac{r^2}{4}\right) \frac{C_{Lj}^2}{4} \right)} - \alpha_j} \left(\stackrel{\text{say}}{=} \frac{g_{j1}}{g_{j2}} \right) & \text{if } j \in \Lambda, \end{cases} \quad (7)$$

and, for $j \in \Lambda$,

$$\alpha_j = \begin{cases} \sqrt{d_L^2 - \beta_j^2} - \frac{C_{Lj}}{2} & \text{if } d_L^2 + C_{Lj}^2 \geq d_j^2 \\ -\sqrt{d_L^2 - \beta_j^2} - \frac{C_{Lj}}{2} & \text{if } d_L^2 + C_{Lj}^2 < d_j^2, \end{cases} \quad (8)$$

$$\beta_j = \frac{\sqrt{s(s-2C_{Lj})(s-2d_j)(s-2d_L)}}{2C_{Lj}}, \quad s = d_L + d_j + C_{Lj}. \quad (9)$$

For the case $k = 2$ we illustrate the meaning of the above definition by Figure 1, which shows a plane passing through the data point x and two cluster centers v_1 and v_2 . In Figure 1 it is assumed that x is closer to v_1 and then $L = 1$. Thus Λ is equal to \emptyset or $\{2\}$.

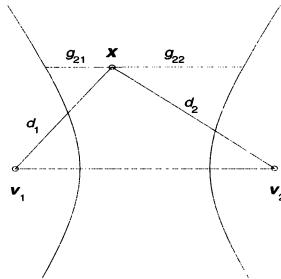


Fig. 1. Centers of clusters and hyperbolae.

From the definition (5) it can be proved that $\Lambda = \emptyset$ means that x exists in the area of the left side of the left hyperbola. The above definition implies that this area is the crisp region belonging to the left cluster. The width of this area depends on the parameter r . When r tends to 2, this area becomes narrow. The area between two hyperbolae is the fuzzy region. In this fuzzy region $j = 2$ is the only element of Λ . The membership values u_1 and u_2 are determined by g_{21} and g_{22} . Note that β_2 is the height of the triangle.

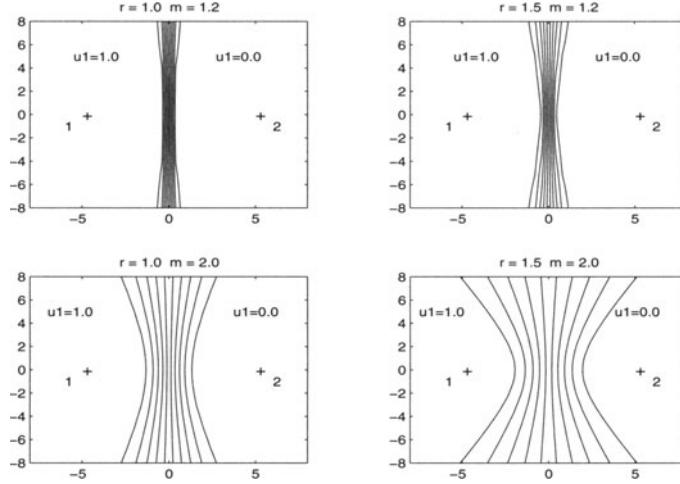


Fig. 2. Contour plots of membership values (2 clusters).

It is easily found that the equations for u_1 and u_2 are the same as those used in the usual fuzzy k -means, except that g_{21} and g_{22} are used instead of d_1 and d_2 . Graphs in Figure 2 are contour plots of u_1 for $r = 1.0, 1.5$ and $m = 1.2, 2.0$. Centers of clusters are shown by '+'.

When $k > 2$, u_j is determined by using the ratio g_{j1}/g_{j2} . As an example, Figure 3 shows some graphs of u_1 and u_5 for the case $k = 5$, where $r = 1.0, 1.5$ and m is fixed to 2.0.

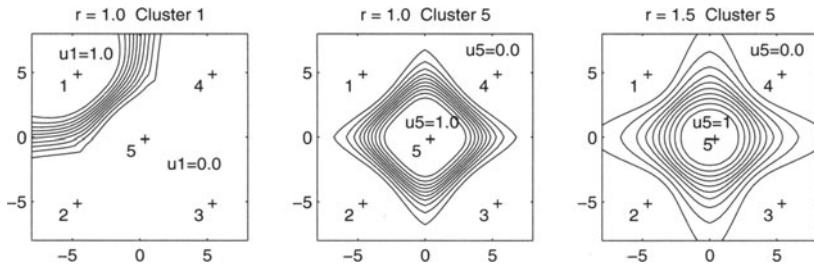


Fig. 3. Contour plots of membership values (5 clusters).

These figures illustrate that U_r produces "fuzzy Voronoi diagrams" which have fuzzy boundaries. The degree of fuzzification depends on the parameters r and m . In many cases m may be fixed to 2.0.

The special case $r = 0$ implies $u_L = 1$, since Λ is always empty. Thus the HFKM with $r = 0$ is identical to the hard k -means.

3 Classification algorithm

The well used recursive procedure of the hard k -means or the usual fuzzy k -means consists of the equation (3) and the classification function. The equation (2) is the classification function for the usual fuzzy k -means. We also consider a similar recursive procedure.

From (3) we have

$$\sum_{n=1}^N u_{jn}^m(x_n - v_j) = 0, \quad (10)$$

which is the same as the equation obtained by setting the partial derivative $\partial J / \partial v_j$ to be zero. In this paper we use the equation (10) instead of (3) and propose the HFKM algorithm.

Algorithm 1 (HFKM).

Initialize v_j for $1 \leq j \leq k$

REPEAT

 Calculate u_{jn} by Definition 2 ($1 \leq j \leq k$, $1 \leq n \leq N$)

 Renew v_j to $v_j + \gamma \sum_{n=1}^N u_{jn}^m(x_n - v_j)$ ($1 \leq j \leq k$)

UNTIL{ Some stop condition is satisfied }

A learning coefficient γ has to be given appropriately. This algorithm can be regarded as a batch mode. A sequential mode algorithm is also available.

Algorithm 2 (HFKM-NN).

Initialize v_j for $1 \leq j \leq k$

REPEAT

 FOR $1 \leq n \leq N$

 Calculate u_{jn} by Definition 2 ($1 \leq j \leq k$)

 Renew v_j to $v_j + \gamma u_{jn}^m(x_n - v_j)$ ($1 \leq j \leq k$)

 END FOR

UNTIL{ Some stop condition is satisfied }

Algorithm 2 is achieved by the neural networks (NN) which have the same structure as the competitive learning neural network by Rumelhart et al. (1986). That is, sizes of the input layer and the output layer are p and k respectively, and weights are elements of v_j . Note that, though Rumerhart et al. (1986) refer to the classification of binary data only, their method can be applied to general data. See also Kohonen(1995) for the application of neural networks to clustering.

Similarly to the hard k -means and the usual fuzzy k -means, Algorithms 1 and 2 are based on the necessary conditions. However, the definition of the classification function U_r is not based on the partial derivative $\partial J / \partial u_j$ unlike the usual fuzzy k -means. Then their convergence property is not clear, though our experiences suggest they will converge. Even if they converge, such algorithms have dependency on initial values. Thus a modified version should be used in practice. One way is to combine a kind of global search.

4 Example

We consider data x_1, \dots, x_{17} shown by circles in Figure 4, where $p = 2$. Algorithm 1 is applied with $k = 2$. Figure 4 shows the values of the membership function of the left cluster for $m = 2.0$ and $r = 1.0, 1.5, 1.8$.

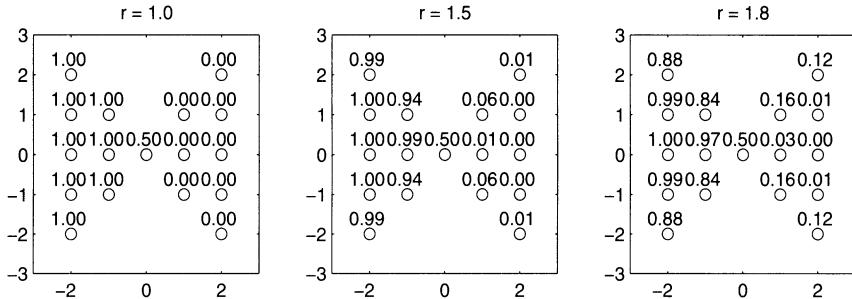


Fig. 4. Results of hyperbolic fuzzy k -means.

When $r = 1.0$, a crisp result is obtained except for the central point. Since data is symmetric, this result is quite natural. Note that such a result can not be obtained by the hard k -means nor the usual fuzzy k -means.

5 Concluding remarks

As the above example shows, the proposed method is more feasible than the hard k -means or the usual fuzzy k -means.

In Definition 2, g_{j1} and g_{j2} can be replaced by other distances between x and hyperbolas. However, we use not g_{j1} and g_{j2} directly but the ratio δ_j . Thus the usage of g_{j1} and g_{j2} does not cause a big influence usually.

In this paper we have considered the distance given by the Euclidean norm $\|\cdot\|$. But a generalized norm $\|\cdot\|_G$ can be used instead of $\|\cdot\|$. In this case the equation (10) and algorithms have to be modified.

To apply the hyperbolic fuzzy k -means, the parameters r and m have to be given previously. Though the appropriate value of r should depend on the situation, automatic selection of r , where m is fixed, is a problem to be studied. Moreover theoretical study on convergence of algorithms is also required.

References

- BEZDEK, J.C. (1981): *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.
- KOHONEN, T. (1995): *Self-Organizing Maps*, Springer.
- RUMELHART, D.E. and et al. (1986) : *Parallel Distributed Processing*, The MIT Press, Massachusetts.

A Toolkit for Development of the Domain-Oriented Dictionaries for Structuring Document Flows^{*}

Pavel P. Makagonov¹, Mikhail A. Alexandrov² and Konstantin Sboychakov¹

¹ Moscow Mayor's Directorate, Moscow City Government,
Novi Arbat 36, Moscow, 121205, Russia. Email: ummpp@mariia3.munic.msk.su

² Center for Computing Research (CIC), National Polytechnic Institute (IPN),
Av. Juan de Dios Batiz, C.P. 07738, Mexico DF.
Email: dyner@pollux.cic.ipn.mx

Abstract. An approach to thematic document classification, clusterization and investigation of document flows and collections based on domain-oriented dictionaries (DODs) is considered. It is simple enough to be used by, say, a secretary that frequently needs to classify and search large amounts of documents. However, for good results, such an approach requires a solid technology for construction and maintenance of the DODs; this task is to be performed by experts or advanced users. A DOD represents a specific subject topic and is constructed on the basis of the analysis of a collection of documents representing this topic, selected by a group of experts. The toolkit facilitates the development of a hierarchical system of DODs by the application of a set of heuristic criteria for the selection of the keywords from such a document collection representing one subject domain. In the paper, the application of the DODs developed with the toolkit for information retrieval is illustrated with examples.

1 Introduction

Many of the current problems of the information analysis, such as classification, clusterization, retrieval, etc., have a complex multi-aspect interdisciplinary nature, with economic, ecological, humanitarian, and other aspects being mixed and interdependent. In many cases, traditional approaches to the information analysis do not give satisfactory results.

Our approach to these tasks is based on the application of the traditional techniques to so called document images rather than the raw texts of the documents, the multidisciplinary information being coded into such images. Namely, a document image relative to a specific domain *Dom* is the set of the words characteristic for this domain that occur in the document, with their respective numbers of occurrences in the document. Thus, each document has different images, one for each domain. The information on what words or word combinations are characteristic for a specific domain is encoded in so-called domain-oriented dictionary (DOD). Thus, crucial for our approach

* Work done under partial support of CONACyT, Mexico.

is the quality of the system of the DODs used to distinguish the domains. A solid technology is required for their construction and maintenance.

In this paper, we describe the technology for the construction of such a dictionary by means of the statistical analysis of a set of documents representing a specific domain. The toolkit that we describe facilitates both the selection process and the construction of the dictionary basing on the selected documents. Thus, construction of a DOD for a given large domain consists of four major steps: construction of the general lexicon statistical word list, selection of the documents, the construction of the DOD proper by their analysis, and the construction of sub-DODs that represent subdomains of the given domain. Note that the sub-DODs, in turn, can be subdivided in sub-sub-DODs, which gives a hierarchy of subject topic domains with different degrees of detail or generalization.

Our approach to the selection of the documents and the construction of the DOD is similar to (Bolso, 1998; Lelu, 1998; Takakura, 1998); however, in these works some important criteria, such as the application of the Gini coefficient described below, are not considered. Also, these works do not consider the task of the subdivision of the general DOD into sub-DODs.

2 Construction of the general lexicon

For the development of the DODs, some statistical information on the general lexicon mixture of the given language is required. The general lexicon mixture is a collection of a very large number of texts in different domains, such as ordinary newspaper articles. The dictionary that characterizes the statistics of the general lexicon is a list of all words used in the general text corpus with their respective frequencies. There are two issues to be paid attention to.

1. *Morphological normalization.* Due to the morphological inflection and regular word formation (*correct*, *corrects*, *correcting*, *corrected*, *corrector*, *correctly*, etc.) all inflected word forms in all texts are reduced to a single representation (*correct*-).
2. *Estimation of the necessary corpus size.* The size of the text corpus sufficient for reliable (representative) statistics can be estimated using the notion of the statistical rank of a word.

Let L be a list of words W with their respective frequencies $F_T(W)$ in some text T . For a word $W \in L$, its rank $rank_L(W)$ is defined as its position in L ordered by the frequency, i.e., $rank_L(W) = |\{x \in L : F_T(x) \geq F_T(W)\}|$, where $|s|$ is the number of elements in the set s .

According to the well-known empirical Zipf law, in any natural language text T , for each word W , the frequency $F = F_T(W)$ is approximately related with its rank $R = rank_T(W)$ by the formula $\log F \approx A - K \times \log R$ for some appropriate language-dependent constants A and K . For example, for the

general lexicon of Spanish $K \approx 1.0$ (determined by a sampling of 100,000 running words). For the general lexicon of Russian, $K \approx 0.866$ (determined by a sampling of 22,000,000 running words); the values were determined with the method of least squares. These constants are approximately the same for any large text sampling within the given language.

On the other hand, for a small text the distribution of the frequencies generally differs from the Zipf law. Thus, the agreement with the Zipf law serves as a good estimation of the statistical representativity of a text corpus. In other words, a text corpus T is large enough if the frequencies $F_T(W_i)$ of its words in average obey well enough the Zipf law.

3 Selection of the domain documents

For the construction of the DOD, first of all a specific domain is selected and fixed, for example, the domain of *Computer Sciences*, or that of *Telecommunications*. Then the documents representing the chosen domain are selected by a group of specialists in the area. As a result, a collection of documents representing the domain is compiled.

The quality of the resulting collection is estimated by a measure of the consistency of their vocabulary. This measure (not described here in detail) reflects the share of their common vocabulary relative to their total one.

4 Construction of the DOD

For a given domain its DOD reflects the vocabulary of this domain and distinguishes the given domain from the other ones. Given a domain¹ Dom , the process of the development of the corresponding DOD begins with the compilation of a list $L(Dom)$ of the words used in the documents of Dom , with two frequencies for each word: its frequency in these texts and in the texts of the general mixture Com . The list is compiled using the morphological normalization procedure described in the section 2. Then each word W from the list $L(Dom)$ is tested on the possibility of its inclusion in the DOD. This test consists of the following three criteria.

Criterion 1. Only those words W are included in the DOD for which $F_{Dom}(W) \gg F_{Com}(W)$, namely, $F_{Dom}(W) > k \times F_{Com}(W)$. The coefficient k is determined after additional investigation. Its value is related with the statistical estimation of the mean error in the measuring of the frequencies due to a limited size of the sample text. A good default value is 2. Here $F_{Dom}(W)$ and $F_{Com}(W)$ are the frequencies of the word W in the domain texts and in the general mixture texts, respectively.

¹ A domain is represented with a set of texts devoted to a certain topic. We do not need, though, to distinguish individual texts and thus effectively consider Dom as a single large text obtained as their concatenation.

To formulate the second criterion, we will need the notion of the Gini index $G_T(W)$ for a given word W relative to a set of texts $\mathbf{T} = \{T_i\}$. Let n_i be the number of occurrences of the word W in the text T_i . Let us assume that n_i are arranged in ascending order; otherwise the texts are to be re-numbered correspondingly. For each i , let N_i be the accumulated number of occurrences: $N_i = \sum_{j \leq i} n_j$. Obviously, for a uniform distribution of the numbers of occurrences $n_i = \text{const}$, N_i represent a straight line: $N_i = \text{const} \times i$. In all other cases, $N_i \leq \text{const} \times i$. The Gini index $G(W) = G_T(W)$ for a given word W relative to a set of texts \mathbf{T} is then determined as the relative difference between the area under the chart for N_i and under the uniform chart (straight line) $\text{const} \times i$. Namely, $G(W) = 1 - \frac{1}{S} \sum_i N_i$, where S is the area under the uniform chart $\text{const} \times i$, $S = \frac{1}{2}(|\mathbf{T}| + 1) \sum_i n_i$, where $|\mathbf{T}|$ is the total number of the texts and $\sum_i n_i$ is the total number of occurrences of the word W in all texts of \mathbf{T} .

Criterion 2. Only those words W are included in the DOD for which the Gini index is between two fixed thresholds, low G_L and high G_H : $G_L < G(W) < G_H$. Their values are fixed empirically by an expert, good default values being 0.8 and 1, respectively.

The application of this criterion requires the number of values $n_i \neq 0$ (i.e., the number of the texts in \mathbf{T} that contain the word W) not to be too small. On the other hand, if the DOD being constructed is to be later subdivided into sub-DODs (see section 5 below), then the word W should not occur in all (or too many) of the texts of \mathbf{T} . Namely, to obtain not less than 5 to 10 sub-DODs, each word should occur in not more than approximately 10% to 20% of the texts. Thus, for interesting DODs, the following criterion holds:

Criterion 3. Only those words W are included in the DOD for which the number N of texts in which they occur, $N = |\{i : n_i \neq 0\}|$, is between two fixed thresholds: $N_L < N < N_H$.

If the expert estimates the number of words in the resulting DOD as too large, the parameters of the three criteria should be adjusted to be more restrictive and the whole procedure repeated.

5 Construction of sub-DODs

For more detailed classification of the documents, it should be investigated whether the given domain can be divided into subdomains reflecting various aspects of the main topic of the domain. For example, the domain of *chemistry* can be subdivided into *organic chemistry* and *inorganic chemistry*.

For this, a matrix $M = \{n_{ij}\}$ is considered whose rows correspond to the texts and columns to the words of the DOD. The element n_{ij} is the number of occurrences of the word $W_j \in \text{DOD}$ in the text $T_i \in \mathbf{T}$. The texts and words in the DOD and \mathbf{T} , correspondingly, can be re-numbered so that the matrix will become a quasi-block matrix. By a quasi-block matrix, we mean a matrix that has several rectangular areas (blocks), possibly overlapping, such that

for each row and each column, the majority of the values inside these blocks are significantly greater than the majority of the values in this row or column outside these blocks. Such re-numbering is achieved by a clustering algorithm. The blocks themselves are clusters; their projections on the word axis are sub-DODs reflecting sub-domains of the whole domain, and their projections on the text axis are the text sets characterizing these sub-domains. For example, a matrix for the domain of *chemistry* might be decomposed in two blocks, one of them corresponding to the words and texts on *organic chemistry* and the other to the words and texts on *inorganic chemistry*.

However, if the matrix is filled rather uniformly and can not be decomposed in any reasonable blocks, then no sub-domains can be found in the given domain using the given set of texts \mathbf{T} .

The program *Visual Heuristic Cluster Analysys* (Makagonov, 1998) uses this algorithm for working with a large database "Sustainable Cities of Russia" in Moscow City Government.

6 Example

Constructing the DODs for the International Conference A-PORS'97, Melbourne, Australia, 30 November to 4 December 1997. The program *Text Recognizer* was used to classify 20 abstracts presented at the Conference to the DODs that correspond to the sessions denoted as Mc, Md, Ta, Tb, Tc, Wa in the Conference program. First, 6 DODs were constructed, one for each of the session topics, by training on the abstracts that were included by the Program Committee in the corresponding sessions.

In Fig. 1, the rows of the matrix are indexed by the section topics shown as numbers to the left of the matrix and as abbreviated names to the right of it, and the columns of the matrix are indexed with the numbers of the abstracts. The matrix is clustered, with the clusters representing the document groups corresponding to each topic. These groups are indicated in the upper part of the figure, e.g., the group corresponding to Tc consists of the abstracts number 13 to 17 (after re-numbering during the clustering process). The values of the frequencies of the words are shown with the colors of the matrix cells: the higher the frequency the darker the color.

The results show that the topics of the sections are rather well separated, the abstracts are in rather good accordance with the corresponding topics, and, what is most important for us, the DODs built with our toolkit correspond well to the topics. What is more, Fig. 1 shows that the Program Committee incorrectly classified some abstracts, namely, 2, 8, 12, 13, 14, 17, and 20. For example, the Program Committee should have assigned the abstract number 20 to the session Md instead of Wa . Also, Fig. 1 shows that some of the abstract reflect several topics simultaneously, e.g., 13, 14, and 17. This is an example of discovering new knowledge with our methods.

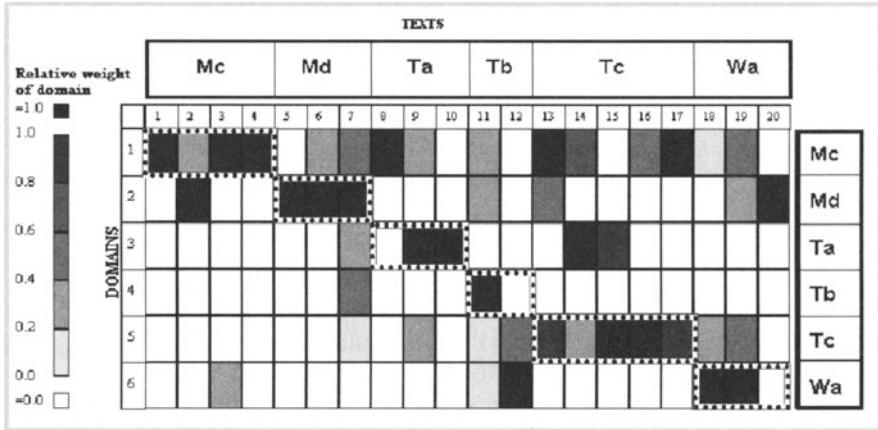


Fig. 1. Concordance of abstracts and DODs inside sessions.

7 Conclusions

The problem of the clustering of large sets of text documents can be solved with DODs even when neither the contents nor the number of the subtopics are known beforehand. The criteria were discussed for selection of the words for the inclusion in the DOD. An algorithm for subdivision of a general DOD into more specific sub-DODs was presented. This algorithm is based on the clustering procedure applied to the matrix of the occurrences of the words in the texts. The described technology has been implemented as software toolkit for experts and end users.

References

- BOLSO, S., and A. MORRONE. (1998): A frequency dictionary of polyforms as a linguistic data base for text disambiguation in TALTAC, In: *Data Science, Classification and Related Methods (Proc. of 6-th Intern. Conf. IFCS, Rome, Italy, 1998)*. Rome, 32-35.
- LELU, A., and S. FERHAN. (1998): Clustering a textual data-flow by incremental density-modes seeking. In: *Data Science, Classification and Related Methods (Proceedings of 6-th Intern. Conf. IFCS, Rome, Italy, 1998)*. Rome, 206-209.
- MAKAGONOV, P., and K. SBOYCHAKOV. (1998): Man-machine methods for solution of weakly-formalized problems in humanitarian and natural fields of knowledge (visual heuristic cluster analysis). In: Pedro Galicia (Ed): *Proceedings of International Computer Symposium CIC'98 (Mexico, 1998)*. National Polytechnic Institute, Mexico, 346-358.
- TAKAKURA, S. (1998): Study of same methods of analysis of textual data in Japanese. In: *Data Science, Classification and Related Methods (Proceedings of 6-th Intern. Conf. IFCS, Rome, Italy, 1998)*. Rome, 297-298. RENV

Classification of Single Malt Whiskies

David Wishart

Department of Management, University of St. Andrews,
St. Katharine's West, The Scores, St. Andrews KY16 9AL, Scotland
(e-mail: d.wishart@st-andrews.ac.uk website: www.clustan.com)

Abstract. Tasting notes in 10 recently published books on malt whisky were coded and analysed for 84 single malt whiskies. Over 400 aromatic and taste descriptors were identified and grouped into 12 sensory features, from which a synonymy of the whisky literature was developed. The 84 malt whiskies were then clustered into 10 groups using the FocalPoint clustering method in ClustanGraphics. An industry survey to validate the classification is described, and applications in product design, brand management and marketing are discussed. A tutored tasting of selected single malt whiskies follows the technical presentation.

1 Data sources

The genesis of this paper was to use ClustanGraphics (Wishart (1999)) to develop a classification of malt whiskies as a worked example of market segmentation, following Lapointe and Legendre (1994) who analysed descriptions of the colour, nose, body, palate and finish of 109 single malt whiskies reviewed by Jackson (1989). Rather than base our classification on Jackson alone, we sought to reduce the degree of personal subjectivity by reviewing ten books on malt whisky currently available (Arthur (1997), Broom (1998), Jackson (1995), Lerner (1997), MacLean (1997), Milroy (1995), Murray (1997), Nown (1997), Shaw (1997) and Tucek and Lamond (1997)). Tasting notes published by the distilleries were also reviewed, where available.

2 Benchmark whiskies and coverage

Most distilleries produce several brands that are differentiated by length of time in cask, special conditioning or finishing, e.g. to impart flavours such as oak, sherry, port or Madeira to the whisky. As our objective was to develop a classification of malts that are readily available to consumers, we felt that we should select a benchmark malt whisky from each distillery. We firstly excluded rare malts and any premium brands that are specially aged, cask conditioned or finished. We also decided not to cover distilleries that had been demolished or are not currently in production.

Not all of our 10 authors reviewed the same distillation from each distillery, as some limit their tasting notes to house style only (e.g. Milroy (1995)). Where more than one distillation is produced we selected the most

widely available brand, usually of 10-15 years maturation in cask. New distilleries that currently offer young malts (Arran and Drumguish) were included for future reference, as they evolve. Vatted malts (blends of pure malts), and malt whiskies produced in Ireland, Japan, New Zealand and Wales were excluded. We thus arrived at 84 single malt whiskies of around 10-15 years maturation, most of which are widely available in the U.K.

3 Standardised flavour profile

In developing a flavour profile, our objective was to standardise wide and open-ended terminology ranging from the precise and analytical to highly imaginative, idiosyncratic even eccentric descriptions. In essence, we used the 10 books as an expert tasting panel and developed a common vocabulary and method of consensus scoring between them.

Every adjective used to describe the taste and smell of a malt whisky was noted. A vocabulary of over 400 aromatic and taste descriptors was thereby compiled from the tasting notes in the 10 books reviewed (see Wishart (2000)). These were grouped into 12 broad aromatic features: Body (Light/Heavy), Sweetness (Dry/Sweet), Smoky (Peaty), Medicinal (Salty), Feinty (Sulphury), Honey (Vanilla), Spicy (Woody), Winey (Sherry), Nutty (Oaky/Creamy), Malty (Cerealy), Fruity (Estery) and Floral (Herbal).

Because we did not have tasting notes by all ten authors on every malt, a system of consensus coding was used which involved an element of judgement. For each malt, a data sheet was compiled with all the tasting notes reviewed, and the strength of each aromatic feature was judged according to the consensus of the 10 expert authors' opinions, on a 5-point scale as follows: (1) Not Present, (2) Slight Hint, (3) Medium Note, (4) Definite Note, (5) Pronounced Feature. The data sheets and coding were then sent to each distillery, to check that they were accurate and complete.

4 Preliminary classification

Cluster analysis groups malts into the same cluster when they have broadly the same flavour characteristics across all 12 sensory features. Our 84 single malts were first classified by Ward's method, which optimises the Euclidean Sum of Squares within clusters by hierarchical cluster analysis. This forms clusters which appear spherical in shape with the members all close to the cluster mean and, at the same time, maximizes the differentiation between clusters. We were therefore confident that all the whiskies within a cluster would be broadly similar in terms of their flavour profiles as coded.

Our preliminary classification was sent to the distillers covered in our study, and to about 50 whisky experts including the authors of the books reviewed, malt whisky societies, independent bottlers, blenders, retailers and

academic researchers. Nearly a third replied, indicating a fair degree of interest within the whisky industry. Of those who replied, over 90% thought the preliminary classification was reasonable overall, and this was very encouraging.

However, when asked whether any should be in different clusters, 32 whiskies were noted as possibly needing review. While this was somewhat disappointing, most of those who responded had attempted to validate our preliminary classification, and this encouraged us to refine it further. We also received valuable expert information as to where the preliminary classification might be defective or inefficient.

Our preliminary classification was next presented at six conferences and seminars, including the Scotch Malt Whisky Society who published it in their 1998 Newsletter. Useful advice from statistical and whisky experts was thus received and incorporated into the further analysis. A number of distilleries and whisky retailers were visited to discuss the classification, and meetings were held with master blenders and brand heritage managers.

5 Final survey and consultation

In the light of the comments received, we revised the data and again used hierarchical cluster analysis by Ward's method. However, we also used a new k-means clustering procedure to check whether a better cluster solution could be obtained from exhaustive random trials on a number of different starting classifications (Wishart (2000)). For this study, we tested 200,000 different random trials for each of 8-clusters to 12-clusters. The best solution in each analysis was then examined in relation to the expert industry advice we had received on our preliminary analysis.

The cluster solution that seemed to accord most closely with the results from our first industry survey was at the 10-cluster level. It was the most frequent final classification in 200,000 random trials, occurring in 2% of the trials, and had the least overall Euclidean Sum of Squares. Our final single malt whisky classification is as follows:

Cluster A (*Full-Bodied, Medium-Sweet, Pronounced Sherry with Fruity, Spicy, Malty Notes and Nutty, Smoky Hints*): Dailuaine, Dalmore, Glendronach, Macallan, Mortlach, Royal Lochnagar; **Cluster B** (*Medium-Bodied, Medium-Sweet, with Nutty, Malty, Floral, Honey and Fruity Notes*): Aberfeldy, Aberlour, Ben Nevis, Benrinnes, Benromach, Blair Athol, Cragganmore, Edradour, Glenfarclas, Glenturret, Knockando, Longmorn, Scapa, Strathisla; **Cluster C** (*Medium-Bodied, Medium-Sweet, with Fruity, Floral, Honey, Malty Notes and Spicy Hints*): Balvenie, Benriach, Dalwhinnie, Glendullan, Glen Elgin, Glenlivet, Glen Ord, Linkwood, Royal Brackla; **Cluster D** (*Light, Medium-Sweet, Low or No Peat, with Fruity, Floral, Malty Notes and Nutty Hints*): An Cnoc, Auchentoshan, Aultmore, Cardhu, Drumguish, Gengoyne, Glen Grant, Mannochmore, Tamdhu, Tobermory; **Cluster E** (*Light,*

Medium-Sweet, Low Peat, with Floral, Malty Notes and Fruity, Spicy, Honey Hints): Bunnahabhain, Glenallachie, Glenkinchie, Glenlossie, Glen Moray, Inchgower, Inchmurrin, Tomintoul; Cluster F (Medium-Bodied, Medium-Sweet, Low Peat, Malty Notes and Sherry, Honey, Spicy Hints): Ardmore, Bushmills, Deanston, Glen Deveron, Glen Keith, Glen Rothes, Old Fettercairn, Singleton, Tomatin, Tormore, Tullibardine; Cluster G (Medium-Bodied, Sweet, Low Peat and Floral Notes): Arran, Dufftown, Glenfiddich, Glen Spey, Miltonduff, Old Rhosdhu, Speyburn; Cluster H (Medium-Bodied, Medium-Sweet, with Smoky, Fruity, Spicy Notes and Floral, Nutty Hints): Balblair, Craigellachie, Glen Garioch, Glenmorangie, Oban, Old Pulteney, Teaninch; Cluster I (Medium-Light, Dry, with Smoky, Spicy, Honey Notes and Nutty, Floral Hints): Bowmore, Bruichladdich, Highland Park, Isle, of, Jura, Ledaig, Springbank; Cluster J (Full-Bodied, Dry, Pungent, Peaty and Medicinal, with Spicy, Feinty Notes): Ardbeg, Caol Ila, Clynelish, Lagavulin, Laphroaig, Talisker.

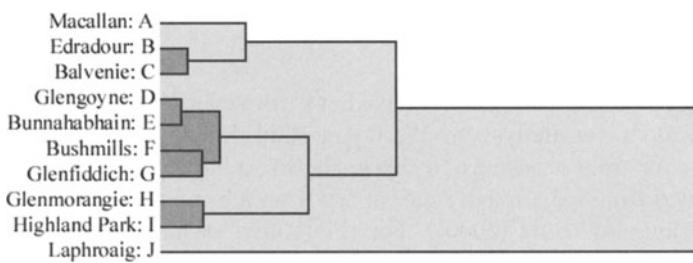


Fig. 1. Final 10 clusters from the classification of 84 single malt whiskies, optimally ordered and illustrated by reference exemplars, with the 5-cluster partition shaded.

The order of the 10 clusters A-J maximizes the row-wise rank correlation of the underlying proximity matrix (Wishart (1999)). Readers who are familiar with malt whiskies may recognise the two extremes of strongly sherried malts (cluster A) and the heavily peated, mainly Islay malts (cluster J). Adjacent to these polar benchmarks are the lightly sherried (clusters B and C) and lightly peated (clusters H and I) malts, with the light-bodied, floral and malty clusters, including four largely unpeated groups (clusters D-G) falling in the middle.

A ClustanGraphics tree that summarises the final classification of the 10 whisky types is shown in Figure 1, in which the last 5 clusters are shaded differentially. This was obtained using Ward's method with the 10-cluster FocalPoint model. The cluster of pungent, peaty Islay malts (J) is most distinctive, being maintained as a separate group to the end of the analysis, with clusters A-C the next most distinctive, followed by clusters H-I.

6 Industry applications

The classification may be of use to retailers and consumers. For example, if you like a particular malt then those in the same cluster should be of interest to your palate. It could be of use to distillers, to describe new distillations or to design new malts for marketing purposes.

The cluster analysis also exemplifies the clusters. This involves finding a reference malt whisky for each cluster, or the whisky which is most typical of its group. It can be useful in marketing, for stratified statistical sampling or for selecting a representative sampling panel. For example, when organising a whisky tasting it is helpful to choose a malt from each cluster to illustrate the range.

In compiling a vocabulary of over 400 aromatic and taste descriptors, grouping them into 12 sensory features, and analysing the resulting classification in these terms, we have developed a standard synonymy of the whisky vocabulary which may now be used to describe malt whiskies more consistently. This framework may be of value to the whisky industry, but it clearly needs further development and evaluation.

Understanding whisky consumers, and what they look for in their favourite brands, is a marketing prerequisite for developing whisky products with enhanced consumer appeal. John McGrath, Diageo Group Chief Executive, puts it simply: "You have to know a brand to grow a brand". Brand analysis, or defining product "types" by customer perceptions about them, reveals a brand's "footprint", or its position relative to its competitors. This type of modelling is valuable for brand management and competitor analysis. Clustering supermarket products by linked purchasing patterns can also be used to plan store layouts to maximise impulse purchasing.

7 Final industry survey

Our revised classification was again sent to the distillers and whisky experts in February 2000. Of those who replied, over 90% thought that the classification was reasonable overall, that the standard flavour profile was a reasonable description of flavours, and that the classification would be a useful guide to consumers. A number of whiskies were again identified as being in the wrong clusters, six being noted by two or more respondents. The data were re-examined in the light of these comments, and the final classification was further adjusted to take account of most of these apparent misclassifications.

References

- ARTHUR, H. (1997), *The Single Malt Whisky Companion: A Connoisseur's Guide*, Apple Press, London.
- BROOM, D. (1998), *Whisky: A Connoisseur's Guide*, Carlton, London.

- JACKSON, M. (1995), *Malt Whisky Companion*, Dorling Kindersley, London. (Also published in 1989).
- LAPOINTE, F-J., and LEGENDRE, P. (1994), A Classification of Pure Malt Scotch Whiskies, *Applied Statistics*, 43, 237–257.
- LERNER, D. (1997), *Single Malt and Scotch Whiskies*, Könemann, Köln.
- MACLEAN, C. (1997), *Malt Whisky*, Mitchell Beazley, London.
- MILROY, W. (1995), *The Malt Whisky Almanac*, Neil Wilson Publishing, Glasgow.
- MURRAY, J. (1997), *The Complete Guide to Whisky*, Carlton, London.
- NOWN, G. (1997), *Malt Whisky: A Comprehensive Guide for both Novice and Connoisseur*, Salamander, London.
- SHAW, C. P. (1997), *Classic Malts*, HarperCollins, Glasgow.
- TUCEK, R., and LAMOND, J. (1997), *The Malt Whisky File*, Canongate Books, Edinburgh.
- WISHART, D. (1988), Using hierarchical classification in information retrieval and diagnosis. In H.-H. Boch (Ed): *Classification and related methods of data analysis*, Elsevier Science, 1988, 717–724.
- WISHART, D. (1999): ClustanGraphics3: Interactive Graphics for Cluster Analysis. In: W. Gaul, and H. Locarek-Junge (Eds.): *Classification in the Information Age*, Springer, Heidelberg, 268–275. See, also: www.clustan.com.
- WISHART, D. (2000), Two-stage k-means clustering method with outlier detection and model calibration. Submitted to G. Ritter (Ed): *Classification, Automation, and New Media*, Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation, University of Passau, Springer, Heidelberg (under review).
- WISHART, D. (2000): *Malt Whisky Classified*, Pavilion Press, London (forthcoming, also at www.WhiskyClassified.com).

Robust Approach in Hierarchical Clustering: Application to the Sectorisation of an Oil Field

Jean-Paul Valois

ELF AQUITAINE Exploration Production
F 64018 - PAU Cedex
(e-mail: jean-paul.valois@elf-p.fr)

Abstract. Production data of oil fields are provided as decline curves (oil and water production vs time), that the user wants to gather in a limited number of clusters. Preprocessing of data is required to remove noise, and provides a complete data set, involving for each statistical unit (wells) extraction of attributes from smoothed or modelized curves. Hierarchical clustering is performed in two steps to avoid smaller or outlier cluster ; firstly the centroid clustering method is used to recognize and then discard clusters having a lower frequency, this is followed by application of the Ward-method. Finally, using the central part of these previous (Ward) clusters, discriminant analysis is performed, including all the discarded units. This sequence avoids the disturbing influence of outlying units, and also gives the probability for each unit to be classified in the clusters.

1 Introduction

Clustering methods involve a lot of different techniques (Jain and Dubes (1988)). One of the more frequently used is the K-means, which is required when the statistical units are too numerous (more than thousand). This method is sensitive to the initial seeds. Therefore, such programs generally include some precautions for outlying clusters.

Hierarchical clustering does not require any seeds to be defined, and is therefore usually believed as providing more stable results. The displayed tree is a convenient way to choose the number of clusters. Among many algorithms, the centroid and Ward methods seem to be the most frequently used. The first one places the outliers into separated clusters, and thus the result is a maelstrom of relatively small clusters (outliers) and a few major clusters. The Ward method is generally preferred, because it provides more balanced clusters. Nevertheless the influence of outlying units on this algorithm is generally not discussed.

More recent methods of clustering such as the Kohonen algorithm, spatially constrained clustering, or fuzzy clustering, have not yet been applied in this sequence.

2 Available data and industrial problem.

Oil field production data provide, for each well, decline curves. These indicate the oil (Q_o , quantity of oil daily produced) and the water (BSW, percent of water in the total liquid phase) vs time. Mature fields can have more than 500 wells, and the production can have begun 40 years ago or more. The selected example has 179 wells, with production between 1961 and 1987. This is a synthetic case, constituted by selecting wells from a real case ; no gas was produced. As is generally the case, all the wells were not open for identical periods.

The main industrial problem is to describe the data (where are the best wells located, and so on, see Valois (2000) concerning some exploratory aspects of this example). A second set of questions concerns the dividing of fields into compartments ; because companies wish to define more or less promising zones for future production, we will discuss this aspect in the paper. The further objective which is to locate sites where supplementary wells could be drilled, is not concerned here. Nevertheless the results of an efficient compartment definition will be a key step towards future well definition.

3 Transformation of data

Because all the wells are not producing during the same period, the raw data table is incomplete. In addition, the data are noisy. Irregularities occur in oil production, due to mechanical problems, human interventions ("work-over"), or interferences between wells when production is halted and rebegun. Water percentages are influenced by equipment modifications and because water is not a commercial product, it is not always measured.

The first step of data management is to model the Q_o curve. An hyperbolic model is adjusted to the data. The following equation is classically used in the oil industry, because some of the coefficients have a physical meaning, at least in the more simple cases :

$$Q_o = Q_{omax} / (1 + EXPR)^{**a}, \text{ where } EXPR = \text{der}Qo * (\text{date} - \text{firstdate}) / a$$

Q_{omax} is taken from the data, a and $\text{der}Qo$ are given by the modeling procedure.

No simple model exists for the BSW evolution, in particular for complex oil reservoirs. We perform the LOWESS smoothing proposed by Cleveland (1993).

Validation of the Q_o modeling is obtained by comparing the cumulative curve (N_p) including all the wells, using the raw data, with the one using the predicted values. Absolute residuals have a mean of 1.86 percent, 9/10 of these are below 4 percent.

From these modeled or smoothed curves, we extract a set of attributes to solve the problem of the incomplete data set. For oil production, attributes are the Q_{omax} , the Q_o value found 1, 2 or 3 years after the Q_{omax} , the

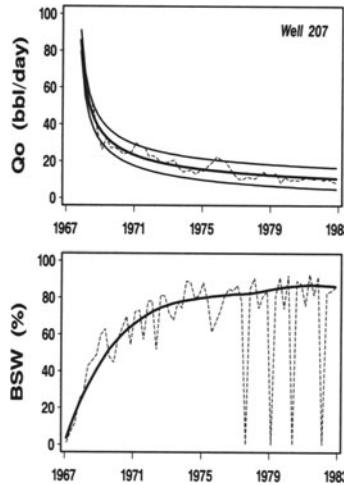


Fig. 1. Modeled (left) or smoothed (right) curves, vs time, dashed line: raw data.

ratio between Qomax and the final Qo, and so on. Some other attributes are indicated by Voineff et al (1996). From the BSW smoothed curve, the attributes are the initial and the final values, the time to obtain different levels such as 1, 10, 25, 50 percent of water. For the wells that did not reach 25 or 50 percent of water, an iterative regression is performed, to complete the missing values.

This yields a complete data set. Moreover the initial data were a 3D incomplete table (wells, measurements, time) and now we obtain a completed 2D table (wells, attributes).

Principal Component Analysis (Harman (1976)) is then performed using this complete data set (units = wells, variables = attributes). The aim of this step is to exclude a part of variance, interpreted as residual noise, the variance retained here is 94 percent. Using the factors scores, the corresponding phenomena can more easily be weighted.

4 First clustering

A first clustering is performed using the factor scores as columns.

A preprocessing step (Jain and Dubes, 1980) is performed, in order to detect and finally discard outlying units or smaller clusters (Bayne et al. (1980)). The Ward method is not suitable for this (Milligan and Schilling (1985)), and here we are using the centroid method. The scatter plot (Figure 2, left) displays the cumulative frequency of the clusters, ranked in increasing size order ; this display shows that the centroid method put 10 percent of

the data set into the 5 smallest clusters. These are practically undesirable and besides could be suspected to be outliers: the end user is then likely to arbitrarily reinclude them in one of the major clusters.

5 Second clustering

The clustering is then recomputed excluding these 10 percent of the data set; the Ward algorithm is performed here, because we want to obtain balanced major clusters and reduce the square-error, Figure 2 (right) displays the obtained tree. The size (1, 2, 3 points...) of the left side column indicates the location of each unit in the Ward run using the complete data set (1 point = cluster 1 and so on). Tone exhibits the agreement (grey) or disagreement (black) between both the Ward runs (using the outliers or not).

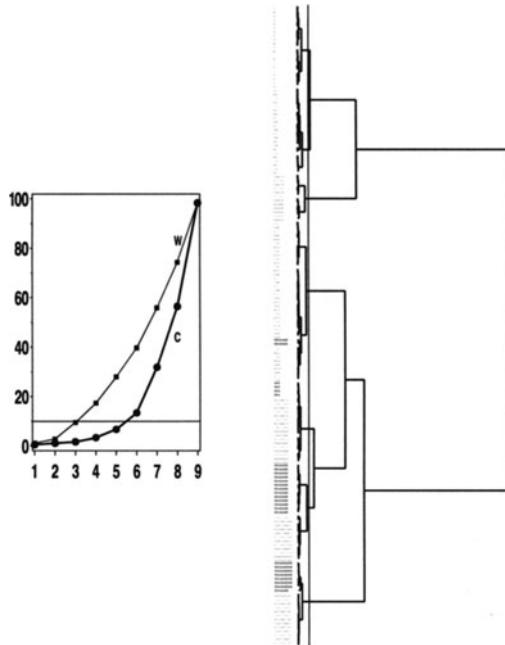


Fig. 2. Cumulative frequency of the clusters (left), tree provided by Ward method using 90 percent of the data set (right).

This obtained tree is here more suitable, because each cluster includes at least 12 units. But it can be outlined at this point that some differences are found between both the Ward runs (using the outliers or not, black items in Figure 2).

6 Discriminant Analysis

The preceding step of Ward hierarchical clustering did not use all the wells, and we now need to reposition the discarded wells in the classification. Discriminant analysis provides a well defined method to do this. It requires many units for definition of the clusters.

To define these, we take into account the central part of the clusters given by the last Ward run. To do this, we compute the geometrical center of each cluster, and then the euclidian distances from each unit to the center of the cluster it is belonging to. Distances are standardized, and the wells closer to 1 s.d. are retained to provide the reference clusters, the others are discarded. The level 1 s.d. is chosen empirically by looking at an histogram.

Discriminant analysis is then performed, using the thus designed central part of the clusters as active units. All the rejected units are taken as passive ones. The smallest cluster has now 15 units.

The discriminant analysis correctly classifies 98.5 percent of the active units. In addition, it defines the probability for each unit to belong to each cluster. These probabilities are available for a final step, as possible example of which would be to change, a posteriori, a few cluster attributions, taking into account spatial constraints (i.e. the behaviour of spatially closer wells). Another way to use these probabilities is as variables in a last P.C.A. in order to display the proximities between clusters in the F1,F2 scatter plot as suggested by Kaufman and Rousseeuw (1990, pp 195-196).

7 Discussion

An external validation can be proposed for this example. The 179 wells were extracted from a real case including several hundreds of wells, previously classified into 9 classes. Cramer's coefficient measures the convergence between these a priori categories, and the clusters found. A value of 0.80 is found using the global data set, compared to 0.84 using the robust approach we have developed.

At different steps, the proposed procedure takes into account the density of the population ; it could be a contribution to a robust approach of hierarchical clustering. Using the Ward method, some instabilities have been noted in the presence of outlying units. More stable clusters have here been searched in two successive steps. This way clearly exhibits the different behaviour of the Ward and centroid methods in the presence of outliers.

Another way could be use the preliminary P.C.A. for detecting outlying units, as suggested by Gnanadesikan (1977), for example by using the 'projections revelatrices' (Caussinus, 1992).

The choice of the number of clusters is not discussed here. The suitable number clearly depends here on the presence of outlying units, and on the efficiency of the used method to discard these outliers. In the proposed

methodology, the precise number of clusters taken from the centroid method is not sensitive, because outliers are flagged even with a limited number of clusters. In the next (Ward) step, we use a number of clusters according to a preceding exploratory study (Valois, 2000), the validation of this number is not the topic of the paper, an abundant litterature discusses this point.

8 Conclusion

In order to classify decline curves from oil industry, several preliminary steps of data transformation are needed, including hyperbolic modelization or smoothing, to produce a complete 2D data table (wells, attributes). The challenge is then to provide balanced clusters, without outliers or small sized clusters. The main point of the proposed approach is to select the most adapted algorithm for each of the successive steps : the preprocessing phase (centroid), the main clustering step (Ward) and finally the validation (discriminant analysis). This approach can be used for any set of quantitative data. It is easy to carry out, because specific computing is very limited. It maintains all the main advantages of hierarchical clustering while providing more complete tools for interpreting the results, namely probabilities from the discriminant analysis.

References

- BAYNE, C.K., BEAUCHAMP, J.J., BEGOVITCH , C.L., and KANE, V.E. (1980): Monte-Carlo comparisons of selected clustering procedures. *Pattern Recognition*, 12 , 51–62.
- CAUSSINUS, H. (1992): Projections revelatrices. In: J.J. Droebecke, B. Fichet and Ph. Tassi (Eds.): *Modèles pour l'analyse des données*. Economica, Paris, 241–266.
- CLEVELAND, W.S. (1993): *Vizualising data*. Hobart Press, Summit, New Jersey.
- GNANADESIKAN R. (1977): *Methods for Statistical Data Analysis of multivariate observations*. Willey-Interscience, New-York.
- JAIN, K.A. and DUBES, R.C. (1988): *Algorithms for clustering data*. Prentice Hall, Englewood Cliffs, New Jersey.
- HARRIS, C.W. and KAISER H.F. (1964): Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29, 4, 347–362.
- HARMAN, H.H. (1976): *Modern Factor Analysis*. Univ. of Chicago Press, Chicago.
- KAUFMAN, L. and ROUSSEEUW P.J. (1990): *Finding groups in data, An introduction to cluster analysis*. Willey-Interscience, New-York.
- MILLIGAN, G.W. and COOPER, M.C. (1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50, 159–179.
- VALOIS, J.-P. (2000): L'approche graphique en analyse des données, *Journal de la Soc. Fr. de Stat..* to be published.
- VONEIFF, G.W. (1996): A new Approach to Large Sale Infill evaluations applied to the Ozona (Canyon) Gas Sands. In: *Permian Basin Oil and gas Conference*, Midland, Texas, 495–510, rf. SPE: 35203.

A Minimax Solution for Sequential Classification Problems

Hans J. Vos

Faculty of Educational Science and Technology, Twente University.
P.O. Box 217, 7500 AE Enschede, The Netherlands
(e-mail: vos@edte.utwente.nl)

Abstract. The purpose of this paper is to derive optimal rules for sequential classification problems. In a sequential classification test, for instance, in an educational context, the decision is to classify a student as a master, a partial master, a nonmaster, or continue testing and administering another random item. The framework of minimax sequential decision theory is used by minimizing the maximum expected losses associated with all possible decision rules at each stage of testing. The main advantage of this approach is that costs of testing can be explicitly taken into account.

1 Introduction

A fixed-length classification test is used to classify subjects on the basis of a fixed number of responses into two or more mutually exclusive categories (e.g., suitable and unsuitable). The test on the basis of what it is decided into which category a subject is classified might also be a sequential classification test. In this case, in addition to the classification decisions, also the decision of continue testing and administering another random item is available. Sequential classification tests are designed with the goal of reducing the average test length while at the same time holding the number of misclassified subjects to an acceptable minimum.

The purpose of this paper is to derive optimal rules for sequential classification problems. Decision rules are hereby prescriptions specifying for each possible response pattern what decision has to be taken. The framework of minimax sequential decision theory (e.g., Ferguson (1967)) is used by minimizing the maximum expected losses associated with all possible decision rules at each stage of testing. To illustrate the approach, the derived rules will be applied to an educational problem where three possible classification decisions are distinguished, namely declaring mastery, partial mastery, or nonmastery.

2 Framework of minimax sequential decision theory

In minimax sequential decision theory, three basic elements are distinguished. First, a psychometric model relating the probability of a correct response

to student's (unknown) true level of functioning. Second, a loss structure evaluating the total costs and benefits for each possible combination of classification outcome and student's true level of functioning. Finally, costs of testing (i.e., costs of administering one additional item) must be explicitly specified. Doing so, maximum expected losses associated with the mastery, partial mastery, and nonmastery classification decisions can be calculated straightforward at each stage of testing. As far as the maximum expected loss associated with continue testing concerns, this quantity is determined by averaging the maximum expected losses associated with each of the possible future decision outcomes relative to the probability of observing those outcomes. Optimal rules (i.e., minimax sequential rules) are now obtained by choosing the decision that minimizes maximum expected loss at each stage of testing using techniques of dynamic programming (i.e., backward induction). This technique starts by considering the final stage of testing and then works backward to the first stage of testing.

3 Notation

In the following, a sequential classification test is supposed to have a maximum length of n ($n \geq 1$). Let the observed item response at each stage of testing k ($1 \leq k \leq n$) for a randomly sampled student be denoted by a discrete random variable X_k , with realization x_k , which can take the values 0 and 1 for respectively an incorrect and correct response. Furthermore, let the number-correct score be denoted by a discrete random variable $S_k = X_1 + \dots + X_k$, with realization $s_k = x_1 + \dots + x_k$ ($0 \leq s_k \leq k$). Student's true level of functioning is unknown due to measurement and sampling error. Therefore, let student's true level of functioning be denoted by a continuous random variable T , with realization $t \in [0, 1]$. Finally, two criteria levels t_0 and t_1 ($0 \leq t_0 < t_1 \leq 1$) on T must be specified by the decision-maker using methods of standard-setting (e.g., Nedelsky (1954)). A student is considered a true nonmaster or true master if t is smaller or larger than t_0 and t_1 , respectively, whereas a student is considered a true partial master if $t_0 < t < t_1$.

4 Loss structure, costs of testing, and psychometric model

Here, the threshold loss function (e.g., Vos (1998)) is adopted as the loss structure involved. The choice of this loss function implies that the “seriousness” of all possible consequences of the classification outcomes can be summarized by possibly different constants, one for each of the possible classification outcomes. For our sequential classification problem, threshold loss can be formulated at each stage of testing k as follows:

Table 1. Table for threshold loss function at stage k of testing.

Classification decision	True level		
	$T \leq t_0$	$t_0 < T < t_1$	$T \geq t_1$
Declare nonmastery	kc	$l_{12} + kc$	$l_{13} + kc$
Declare partial mastery	$l_{21} + kc$	kc	$l_{23} + kc$
Declare mastery	$l_{31} + kc$	$l_{32} + kc$	kc

The value c represents the costs of administering one random item. Table 1 was rescaled in such a way that the losses l_{11} , l_{22} , and l_{33} associated with the correct classification outcomes were equal to zero.

Finally, at each stage of testing k , the binomial model will be adopted for specifying the statistical relation between s_k and t , $f(s_k | t)$. Its distribution is given by $\binom{k}{s_k} t^{s_k} (1-t)^{k-s_k}$.

5 Sufficient conditions for monotone solutions

The optimal rules in this paper are assumed to have monotone forms, that is, they take the form of cutting scores on the test. Generally, two conditions sufficient for setting cutting scores must be satisfied (e.g., Ferguson (1967)). First, $f(s_k | t)$ must have a monotone likelihood ratio (MLR). This condition implies that the higher s_k , the more likely it will be that t is high too. Second, the condition of monotonic loss must hold, that is, there must be an ordering of the classification decisions such that for each pair of adjacent classification decisions the loss functions possess at most one point of intersection.

6 Applying the minimax principle to the fixed-length classification problem

Given response pattern (x_1, \dots, x_k) , the minimax principle will first be applied to the fixed-length classification problem (i.e., declaring mastery, partial mastery, or nonmastery). It is assumed that there exist two cutting scores on S_k , say $s_0(k)$ and $s_1(k)$ ($0 \leq s_0(k) < s_1(k) \leq k$), such that mastery is declared when $s_k \geq s_1(k)$, partial mastery when $s_0(k) < s_k < s_1(k)$, and nonmastery when $s_k \leq s_0(k)$. Furthermore, assuming the conditions of monotonicity are satisfied, it follows (e.g., Ferguson (1967)) that if the maximum expected loss associated with declaring mastery is smaller than the maximum expected loss associated with declaring partial mastery, then it is also smaller than the maximum expected loss associated with declaring nonmastery.

Let $y = 0, 1, \dots, k$ represent all possible values s_k can take, it then can easily be verified from Table 1 and the assumed psychometric model that mastery is declared when number-correct score s_k is such that

$$\begin{aligned}
& \sup_{t \leq t_0} (l_{31} + kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} + \sup_{t_0 < t < t_1} (l_{32} + kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} \\
& + \sup_{t \geq t_1} (kc) \sum_{y=0}^{s_k-1} \binom{k}{y} t^y (1-t)^{k-y} < \sup_{t \leq t_0} (l_{21} + kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} \\
& + \sup_{t_0 < t < t_1} (kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} + \sup_{t \geq t_1} (l_{23} + kc) \sum_{y=0}^{s_k-1} \binom{k}{y} t^y (1-t)^{k-y}.
\end{aligned}$$

Since the cumulative binomial distribution function is decreasing in t , and rearranging terms, it follows that mastery is declared when s_k is such that

$$(l_{31} - l_{21} - l_{32}) \sum_{y=s_k}^k \binom{k}{y} t_0^y (1-t_0)^{k-y} + (l_{23} + l_{32}) \sum_{y=s_k}^k \binom{k}{y} t_1^y (1-t_1)^{k-y} < l_{23}.$$

If the above inequality does not hold, it can easily be verified from Table 1 that partial mastery is declared when s_k is such that

$$\begin{aligned}
& \sup_{t \leq t_0} (l_{21} + kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} + \sup_{t_0 < t < t_1} (kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} + \\
& \sup_{t \geq t_1} (l_{23} + kc) \sum_{y=0}^{s_k-1} \binom{k}{y} t^y (1-t)^{k-y} < \sup_{t \leq t_0} (kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} + \\
& \sup_{t_0 < t < t_1} (l_{12} + kc) \sum_{y=s_k}^k \binom{k}{y} t^y (1-t)^{k-y} + \sup_{t \geq t_1} (l_{13} + kc) \sum_{y=0}^{s_k-1} \binom{k}{y} t^y (1-t)^{k-y},
\end{aligned}$$

and that nonmastery is declared otherwise. Rearranging terms, it follows that partial mastery is declared if s_k is such that

$$(l_{12} + l_{21}) \sum_{y=s_k}^k \binom{k}{y} t_0^y (1-t_0)^{k-y} + (l_{13} - l_{23} - l_{12}) \sum_{y=s_k}^k \binom{k}{y} t_1^y (1-t_1)^{k-y} < l_{13} - l_{23},$$

and that nonmastery is declared otherwise.

7 Derivation of minimax sequential rules

Given response pattern (x_1, \dots, x_k) , let $d_k(x_1, \dots, x_k)$ denote the decision rule yielding the minimum of the maximum expected losses associated with the three classification decisions, and let the maximum expected loss associated with this minimum be denoted as $V_k(x_1, \dots, x_k)$. At each stage of testing k ,

$d_k(x_1, \dots, x_k)$ can then be obtained by using the following backward induction computational scheme:

First, since the continue testing option is not available at the final stage of testing n , it follows immediately that the minimax sequential rule at stage n is given by $d_n(x_1, \dots, x_n)$, and its associated maximum expected loss is given by $V_n(x_1, \dots, x_n)$. Next, the minimax sequential rule at stage $(n-1)$ is computed by comparing $V_{n-1}(x_1, \dots, x_{n-1})$ with the maximum expected loss associated with the continue testing option. As noted before, the maximum expected loss associated with administering one more item at stage $(n-1)$ is computed by averaging the posterior expected losses associated with each of the possible future decision outcomes at stage n with weights corresponding to the probabilities of observing those outcomes (i.e., the posterior predictive distributions).

Let $P(X_n = x_n | x_1, \dots, x_{n-1})$ denote the posterior predictive distribution of X_n , given response pattern (x_1, \dots, x_{n-1}) , the maximum expected loss associated with administering one more item, $E[V_n(x_1, \dots, x_{n-1}, X_n) | x_1, \dots, x_{n-1}]$, is then computed as:

$$\sum_{x_n=0}^{x_n=1} V_n(x_1, \dots, x_n) * P(X_n = x_n | x_1, \dots, x_{n-1}).$$

The minimax sequential rule at stage $(n-1)$ is now given by: Administer one more item if $E[V_n(x_1, \dots, x_{n-1}, X_n) | x_1, \dots, x_{n-1}]$ is smaller than $V_{n-1}(x_1, \dots, x_{n-1})$, and take decision $d_{n-1}(x_1, \dots, x_{n-1})$ otherwise.

To compute the maximum expected loss associated with the continue testing option, it is convenient to introduce the risk at each stage of testing k , which will be denoted as $R_k(x_1, \dots, x_k)$. Let the risk at the final stage of testing n be defined as $V_n(x_1, \dots, x_n)$. Generally, given response pattern (x_1, \dots, x_{k-1}) , the risk at stage $(k-1)$, $R_{k-1}(x_1, \dots, x_{k-1})$, is then computed inductively as a function of the risk at stage k :

$$\min\{V_{k-1}(x_1, \dots, x_{k-1}), E[R_k(x_1, \dots, x_{k-1}, X_k) | x_1, \dots, x_{k-1}]\}.$$

The maximum expected loss associated with administering one more item after $(n-2)$ items have been administered, $E[R_{n-1}(x_1, \dots, x_{n-2}, X_{n-1}) | x_1, \dots, x_{n-2}]$, can then be computed as the expected risk at stage $(n-1)$:

$$\sum_{x_{n-1}=0}^{x_{n-1}=1} R_{n-1}(x_1, \dots, x_{n-1}) * P(X_{n-1} = x_{n-1} | x_1, \dots, x_{n-2}).$$

The minimax sequential rule at stage $(n-2)$ is now given by: Administer one more item if $E[R_{n-1}(x_1, \dots, x_{n-2}, X_{n-1}) | x_1, \dots, x_{n-2}]$ is smaller than $V_{n-2}(x_1, \dots, x_{n-2})$; otherwise, decision $d_{n-2}(x_1, \dots, x_{n-2})$ is taken. Following the same computational backward scheme, the minimax sequential rules at stages $(n-3), \dots, 1$ are computed.

8 Computation of posterior predictive distributions

To compute the posterior predictive distribution $P(X_k \mid x_1, \dots, x_{k-1})$, a prior distribution must be specified representing our best prior beliefs concerning student's true level of functioning. Since the minimax principle is very attractive (e.g., Coombs, Dawes, and Tversky (1970)) when the only information is s_k , the uniform prior on the standard interval $[0,1]$ is taken as the prior in this paper. In combination with a binomial distribution for the psychometric model, it is then known (e.g., Ferguson (1967)) that $P(X_k = 1 \mid x_1, \dots, x_{k-1}) = (1 + s_{k-1})/(k+1)$, and thus, that $P(X_k = 0 \mid x_1, \dots, x_{k-1}) = (k - s_{k-1})/(k+1)$. For given loss parameters and maximum test length n , the appropriate decision (i.e., mastery, partial mastery, nonmastery, or continue testing) can now be determined at each stage of testing k for different number-correct score s_k . A computer program called MINIMAX has been developed for this purpose, which is available from the author upon request.

9 Discussion

Within the framework of minimax sequential decision theory, optimal rules were derived for sequential classification problems and applied to the educational problem of deciding on mastery, partial mastery, nonmastery, or continue testing and administering another random item. It should be emphasized, however, that the sequential decision-making procedures advocated here have a larger scope. For instance, sequential classification problems may be important in psychodiagnostic for making decisions concerning the effectiveness (e.g., significantly effective, moderately effective, or ineffective) of a new treatment for patients suffering from some mental health problem. Costs of exposing a random patient to the new treatment might be even more realistic in such clinical settings, since these costs might be quite large.

In addition to the simple binomial model, also more sophisticated psychometric models may be used in deriving optimal rules. For instance, given student's true level of functioning, the probability of a correct response might also be modeled by the Rasch model from item response theory.

References

- COOMBS, C.H., DAWES, R.M., and TVERSKY, A. (1970): *Mathematical Psychology: An Elementary Introduction*. Englewood Cliffs, New Jersey.
- FERGUSON, T.S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York.
- NEDELSKY, L. (1954): Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement*, 14, 3–19.
- VOS, H.J. (1998): Compensatory Rules for Optimal Classification With Mastery Scores. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer, Heidelberg, 211–218.

Comparison of Ultrametrics Obtained With Real Data, Using the P_L and VAL_{Aw} Coefficients

Isabel Pinto Doria¹², Georges Le Calvé³, and Helena Bacelar-Nicolau¹

¹ LEAD, Faculdade de Psicologia e de Ciências da Educação
University of Lisbon, Lisbon, Portugal

² Escola Superior de Tecnologia da Saúde de Lisboa, Lisbon, Portugal

³ Université de Rennes II, Rennes, France

Abstract. We compare 20 ultrametric matrices generated by the classifications obtained from 20 similarity indices for binary variables on the same group of data, that were studied by Hubálek (1982). To measure the similarity between the ultrametric matrices we use the P_L coefficient proposed by Le Calvé (1977) and the Validity of Affinity Coefficient WW, VAL_{Aw} proposed by Bacelar-Nicolau (1988). By means of hierarchical cluster analysis and principal component analysis on the similarity matrices obtained with those two coefficients, we draw conclusions about the 20 similarity indices and compare results for P_L and VAL_{Aw} coefficients. The results obtained with these two coefficients are very similar and are also similar to the results obtained by Hubálek. Finally we introduce in this ultrametrics/coefficients comparative study the simple matching coefficient, Sokal and Michener (1958), and observe, using P_L or VAL_{Aw} coefficients, its particular behaviour in relation to the other indices.

1 Introduction

In the following study we compare various classifications/ultrametrics on the same objects, that are obtained with different/various similarity indices and the same clustering algorithm. To do so we use the P_L and the VAL_{Aw} coefficients. The P_L coefficient is a similarity index between variables proposed by Le Calvé (1977) that generalises an index by Lerman (1972). The Validity of Affinity Coefficient WW, VAL_{Aw} , is a similarity index proposed by Bacelar-Nicolau (1988) and is also based on the same Lerman index. These two probabilistic coefficients are general similarity indices that allow us to compare variables with different structures, in particular, ultrametrics (Doria et al., 1999).

We base our comparison on the example used by Hubálek (1982) that results from a concrete problem: a comparison of 20 similarity indices on dichotomic data. For the definition of the similarity indices, see Table 1, where a, b, c and d refer to the four cells of a 2×2 contingency table, designating respectively, the number of co-presences, presence/absence, absence/presence and co-absences of the attributes.

Table 1. Similarity coefficients for binary variables proposed by several authors.

Similarity coefficient	Author
$A_3 = \frac{a}{b+c}$	Kulczynski (1927).
$A_4 = \frac{a}{a+b+c}$	Jaccard (1901), Sneath (1957)
$A_5 = \frac{a}{a+(b+c)/2}$	Dice (1945), Sørensen (1948)
$A_6 = \frac{a}{a+2(b+c)}$	Sokal & Sneath (1963)
$A_7 = [\frac{a}{a+b} + \frac{a}{a+c}] / 2$	Kulczynski (1927)
$A_{11} = \frac{a}{[(a+b)(a+c)]^{1/2}}$	Driver & Kroeber (1932)
$A_{14} = \frac{a}{a+b+c+d} = \frac{a}{n}$	Russell & Rao (1940)
$A_{15} = \frac{a}{(ab+ac)/2+bc}$	Mountford (1962)
$A_{18} = [\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d}] / 4$	Sokal & Sneath (1963)
$A_{25} = \frac{ad}{(a+b)(c+d)(a+c)(b+d)]^{1/2}}$	Sokal & Sneath (1963)
$A_{30} = \frac{ad-bc}{ad-bc}$	Yule (1912), Pearson & Heron (1913)
$A_{32} = \frac{\sqrt{ad+a-b-c}}{\sqrt{ad+a+b+c}}$	Baroni-Urban & Buser (1976)
$A_{34} = \pm(\frac{\chi^2}{\chi^2_{max}})^{1/2}$ with the sign of $(ad - bc)$	Cole (1949): 'C ₇ '; Hurlbert (1969)
$A_{35} = \pm[\frac{\chi^2 - \chi^2_{min}}{\chi^2_{max} - \chi^2_{min}}]^{1/2}$	Hurlbert (1969)
$A_{36} = \frac{ad-bc}{ad+bc}$	Yule (1900)
$A_{37} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	Yule (1912)
$A_{38} = \cos[\frac{180\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}]$	Pearson & Heron (1913)
$A_{39} = \frac{4(ad-bc)}{(a+d)^2 + (b+c)^2}$	Michael (1920)
$A_{40} = \frac{na}{(a+b)(a+c)}$	Forbes (1907)
$A_{43} = \frac{(a-a')}{(a+a')} = \frac{na-(a+b)(a+c)}{na+(a+b)(a+c)}$	Tarwid (1960)

We compute the coefficients P_L and VAL_{Aw} between the 20 ultrametrics generated by the 20 similarity indices for binary variables. So, we obtain two similarity matrices between these indices. Assimilating them to scalar products, their diagonalisation is used as a principal component analysis (PCA). We also use hierarchical cluster analysis (HCA) directly on these matrices.

We have performed the same computations with the coefficients P_L and VAL_{Aw} on the 20 ultrametrics in order to compare the behaviour of these coefficients on real data, as well as with Hubálek's results. Finally we compare the simple matching coefficient, Sokal and Michener (1958), with the other similarity indices studied using the P_L and the VAL_{Aw} coefficients.

2 The P_L and VAL_{Aw} coefficients

Definition of the P_L coefficient (Le Calvé (1977)): A probabilistic similarity coefficient, designated by $P_{Lx,y}$, is defined by the probability of $Z_{Lx,y}$ being smaller than z_l : $P_L(x, y) = P(Z_{Lx,y} \leq z_l(x, y)) = \Phi(z_l(x, y))$. The random variable Z_L is the standardised similarity, $Z_L = (L_{X,Y}(w) - \mu) / \sigma$, and

Table 2. Results obtained with hierarchical cluster analysis ($P_L + Average\ Linkage$): the best clusters in conformity with the “level statistics” criterion (Lerman, 1970; Bacelar-Nicolau, 1972).

Partitions obtained on the k -th level	Statistic: STAT(k)
{A ₃ ,A ₃₉ ,A ₆ ,A ₄ ,A ₁₄ ,A ₅ ,A ₁₁ ,A ₂₅ ,A ₃₂ }	STAT(17)
{A ₇ ,A ₃₀ ,A ₁₅ ,A ₃₇ ,A ₃₈ ,A ₃₆ ,A ₄₀ ,A ₄₃ ,A ₁₈ }, {A ₃₄ ,A ₃₅ }	11.4820
{A ₃ ,A ₃₉ ,A ₆ ,A ₄ ,A ₁₄ ,A ₅ ,A ₁₁ ,A ₂₅ },{A ₃₂ }, {A ₃₄ ,A ₃₅ }	STAT(15)
{A ₇ ,A ₃₀ ,A ₁₅ ,A ₃₇ ,A ₃₈ ,A ₃₆ ,A ₄₀ ,A ₄₃ ,A ₁₈ }, {A ₃₄ ,A ₃₅ }	11.1466

$L_{X,Y}(\theta, \theta') = <\theta X \theta^t, \theta' Y \theta'^t>$, $\forall w \in (\theta, \theta')$, considering the set of all permutations couples, $\Omega = \Theta(I) \times \Theta(I)$ defined on I , provided with a probability measure uniformly distributed. P_L is the probability distribution function of that standardised similarity being observed.

Definition of the VAL_{Aw} coefficient (Bacelar-Nicolau (1988)): A probabilistic similarity coefficient of validity linkage (VL) type, associated to both the affinity coefficient and the limit theorem of Wald and Wolfowitz, is defined by $VAL_{Aw}(x, y) = \alpha_{Aw}(x, y) = \text{Prob}(A_W(x, y) \leq a_W(x, y)) \approx \Phi(a_W(x, y))$, which is called the Validity of Affinity Coefficient WW, VAL_{Aw} . The random variable $A_W(x, y) = (A(x, y) - \mu)/\sigma$ has asymptotic standard normal distribution, the affinity coefficient $a(x, y)$ is the actual value of the random variable $A(x, y)$ and Φ denotes the standard normal distribution function.

3 Comparison of ultrametrics and coefficients: Results

The data we have analysed refer to the co-occurrence of 7 types of Chaetomium (Ascomycetes) mushrooms in 869 bird nests, Hubálek (1982). We have made 20 classifications of the seven mushrooms with hierarchical cluster analysis. We used the average linkage with each one of the 20 indices, having then obtained 20 classifications on the same objects, that only differ from the way the similarity was observed. We have then obtained 20 ultrametrics/dendograms (they show clearly that different results have been obtained with different coefficients) and we want to compare them.

To do that we have used the P_L index on the ultrametrics which constitute the scores matrices. Once the P_L value is calculated among the ultrametrics we have obtained a similarity matrix P_L (20×20). To represent it we have used two data analysis techniques:

- Hierarchical cluster analysis with average linkage (see Table 2).
- Principal component analysis of this similarity table. The results are in Figure 1.

We have performed the same computations with the coefficient VAL_{Aw} on the 20 ultrametrics. The best HCA ($VAL_{Aw} + Average\ Linkage$) clusters, in conformity with the “level statistics” criterion, are obtained on the 9th level (STAT(9)=11.90; {A₃,A₆,A₄,A₅,A₁₁,A₂₅,A₁₄,A₃₉,A₃₂}, {A₇,A₃₀,A₁₅,A₃₇,A₃₆}, {A₃₄,A₃₅}).

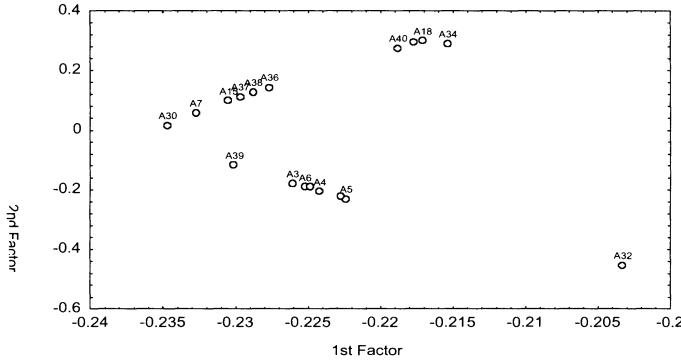


Fig. 1. Plot of the 1st component plane obtained by PCA of the similarity matrix P_L .

$A_{38}, A_{18}, A_{40}, A_{43}, A_{34}, A_{35}\}$) and on the 7th level ($\text{STAT}(7) = 11.30$; $\{A_3, A_6, A_4, A_5, A_{11}, A_{25}, A_{14}, A_{39}\}$, $\{A_{32}\}$, $\{A_7, A_{30}, A_{15}, A_{37}, A_{36}, A_{38}, A_{18}, A_{40}, A_{43}\}$, $\{A_{34}, A_{35}\}$). The HCA results obtained with these coefficients, P_L and VAL_{Aw} , are very similar.

The similarity matrices P_L and VAL_{Aw} are not positive semi-definite. The 1st component explains, in both principal component analyses, most of the variability (PCA+ P_L : 93.77%; PCA+ VAL_{Aw} : 91.14%). The 2nd component has a weak percentage of explained inertia associated to it (PCA+ P_L : 6.96%; PCA+ VAL_{Aw} : 10.54%). In both principal component analyses we see that (i) all variables/binary coefficients have negative coordinates on the 1st axis, and (ii) the 2nd axis separates the variables. The PCA results obtained with these coefficients are very similar. Besides, for each coefficient, the best partitions coincide with the results of the PCA on the 1st factorial plane (Figure 1).

Table 3. Cluster descriptions.

Cluster 1	<ul style="list-style-type: none"> • All indices present nil association at zero a (they are null when $a = 0$), excluding A_{39}. • Indices $A_3, A_4, A_6, A_{25}, A_{32}$, and A_{39} are not linear. • d does not take part of indices $A_3, A_4, A_5, A_6, A_{11}$, and A_{14}.
Cluster 2	<ul style="list-style-type: none"> • Some of these indices present nil association. • Indices $A_{15}, A_{36}, A_{37}, A_{38}$, and A_{43} are not linear. • d takes part of the majority of cluster's 2 indices: $A_{18}, A_{30}, A_{36}, A_{37}, A_{38}, A_{40}$, and A_{43}.
Cluster 3	<ul style="list-style-type: none"> • These indices present all proprieties: linearity, nil association.

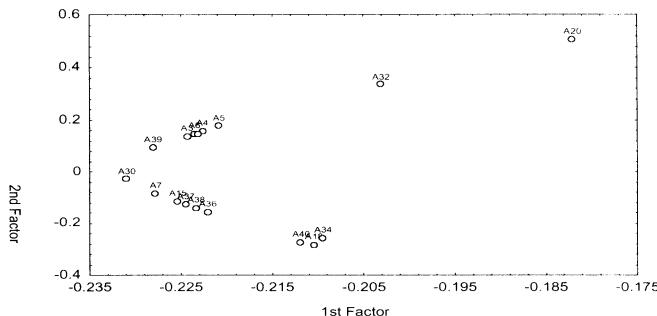


Fig. 2. Plot of the 1st component plane obtained with the PCA of the similarity matrix P_L (21×21).

Considering the partition obtained with the algorithms *P_L + Average Linkage* and *VAL_{Aw} + Average Linkage* (Cluster 1: {A₃, A₃₉, A₆, A₄, A₁₄, A₅, A₁₁, A₂₅, A₃₂}, Cluster 2: {A₇, A₃₀, A₁₅, A₃₇, A₃₈, A₃₆, A₄₀, A₄₃, A₁₈}, and Cluster 3: {A₃₄, A₃₅}), we can say that clusters are separated more by their mathematical proprieties than by their mathematical aspect (Table 3).

When we compare these results with the best partitions obtained by Hubálek (Pearson's $r + Average\ Linkage$), the main difference between them is observed at indices A₇ and A₁₈. We think A₁₈ is better placed in cluster 4 ($r = 0.95$), as we obtained, considering this coefficient's results when classifying the mushrooms.

4 What happens to the Simple Matching Coefficient when compared with the 20 coefficients?

It will be interesting to study the simple matching coefficient, Sokal and Michener (1958), as it has a simple mathematical aspect, $A_{20} = (a + d)/n$, close to $A_{14} = a/n$, only in appearance. When Hubálek classifies the mushrooms, he observes that this coefficient gives very different results from the others studied. This special aspect is confirmed when we introduce this new coefficient in the PCA (Figure 2). And this may happen because in this particular example d is too large in relation to a . So this index could give different results from the others in the case of rare phenomena.

5 Conclusions

When applied to the ultrametrics obtained with these 20 binary indices on the same data, the three coefficients P_L , VAL_{Aw} , and r gave very similar results or partitions. This shows a strong hierarchical cluster analysis structure. When

we look at the partitions we see that clusters are separated more by their mathematical proprieties than by their mathematical aspect.

The simple matching coefficient, Sokal and Michener (1958), when compared with the 20 coefficients studied, using P_L or VAL_{Aw} coefficients with PCA, behaves in a different way, which agrees with the Hubálek's statement. Perhaps this happens because we are in presence of a rare phenomenon.

Acknowledgements

We thank Mr. S. Camiz for making his EUClidean APProximation Software available. This study has been partially supported by Franco-Portuguese Programme ADAM (Embassy of France and Portuguese Ministry of Science and Technology-ICCTI, coordinated by H. Bacelar-Nicolau (LEAD/FPCEUL, Lisbon) and G. Saporta (CNAM, Paris)) and the CEAUL/FCT.

References

- BACELAR-NICOLAU, H. (1972): Analyse d'un Algorithme de Classification. *Thèse de 3ème Cycle*, Univ. Pierre et Marie Curie, Paris.
- BACELAR-NICOLAU, H. (1988): Two Probabilistic Models for Classification of Variables in Frequency Tables. In: H.-H. Bock (Ed.): *Classification and Related Methods of Data Analysis*. Elsevier Sciences, Publishers B.V., North Holland, 181–186.
- DORIA, I., LE CALVÉ, G., and BACELAR-NICOLAU, H. (1999): Comparing Several Similarity Indices On Dichotomics Based on The Associated Ultra-metrics Obtained With Real Data. In: H. Bacelar-Nicolau, F. Costa Nicolau, J. Janssen (Eds.): *Applied Stochastic Models and Data Analysis. Quantitative Methods in Business and Industry Society*. INE, Lisboa, 142–147.
- HUBÁLEK, Z. (1982): Coefficients of Association and Similarity Based on Binary (Presence-Absence) Data: an Evaluation. *Biolog. Rev.*, 57, 669–689.
- LE CALVÉ, G. (1977): Un Indice de Similarité pour des Variables de Type Quelconque. *Revue Statistique et Analyse des Données*, juin.
- LERMAN, I.C. (1970): Sur l'Analyse des Données Préalable à une Classification Automatique - Proposition d'une Nouvelle Mesure de Similarité. *Mathématiques et Sciences Humaines*, 32, 5–15.
- LERMAN, I.C. (1972): Étude Distributionnelle de Statistiques de Proximité entre Structures Algébriques Finies du Même Type; Application à la Classification Automatique. *Cahiers du B.U.R.O.*, 19.
- SOKAL, R.R. and MICHENER, C.D. (1958): A Statistical Method for Evaluating Systematic Relationships. *The University of Kansas Scientific Bulletin*, 38, 1409–1438.

Numerical Comparisons of two Spectral Decompositions for Vertex Clustering

P. Kuntz¹ and F. Henoux²

¹ IRIN-IRESTE - BP 60601 - 44306 Nantes cedex 3 - France

² ENST - Rue Barrault - 75013 Paris - France

pkuntz@ireste.fr

Abstract. We study multi-way partitioning algorithms of a hypergraph which are based on its prior transformation into a geometric object by constructing a one-to-one mapping between the vertex set and a point set in a Euclidean space. The coordinates of the points are generated by a spectral decomposition of a positive semi-definite matrix. Here, we compare the decomposition of the discrete Laplacian of a graph associated with the hypergraph to that of the Torgerson matrix associated with a dissimilarity coefficient. Numerical results are presented on standard test cases of large sizes from the integrated circuit design literature.

1 Introduction

In numerous fields, such as integrated circuit design, the size of relationship networks keeps growing and a preliminary stage of partitioning is often required to deal with the frequently inextricable associated computations. This stage is critical since it is often the only one which takes into account the network as a whole.

Generally speaking, partitioning a graph or a hypergraph aims at providing classes of practicable sizes that ensure both the quality of local cohesion between all classes (abundance of intra-class links) and the quality of dislocation of the classes (rarity of inter-class links). More formally, let us consider a hypergraph $H = (V, E)$ with vertices $V = \{v_1, \dots, v_n\}$ and hyperedges $E = \{e_1, \dots, e_m\}$. A k -partitioning P^k of the vertex set V is a set of k disjoint non empty clusters $\{V_1, \dots, V_k\}$ so that $V_1 \cup \dots \cup V_k = V$ and, most of the time, the optimization problem consists of minimizing the number of hyperedges that go across between two or more clusters in P^k . Several variants have been proposed but in order to favor size balance among the clusters. We focus here on the scaled cost criterion used in integrated circuit design (Chan and al. (1994)) : minimize $c(P^k) = \frac{1}{n(k-1)} \sum_{i=1}^k \frac{|E_i|}{|V_i|}$ where $E_i = \{e \in E | \exists (u, v) \in e \text{ s.t. } u \in V_i \text{ and } v \notin V_i\}$ is the set of hyperedges incident to V_i and different other clusters and $|V_i|$ the cardinal of V_i .

Except for very particular cases, the vast majority of k -partitioning problems are known to be NP-complete (Garey et al. (1976), Lengauer (1990)). Most of the heuristics which are integrated in software are based on iterative improvements of initial partitions (e.g. Fiduccia and Mattheyses (1982)) but,

it is well-known that they are quite sensitive to the initialization and that their probability of getting stuck on a local optimum increases dramatically with the size of the graph.

In an attempt to face such limitations, a conceptually very different approach has known an increasing development in the last decade. The basic underlying idea is to transform the initial abstract graph into a geometric object by constructing a mapping between the vertex set and a point set in a geometric space -generally a Euclidean space- which transfers the connectivity properties between components of the graph (link frequencies, ...) onto the new geometric object and which, consequently, makes new measures and efficient clustering algorithms accessible. A *d-dimensional embedding* of $H = (V, E)$ is here a set of $|V|$ points in a Euclidean space of dimension d which are in one-to-one correspondance with the vertices of V .

The most commonly used embeddings resort to a spectral decomposition of the discrete Laplacian of an associated graph with H (see Alpert and Kahng (1995b) for an overview). They are originally based on analogies between partitioning and placement problems (Hall (1970)) and motivated by the well-established relationships between eigenvectors of the Laplacian and min-cut partitionings. The hypergraph is first transformed into a weighted graph $G_H = (V, E')$ by replacing each hyperedge by a clique of weighted edges. Let $A = (a_{ij})$ be the adjacency matrix of G_H where $a_{ij} > 0$ is the weight of $(v_i, v_j) \in E'$ and $a_{ij} = 0$ if no (v_i, v_j) edge exists in E' . Let $D = (d_{ij})$ be the diagonal degree matrix s.t. $d_{ii} = \sum_{j=1}^n a_{ij}$ and $d_{ij} = 0$ if $i \neq j$. The coordinates of the vertices in the d -dimensional embedding are given by the eigenvectors $\mu_1^Q, \mu_2^Q, \dots, \mu_d^Q$ associated with the d -smallest non trivial eigenvalues $\lambda_1^Q \leq \lambda_2^Q \leq \dots \leq \lambda_d^Q$ of the Laplacian matrix $Q = D - A$ which is positive semi-definite.

The other transformations come from scaling methods. The relationships between hypergraph components are explicitly conveyed via a dissimilarity coefficient δ on $V \times V$ and the graph embedding is set as an isometric problem (e.g. Fraysseix and Kuntz (1992)). We restrict ourselves to embeddings based on the decomposition of the scalar product matrix $W = (w_{ij})$ known as the Torgerson matrix (Torgerson (1958)) :

$w_{ij} = -\frac{1}{2} (\delta^2(v_i, v_j) - \delta_{i.}^2 - \delta_{.j}^2 + \delta_{..}^2)$ where $\delta_{i.}^2 = \frac{1}{n} \sum_{j=1}^n \delta^2(v_i, v_j)$ and $\delta_{..}^2 = \frac{1}{n} \sum_{i=1}^n \delta_{i.}^2$. The coordinates of the vertices in the d -dimensional embedding are given here by the eigenvectors $\mu_1^W, \mu_2^W, \dots, \mu_d^W$ associated with the d -greatest non trivial eigenvalues $\lambda_1^W \geq \lambda_2^W \geq \dots \geq \lambda_d^W$ of W which are positive when δ is a Euclidean metric. This transformation has received less attention in graph partitioning, in particular in integrated circuit design, than the previous approach. Nevertheless, it has been used successfully for highlighting clusters in the organization of large graphs in various fields as for instance recently in social networks and in the study of connections between structures in the central nervous system (Jouve et al. (1998)).

In this communication, we compare the performances of these two geometric embeddings for the k -partitioning problem with the scaled cost criterion on standard large test cases from Computer Aided Design literature. For each test case, we apply a k -partitioning algorithm which takes the proximity of the vertices in the geometric representation into account. From a previous comparison of different clustering algorithms by Alpert and Khang (1995a), we here retain their "dynamic programming for restricted partitioning" approach which integrates both the spectral embedding and the hypergraph structure and we develop an iterative version.

2 Preliminaries to spectral embeddings

As the Laplacian properties used here apply on a graph, the hypergraph is first transformed into a weighted graph $G_H = (V, E')$ with a clique net model : a hyperedge which contains p vertices is transformed into a complete subgraph with an edge between every pair of its vertices each edge having a weight that is a function of p . Ideally, no matter how vertices of the associated clique are partitioned, the cost should be one i.e. equal to the cost of a single cut of the hyperedge. Yet, this condition is impossible to achieve (Ihler et al. (1993)). Nevertheless, numerical experiments show that the geometric approaches are quite stable with regard to the model (Lengaeur (1990)). We retain here the one proposed by Huang with a weight equal to $4/(p(p-1))$ since the expected cost of cutting a hyperedge is one for a 2-partitioning.

The embedding based on the decomposition of the Torgerson matrix closely depends on the choice of a dissimilarity coefficient on V . This choice is a function of both the information available on the hypergraph (physical properties, combinatorial features, ...) and any constraints imposed by the desired application. Moreover, the dissimilarity can be directly defined on the hypergraph as well as on its associated graph G_H . We restrict ourselves here to the dissimilarities on G_H which are based on local combinatorial relationships only. A dissimilarity $\delta(v_i, v_j)$ between any vertices v_i and v_j is here a function of the weights of the edges connected to v_i or v_j :

$$\delta(v_i, v_j) = \frac{\sum_{v_k \in V - \{v_i, v_j\}} |w(v_i, v_k) - w(v_j, v_k)|}{\sum_{v_k \in V} w(v_i, v_k) + w(v_j, v_k)}$$

where $w(v_i, v_j)$ is the weight between v_i and v_j in the clique model when v_i and v_j belong to a same hyperedge, or 0 otherwise.

3 Vertex clustering

Once the abstract graph G_H is embedded in a Euclidean space, numerous clustering algorithms can be applied. Alpert and Khang (1995a) have compared some of them for the most common criteria (diameter, complete linkage,

....) and have experimentally showed that, due to the poor correlation with the initial combinatorial partitioning objective $\text{Min } c(P^k)$, the classical clustering algorithms could lead to debased results. They consequently developed a partitioning approach that integrates both the spectral embedding and the hypergraph structure. This approach is inspired by a heuristic proposed for the traveling salesman problem (Karp (1977)) : it generates, from the geometric point set, a circular ordering $\pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ of the vertices s.t. "close" vertices on the embedding remain "close" on the tour $v_{\pi(1)}, \dots, v_{\pi(n)}$ and resorts to a dynamical programming approach for splitting this ordering according to $c(P^k)$.

The circular ordering in the d -dimensional embedding is generated by the spacefilling curve heuristic of Bartholdi and Platzman (1988) which uses the recursive construction of Sierpinski and, the obtained result is improved by a classical greedy 3-opt heuristic. The objective is then to find an optimal k -partitioning such that, for any pair $v_{\pi(i)}$ and $v_{\pi(j)}$, $i < j$, of a cluster V_h one of the following properties holds :

- for any m in $\{0, \dots, n\}$ s.t. $i \leq m \leq j$ then $v_{\pi(m)} \in V_h$
- for any m in $\{0, \dots, n\}$ s.t. $m \leq i$ or $m \geq j$ then $v_{\pi(m)} \in V_h$

Let $V_{[i,j]} = \{v_{\pi(i)}, v_{\pi(i+1)}, \dots, v_{\pi(j)}\}$, $1 \leq i \leq j \leq n$, be a set of consecutive vertices on a linear ordering deduced from the tour by deletion of an edge. As $c(P^k)$ is an additive function, the problem can be solved by dynamical programming. The cost $c(P^{k+1}(V_{[i,j]}))$ of the best $k+1$ -partition of $V_{[i,j]}$ is obtained by examining all k -partitions $P^k(V_{[i,m]})$, $m < j$, found at the previous step coinjoined with single clusters :

$$c(P^{k+1}(V_{[i,j]})) = \text{Min}_{m < j} c(P^k(V_{[i,m]}) \cup V_{[m+1,j]}).$$

The application of this heuristic on a linear ordering -instead of a circular one - allows to reduce its time complexity in $O(kn^2)$. Our linear ordering results from the deletion of the greatest edge between the two classes $V_{[i,j]}, V - V_{[i,j]}$ of the best bipartition on the circular order.

Moreover, in order to detect k -partitions which reflect the "structure" of the graph, we have implemented an iterative version of this approach. Let V_1, \dots, V_k be k vertex classes deduced from a k -partitioning of V . The algorithm (embedding plus clustering) can be applied on each V_i , $1 \leq i \leq k$, to obtain a bipartition V_i^1, V_i^2 of each of them. Suppose that the minimum of $c(P^2)$ is reached for V_j . Then, a $(k+1)$ -partition of V is the set $\{V_1, \dots, V_j^1, V_j^2, \dots, V_k\}$. This iterative partitioning can be executed from initial k -partitions of different sizes ; we retain here the best results for $2 \leq k \leq 4$.

4 Experimental results

Our experiments compare performances of the two embeddings on six standard test cases from ACM/SGDA¹ (table 1). We execute k -partitionings of

¹ <http://vlsicad.cs.ucla.edu/~cheese/benchmarks.html>

d -dimensional embeddings for each spectral decompositions - denoted LAP and TOR for the dynamical programming clustering algorithm and LAP-Iter and TOR-Iter for the iterative version- with $2 \leq k \leq 10$ and two bounds for d , $1 \leq d \leq 10$ and $1 \leq d \leq 20$ (table 2). The best d -dimensional embedding for each k is in brackets and, the results obtained by Alpert and Kahng are recalled in the first row (A.K.).

Test cases	# Vertex	# Hyperedges	# Connections
Prim1	833	902	2908
bm1	882	903	2910
test04	1515	1658	5975
test06	1752	1541	6638
19ks	2844	3282	10547
prim2	3014	3029	11219

Table 1. Test cases used for experimentations

Test Case	ALG.	Number of clusters (k)									
		2	3	4	5	6	7	8	9	10	
Prim1	A.K.	13.5	14.7	20.4	24.5	28.2	31.2	34.0	38.5	40.4	
	LAP (10)	13.4 (1)	14.5 (2)	17.2 (8)	21.8 (8)	24.8 (8)	28.4 (8)	32.2 (8)	36.1 (8)	38.7 (8)	
	LAP (20)	13.4 (1)	13.9 (11)	17.2 (8)	21.8 (8)	24.8 (8)	28.4 (8)	29.8 (18)	33.3 (18)	36.3 (18)	
	TOR (10)	13.5 (4)	13.9 (7)	21.6 (9)	25.4 (9)	31.3 (8)	38.1 (8)	42.6 (8)	46.4 (8)	50.3 (8)	
	TOR (20)	13.5 (4)	13.9 (7)	21.6 (9)	25.4 (9)	31.3 (8)	38.1 (8)	42.6 (8)	45.1 (16)	48.6 (20)	
	LAP Iter.			16.1	20.3	23.3	25.4	27.5	29.1	30.5	
	TOR Iter.			16.2	20.3	23.4	25.4	27.5	29.1	30.5	
bm1	A.K.	5.53	6.61	9.02	14.8	16.8	17.9	20.0	21.8	23.4	
	LAP (10)	5.53 (1)	6.61 (4)	10.8 (6)	13.1 (6)	14.4 (6)	15.9 (6)	18.6 (6)	20.8 (6)	22.8 (6)	
	LAP (20)	5.53 (1)	6.61 (4)	8.97 (14)	10.5 (14)	12.5 (19)	15.9 (6)	18.6 (6)	20.8 (6)	22.8 (6)	
	TOR (10)	5.53 (1)	6.88 (9)	9.02 (9)	12.6 (9)	15.1 (9)	16.9 (9)	19.7 (9)	22.1 (9)	24.4 (9)	
	TOR (20)	5.53 (4)	6.61 (18)	9.01 (9)	12.5 (9)	15.1 (9)	16.9 (9)	19.7 (9)	22.1 (9)	24.4 (9)	
	LAP Iter.				10.0	11.9	14.7	17.1	19.1	20.8	
	TOR Iter.				10.0	11.9	14.3	16.6	18.8	20.4	
test04	A.K.	5.78	7.11	7.33	10.2	11.0	12.3	13.1	14.1	15.1	
	LAP (10)	5.70 (8)	6.41 (5)	8.85 (7)	10.3 (10)	11.1 (6)	12.4 (7)	13.3 (7)	14.2 (10)	15.0 (10)	
	LAP (20)	5.70 (8)	6.41 (5)	8.21 (20)	9.69 (20)	10.9 (19)	11.2 (16)	12.4 (16)	13.2 (16)	13.7 (15)	
	TOR (10)	5.85 (7)	12.0 (7)	19.4 (7)	24.6 (10)	27.3 (10)	31.3 (10)	32.0 (10)	33.4 (10)	34.9 (10)	
	TOR (20)	5.77 (19)	8.50 (11)	10.4 (11)	14.4 (19)	17.9 (16)	19.5 (16)	21.7 (16)	23.4 (19)	25.2 (16)	
	LAP Iter.			6.37	7.33	8.32	8.93	9.54	10.2	10.8	
	TOR Iter.			6.41	7.33	8.32	8.93	9.54	10.2	11.2	
test06	A.K.	8.80	10.3	11.7	12.7	13.5	15.7	16.5	17.0	17.9	
	LAP (10)	7.68 (6)	9.54 (5)	11.0 (10)	11.9 (6)	12.9 (6)	13.7 (6)	14.9 (10)	15.8 (10)	16.8 (9)	
	LAP (20)	7.68 (6)	9.27 (20)	11.0 (10)	11.9 (6)	12.7 (20)	13.6 (20)	14.9 (10)	15.8 (10)	16.8 (10)	
	TOR (10)	9.07 (4)	15.6 (6)	19.6 (6)	24.0 (9)	27.5 (6)	28.8 (8)	30.3 (8)	32.0 (8)	33.4 (8)	
	TOR (20)	9.07 (4)	15.6 (6)	18.8 (13)	20.9 (13)	23.4 (13)	25.9 (13)	28.3 (13)	30.4 (13)	32.5 (13)	
	LAP Iter.			8.93	9.25	10.1	11.1	12.0	12.6	13.6	
	TOR Iter.			10.0	10.5	11.1	11.8	12.6	13.1	14.3	
prim2	A.K.	4.77	6.98	8.18	9.20	10.0	10.8	11.6	12.2	12.7	
	LAP (10)	5.00 (3)	7.11 (8)	8.25 (8)	9.02 (8)	9.68 (8)	10.4 (8)	10.9 (8)	11.3 (8)	11.9 (8)	
	LAP (20)	5.00 (8)	7.11 (8)	8.25 (8)	9.02 (8)	9.68 (8)	10.4 (8)	10.9 (8)	11.3 (8)	11.9 (8)	
	TOR(10)	11.0 (1)	14.6 (1)	17.0 (4)	19.1 (4)	21.6 (4)	23.8 (7)	24.4 (7)	25.5 (7)	26.6 (7)	
	TOR (20)	5.33 (20)	11.0 (20)	15.0 (13)	15.7 (20)	15.9 (20)	16.3 (20)	16.6 (20)	17.2 (20)	18.0 (20)	
	LAP Iter.			6.48	7.09	7.55	8.02	8.41	8.76	9.06	
	TOR Iter.			6.65	7.20	7.64	8.06	8.37	8.62	8.87	
19ks	A.K.	4.82	5.45	5.64	6.57	7.12	7.68	8.22	8.58	8.88	
	LAP (10)	4.67 (5)	5.05 (4)	5.25 (5)	6.05 (5)	6.68 (5)	7.57 (5)	8.57 (5)	9.59 (9)	9.88 (9)	
	LAP (20)	4.67 (5)	5.05 (4)	5.25 (5)	6.05 (5)	6.68 (5)	7.57 (5)	8.57 (5)	9.14 (13)	9.15 (13)	
	TOR (10)	6.79 (9)	7.08 (9)	8.74 (4)	9.07 (4)	13.7 (4)	16.3 (5)	16.9 (5)	18.6 (5)	19.5 (5)	
	TOR (20)	5.49 (15)	6.89 (20)	7.98 (18)	9.08 (4)	12.5 (18)	13.7 (18)	14.7 (18)	15.6 (18)	16.3 (18)	
	LAP Iter.			4.72	4.79	5.06	5.31	5.52	5.82	6.11	
	TOR Iter.			5.13	5.05	5.26	5.47	5.67	5.92	6.16	

Table 2. Comparisons of Laplacian and Torgerson spectral decompositions for different test cases

For small k , results are quite similar whatever the approach but, when k increases, LAP leads to a better optimum than TOR. Multiple eigenvectors are necessary in constructing k -partitionings but LAP is distinctly less sensitive to the increasing of the d bound : best results can be obtained in high dimensions, but values for $d \leq 10$ are often close to these results. Combinatorial proximities seem better taken into account by a Laplacian spectral decomposition than a Torgerson one for a large hypergraph taken as a whole. However, we observe that the iterative versions LAP-Iter and TOR-Iter are often almost identical for any k . In this case, the spectral decomposition is applied on smallest graphs and is used to highlight the main "macroscopic" components.

We are currently carrying out additionnal experiments on two directions : the integration of cluster size bounds fixed a priori, and the study of the influence of different metrics for the Torgerson spectral decomposition.

References

- ALPERT, C.J. and KAHNG, A. (1995a) : Multiway partitioning via geometric embeddings, ordering and dynamic programming. *IEEE Trans. on CAD*, 14(11), 1342–1358.
- ALPERT, C.J. and KAHNG, A. (1995b) : Recent directions in netlist partitioning : a survey, *Integration : the VLSI Journal*, 19(1–2), 1–81.
- BARTHOLDI J.J. and PLATZMAN L.K. (1988) : Heuristics based on spacefilling curves for combinatorial problems in Euclidean space, *Management Science*, 34(3), 291–305.
- CHAN, P.K., SCHLAG, M.D. and ZIEN, J. (1994) : Spectral k-way partitioning and clustering. *IEEE Trans. on CAD*, 13(9), 1088–1096.
- FIDUCCIA, C. and MATTHEYES, M. (1982) : A linear time heuristic procedure for improving network partitions. *Proc. of the ACM/IEEE 19th Design Automation Conference*, 175–182.
- FRAYSSEIX de, H. and KUNTZ, P. (1992) : Pagination of large-scale networks ; embedding a graph in R^n for effective partitioning. *Algorithm Review*, 2(3), 105–112.
- GAREY, M., JOHNSON, D. and STOCKMEYER, L. (1976) : Some simplified NP-complete graph problems. *Theoretical Computer Science*, 1, 237–276.
- HALL, K. (1970) : An r-dimensional quadratic placement algorithm. *Management Sci.*, 17, 219–229.
- JOUVE, B., ROSENSTIEHL, P. and IMBERT, M. (1998) : A mathematical approach to the connectivity between the cortical areas of the macaque monkey. *Cerebral Cortex*, 8, 28–39.
- KARP, R.M. (1977) : Probabilistic analysis of partitioning algorithm for the traveling-salesman problem in the plane. *Mathematics of Operations Research*, 2(3), 209–224.
- LENGAUERT, T. (1990) : Combinatorial algorithms for integrated circuit layout. *J. Wiley and Sons.*
- TORGERSON, W. (1958) : Theory and methods of scaling. *J. Wiley and Sons.*

Measures to Evaluate Rankings of Classification Algorithms

Carlos Soares¹, Pavel Brazdil¹, and Joaquim Costa²

¹ LIACC/FEP, University of Porto
R. Campo Alegre 823, 4150-800 Porto, Portugal
(e-mail: {csoares,pbrazdil}@ncc.up.pt)

² LIACC/DMA-FCUP, University of Porto
R. Campo Alegre 823, 4150-800 Porto, Portugal
(e-mail: jpcosta@ncc.up.pt)

Abstract. Due to the wide variety of algorithms for supervised classification originating from several research areas, selecting one of them to apply on a given problem is not a trivial task. Recently several methods have been developed to create rankings of classification algorithms based on their previous performance. Therefore, it is necessary to develop techniques to evaluate and compare those methods. We present three measures to evaluate rankings of classification algorithms, give examples of their use and discuss their characteristics.

1 Introduction

Nowadays there are many algorithms for supervised classification that originate from several research areas (e.g. Statistics, Machine Learning, Neural Networks, etc). Therefore, the choice of the most appropriate algorithm for a given problem is not an easy task.

Recently some work has been done on developing methods which use information on the past performance of classification algorithms to generate rankings of those algorithms (Soares (1999), Nakhaeizadeh and Schnabl (1997), Gama and Brazdil (1995)).

Given that several ranking methods are available, a question arises as to how to decide which one is the best. Here we will concentrate on describing measures to evaluate rankings. Each of those measures has been employed in different ranking methods. More details about the comparative study can be found in (Soares (1999), Brazdil and Soares (2000)).

First, we describe three evaluation measures adopted. Then, we discuss those measures and, finally, we present some conclusions and future work.

2 Evaluation measures

Different methods can be used to rank algorithms based on their results on a set of datasets (Soares (1999), Nakhaeizadeh and Schnabl (1997)). These rankings can then be used to select one or more algorithms to apply on a new

problem. In this setting, we refer the ranking generated by a certain method as the *recommended ranking*. To evaluate such ranking, we first determine an *ideal ranking* for the new problem and then calculate a distance measure between the ideal and recommended rankings.

The ideal ranking represents the correct ordering of the algorithms on the new problem. The difficulty in building an ideal ranking of classification algorithms for a given dataset is essentially due to the fact that we can only measure error rates on a sample of the data. Thus, the values obtained represent *estimates* of the *true error rates* of those algorithms and not the true error rates themselves. Since those estimates have confidence intervals which may overlap, it is not always possible to establish a unique total ordering of those algorithms. We have considered two ways to deal with this problem:

- Define the ideal ranking as a partial order.
- Build an ideal ranking in the form of N total orders.

The first concept is used in the Proportion of Significant Differences Violated (PSDV) measure and the second, in two other measures, Average Correlation (AC) and Average Weighted Correlation (AWC).

Before presenting those measures, we describe the experimental setting. We have used six supervised classification algorithms: decision tree classifier (**c5**), decision tree classifier with boosting (**c5boost**), decision tree classifier which can introduce oblique decision surfaces (**ltree**), instance based classifier (**timbl**), linear discriminant (**discrim**) and naive bayes (**nbayes**). These algorithms were applied to datasets **glass** and **hepatitis**¹. Results were obtained using a 10-fold cross-validation procedure (CV). In this procedure, the dataset is divided into ten parts of equal size. Then, iteratively, the algorithm is applied to all but one part and the model obtained is tested on the remaining part. The estimate of the accuracy for each algorithm is calculated by averaging the accuracies obtained in individual folds.

Proportion of Significant Differences Violated. In this approach, the ideal ranking for a given dataset is represented as a set of restrictions of the form $j \gg k$, meaning that algorithm j is significantly better than algorithm k . These restrictions are obtained from pairwise comparisons of the performance of the algorithms formulated in terms of error rate. Each comparison is a hypothesis test concerning the difference between two population means. Here we use a paired t test on the two series of error rate values obtained with the CV procedure by a pair of algorithms. This way we are able to determine, at a certain confidence level, whether those algorithms have significantly different performance on the given dataset. The results for **c5** and **discrim** on the **glass** dataset are given in Table 1. A paired t test on these values yields a p value of 1.36%. At a 5% confidence level, we conclude that the algorithms have significantly different performance. Given that the average

¹ References for the algorithms and datasets are given in Soares (1999).

error rate of `c5` is lower than that of `discrim`, we include the restriction `c5` \gg `discrim` in the corresponding ideal ranking. The rest of the ideal ranking for the `glass` dataset is: `c5` \gg `nBayes`, `lTree` \gg `timbl`, `lTree` \gg `nBayes`, `c5boost` \gg `lTree`, `discrim` \gg `timbl`, `c5boost` \gg `timbl`, `discrim` \gg `nBayes`, `c5boost` \gg `discrim` and `c5boost` \gg `nBayes`.

Algorithm	Fold										Average
	1	2	3	4	5	6	7	8	9	10	
<code>c5</code>	28.6	19.0	23.8	33.3	47.6	19.0	27.3	45.5	22.7	31.8	29.9
<code>discrim</code>	36.4	22.7	36.4	50.0	47.6	28.6	33.3	33.3	38.1	42.9	36.9

Table 1. Error rates obtained by `c5` and `discrim` on `glass` using cross-validation.

To evaluate the recommended ranking we count the number of significant differences in the ideal ranking that are violated. To illustrate this performance measure, we evaluate a hypothetical recommended ranking: `lTree`, `c5boost`, `discrim`, `nBayes/c5` (these algorithms are tied) and `timbl` in the last position. In this recommended ranking, `c5` has lower rank than `discrim` but, in fact, it actually performs significantly better on this dataset (i.e. `c5` \gg `discrim`) as we saw earlier. Also the order between `lTree` and `c5boost` is inverted.

A different kind of violation involves `c5` and `nBayes`. In the recommended ranking they are tied in the 4th position but the ideal ranking contains the restriction `c5` \gg `nBayes`. We consider this to be a less important kind of violation. In fact, as none of the algorithms is preferred, we may chose one randomly. If we do so, we expect that in half of the cases the right decision will be made (i.e. the significant difference is not violated) and in the other half, the choice will be incorrect (i.e. the significant difference is violated). Therefore, each violation of this kind contributes with $1/2$ to the score of the ranking.

In our example, the recommended ranking causes two violations of the first kind and one of the second, thus scoring $2 + 1/2 = 2.5$. This score contains little information on the quality of the ranking. Therefore, we calculate the proportion of significant differences violated. The ideal ranking for the `glass` dataset contains 11 significant differences, so the recommended ranking in question has a score of $2.5/11 = 0.2273$ or 22.73%.

Average Rank Correlation. Our second approach to building the ideal ranking relies on creating N total orders, by ordering the algorithms according to their accuracy on each fold of the cross-validation procedure. These orders can then be compared with the recommended ranking. We use Spearman's rank correlation coefficient, r_S , (Neave and Worthington (1992)) for this purpose. Given that n is the number of algorithms and $D_i = R_i - R_i^*$ is the distance between the recommended and ideal rank of algorithm i , $r_S = 1 - 6 * \sum_m D_i^2 / n(n-1)$.

We calculate the correlations for each of the folds and then average them to calculate the average correlation (AC). The AC score of the recommended ranking for the **glass** dataset mentioned above is presented in Table 2.

Coefficient	Fold										Average
	1	2	3	4	5	6	7	8	9	10	
r_s	0.50	0.27	0.50	0.69	0.76	0.07	0.47	0.76	0.50	0.56	0.51
r'_s	-3.85	-3.90	-12.12	-0.78	-0.71	-2.61	-1.58	0.30	-3.85	-1.05	-3.02

Table 2. AC and AWC scores and corresponding Spearman and weighted correlations for the recommended ranking **ltree**, **c5boost**, **discrim**, **nbayes/c5** (tied) and **timbl** when evaluated on the **glass** dataset.

Average Weighted Correlation. In the third measure, we adapted the calculation of the correlation coefficient to have lower correlation when a ranking method incorrectly ranks a “good” algorithm (that is, one which is highly ranked in the ideal ranking) rather than a “bad” one. This can be achieved by weighing the distance between the ideal and the recommended rank, D_i , in the calculation of Spearman’s rank correlation coefficient. We used the following weighing scheme: $D'_i = \frac{D_i * n}{R_i^*}$, where n is the number of algorithms and R_i^* is the ideal rank. The calculation of the weighted correlation coefficient, r'_s , is done in exactly the same manner as before.

Once again, we calculate the correlation for each fold and then average the values obtained. The AWC score of the recommended ranking presented above when evaluated on the **glass** dataset is presented in Table 2.

3 Discussion

The evaluation measures presented earlier have advantages and disadvantages. Here we do a preliminary analysis grouped according to three issues: accuracy estimation, error importance and intuitiveness.

Accuracy estimation. Building an ideal ranking based on pairwise significant differences between algorithms, as in PSDV, seems an obvious way to avoid the problem of unreliable accuracy estimation. However, one must not forget that the statistical tests underlying this measure can incur Type I and Type II errors (Dietterich (1998)). Furthermore, this problem is aggravated when multiple comparisons are performed (Salzberg (1997)).

In both correlation-based measures, we use the performance information in the folds to minimize the problem of accuracy estimation. To analyze whether we have succeeded or not, we have used the results obtained on the **hepatitis** dataset. On this dataset, the six algorithms considered have

obtained results which are not significantly different according to paired t tests with a 5% significance level. Therefore, the difference between the AC scores of any pair of recommended rankings should be small. In Table 3, we present the evaluation of a ranking obtained by ordering the average error rates (ranking 1), `lmtree` (17.9%), `c5` (18.1%), `nbayes` (18.6%), `c5boost` (19.2%), `timbl` (21.8%) and `discrim` (23.0%), and also the ranking obtained by reversing that order (ranking 2). We observe that the difference is not negligible, although we must be aware that the tests which concluded that there are no differences can incur into errors, as mentioned above.

ranking	Fold										Average
	1	2	3	4	5	6	7	8	9	10	
1	0.50	-0.30	0.39	0.33	0.43	0.61	-0.26	-0.20	0.49	0.01	0.20
2	-0.47	0.33	-0.36	-0.30	-0.43	-0.59	0.26	0.20	-0.59	0.01	-0.29

Table 3. AC scores for two very different rankings on `hepatitis`, where there are no significant differences between the algorithms.

Error importance. The importance of an error is influenced by the position of the algorithm in the ideal ranking and by the distance between the recommended and the ideal rank. PSDV does not take into consideration any of these. Spearman's correlation coefficient, on which both AC and AWC are based, uses the difference between the recommended and the ideal ranks. Therefore, the larger the distance, the lower the correlation and, thus, the score of the ranking. Finally, only AWC takes the ideal rank into account. For example, the AC and AWC scores of the ranking obtained by swapping the 5th and 6th algorithms in an ideal ranking involving 6 items are 0.943 and 0.930, respectively. On the other hand, if we swap the first and the second algorithms, the AC score remains the same but AWC decreases considerably (-0.286).

Intuitiveness. In PSDV, we calculate the proportion of cases for which the recommended ranking conflicts with the ideal one. This is an intuitive measure because if there is a large number of conflicts then the ranking is certainly not a good one, and vice-versa.

As for AC, the values for Spearman's correlation coefficient lie in the interval between -1 and +1, where -1 represents perfect disagreement and 1, perfect agreement. A coefficient of 0 indicates that there is no correlation between the rankings. Although the statistical significance depends on the critical value for the given significance level, we can still compare two coefficients to assess which one represents a higher correlation.

We were not able to determine the distribution of r'_s , as opposed to what happens with r_s . This is due to the changes we have made. Therefore we

cannot test the significance of the weighted correlation. This makes the interpretation of the AWC score harder, although we expect it to behave much like AC except that the results are harder to interpret.

4 Conclusions

We have presented three measures to evaluate rankings of classification algorithms. We have discussed their characteristics and raised interesting issues that need further analysis. Some important improvements are:

- To reduce the problems of incorrect decisions associated with the ideal ranking used in the proportion of significant differences violated, we could use a different hypothesis test, e.g. McNemar's (Dietterich (1998)).
- Although Spearman's correlation coefficient seems to be well suited for the purpose of measuring the distance between two rankings, we could try Kendall's rank correlation coefficient instead.
- Further investigation should be done on the effect of non-significant differences on the correlation-based measures and on the properties of the weighted correlation coefficient. In particular, it is important to normalize AWC and determine its distribution in order to be able to test its statistical significance.

References

- BRAZDIL, P. and SOARES C. (2000): A Comparison of Ranking Methods for Classification Algorithm Selection. To be published in: *Proceedings of the European Conference on Machine Learning*.
- GAMA, J. and BRAZDIL, P. (1995): Characterization of Classification Algorithms. In: Pinto-Ferreira, C. and Mamede, N. (Eds.): *Progress in Artificial Intelligence*. Springer-Verlag, 189–200.
- DIETTERICH, T.G (1998): Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10, 7, 1895–1924 (<ftp://ftp.cs.orst.edu/pub/tgd/papers/nc-stats.ps.gz>).
- NAKHAEIZADEH, G. and SCHNABL, A. (1997): Development of Multi-Criteria Metrics for Evaluation of Data Mining Algorithms. In: D. Heckerman and H. Mannila and D. Pregibon and R. Uthurusamy (Eds.): *Proceedings of the Third International Conference on Knowledge Discovery in Databases*. AAAI Press, 37–42.
- NEAVE, H.R. and WORTHINGTON, P.L. (1992): *Distribution-Free Tests*. Routledge.
- SALZBERG, S.L. (1997): On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach. *Data Mining and Knowledge Discovery*, 1, 317–327 (<http://www.cs.jhu.edu/~salzberg/critique.ps>).
- SOARES, C.P. (1999): *Ranking Classification Algorithms on Past Performance*. M.Sc. Thesis, Faculty of Economics, University of Porto (http://www.ncc.up.pt/~csoares/miac/thesis_revised.zip).

A General Approach to Test the Pertinence of a Consensus Classification

Guy Cucumel¹ and François-Joseph Lapointe²

¹ École des sciences de la gestion, Université du Québec à Montréal,
C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8
(e-mail: cucumel.guy@uqam.ca)

² Département de sciences biologiques, Université de Montréal.
C.P. 6128, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3J7

Abstract. Many techniques have been proposed to combine classifications defined on the same set of objects. All the methods that have been developed are designed to return a solution, but validation of the solution is seldom performed. In this paper we propose a general approach to test the pertinence of a consensus classification and discuss the choices that one has to make at each step of the method.

1 Introduction

Given a profile $\mathbf{P} = (C_1, \dots, C_i, \dots, C_k)$ of k classifications (e.g., n-trees, dendograms, pyramids) defined on a common set of n objects \mathbf{S} , how to obtain a single consensus solution that summarizes the entire profile \mathbf{P} ? Several methods and algorithms to combine classifications have been developped during the last decades (see Leclerc (1998)). Because a given consensus method will always return a solution, even from random classifications, there needs to be a way to assess the pertinence of the consensus solutions. In this paper, we propose a general algorithm for testing such consensus classifications.

2 Defining pertinence

A consensus classification can be derived from any profile \mathbf{P} defined on \mathbf{S} . However, if \mathbf{P} only contains random classifications, the consensus would not be resolved (i.e., structured). A consensus solution will be usefull iff it is representative of the initial profile \mathbf{P} (i.e., iff it summarizes well \mathbf{P}). The more random are the classifications combined, the less structured the consensus is likely to be. Thus, we propose to assess the pertinence of a consensus by comparing the fit between this consensus and the classifications in the initial profile \mathbf{P} to a distribution of "fit-values" for consensus derived from randomly generated classifications (Lapointe and Legendre (1991)). A consensus of k classifications will be declared pertinent iff it is closer from \mathbf{P} than the vast majority of consensus solutions (e.g., 95%) derived with the same method from profiles of k random classifications defined on \mathbf{S} (Cucumel and Lapointe

(1998)). Some methods, like the average consensus, will always return an optimal solution according to a least-squares criterion. But the question remains: is this solution a pertinent one?

3 A general algorithm

The algorithm to assess the pertinence of a consensus classification relies on a randomization approach that can be adapted to different problems. The general method proceeds as follows:

Step 1: input the k classifications from a given profile \mathbf{P} .

Step 2: compute the consensus classification.

Step 3: compute a statistic to measure the fit between the consensus and the profile.

Step 4: generate a profile \mathbf{P}' of k random classifications defined on \mathbf{S} .

Step 5: repeat steps 2 to 4 a large number of times (e.g. 1000 times) for \mathbf{P}' .

Step 6: compute the pertinence of the consensus classification under the null hypothesis. This is accomplished by comparing the actual value of the test statistic to a distribution of that statistic obtained from randomly generated classifications. With a risk of α the null hypothesis is rejected if the test statistic is more extreme than the vast majority (say 95% for $\alpha=0.05$) of the statistic based on random classifications.

In the following sections, we will see how this procedure can be applied to (1) different types of classifications, (2) different kinds of consensus methods, (3) different test statistics, and (4) different models for generating random classifications. Obviously, some of these options are only available for specific types of classifications. More importantly, one must always bear in mind that the randomization models must be compatible with the classifications combined (e.g., a model for generating rooted trees should not be used to compare unrooted ones).

3.1 Different type of classification

The most common type of classification is a hierarchy, represented in the form of a rooted and terminally-labeled tree. Such labeled trees differ from unlabeled ones in that labels are assigned (or not) to its nodes to refer to a set of objects \mathbf{S} ; in terminally-labeled trees, only the terminal nodes are labeled. A tree is said to be rooted (as opposed to unrooted) when one of its nodes is labeled as the root to induce a direction on the branches of the tree. If these branches have lengths associated to them to represent the amount of change between the nodes, the trees can be depicted in the form of weighted trees. When every terminal node is equidistant from the root, the corresponding weighted tree is said to be ultrametric and its path-lengths satisfy the ultrametric inequality (Hartigan (1967)). On the other hand, weighted trees that do not meet the ultrametric property still remain additive when they satisfy the four-point condition (Buneman (1971)). Interestingly, it is equivalent

when computing consensus trees to deal with the graphical representations or their associated path-length matrices. The approach we proposed in this paper applies to all those classifications as well as others not represented in the form of trees, like pyramids (Bertrand and Diday (1985)) and weak-hierarchies (Bandelt and Dress (1989)).

3.2 Different consensus methods

Depending on the type of classifications considered, several consensus methods can be defined. Some techniques, including the strict (Sokal and Rohlf (1981)), semi-strict (Bremer (1990)), median (Barthélemy and McMorris (1986)), and majority-rule consensus (Margush and McMorris (1981)) have been developed to combine unweighted trees (see also Adams (1972), McMorris et al. (1983), Neumann (1983)). However, when branch lengths are of interest, specific consensus techniques require to be used to deal with weighted trees (Lapointe (1998b)). Stinebrickner (1984), Lefkovitch (1985) and Brossier (1990) have developed consensus techniques to combine ultrametric trees (i.e. dendograms), whereas Lapointe and Cucumel (1997) have proposed a consensus procedure to account for other types of weighted trees. All these techniques (for weighted or unweighted trees) are restricted to consensus classifications defined on a common set of objects \mathbf{S} . Other methods are to be used for the construction of consensus supertrees from classifications bearing overlapping sets of taxa (Gordon (1986), Baum (1992), Steel (1992)), or the computation of common pruned trees (Finden and Gordon (1985)) and reduced consensus trees (Wilkinson (1994)). Finally, one could also use special consensus methods to combine classifications presented in the form of pyramids or weak-hierarchies (McMorris and Powers (1991)).

3.3 Different test statistics

When trees are considered, it is always necessary to distinguish topological indices from tree metric indices (see Lapointe (1998a)); the former are designed for unweighted trees, whereas the latter are for weighted-tree comparisons. Test statistics available for topological comparisons include the partition metric (Robinson and Foulds (1981)), the neighborhood interchange metric (Robinson (1971)) and the quartet metric (Estabrook et al. (1985)), among many others (Bosibud and Bosibud (1972), Margush (1982), Day (1983), Penny and Hendy (1985), Steel (1988)). When path-lengths matrices need to be compared, modified versions of the topological indices can be used (Robinson and Foulds (1979)), in addition to specific indices designed for dendograms (Sokal and Rohlf (1962), Fowlkes and Mallows (1983)), or for any weighted trees (Steel and Penny (1993)). For other test statistics used as consensus indices, the reader should consult Rohlf (1982).

3.4 Different models of random classifications

Any classification can be decomposed into three distinct parts (Lapointe and Legendre (1990)): topology, label positions, and branch lengths (if any). The topological part represents the shape of the classification. The labels attached to the nodes refer to different objects and their relative positions provide information about their relationships. The branch lengths represent the amount of change (relative or absolute) among different objects. Depending on the criterion selected, classifications can differ in terms of topological relationships, label positions, or branch lengths. Consequently, different random-tree generation algorithms are distinguishable (see Furnas (1984), Oden and Shao (1984), Quiroz (1989), Lapointe and Legendre (1991)) depending on the type of classifications compared (e.g., labeled or unlabeled, weighted or unweighted). The sampling distribution of the classifications is also important in the computations. In the case of rooted trees, three distribution models are usually defined (Simberloff et al. (1981), Lapointe and Legendre (1995)). The first and simplest model is to generate every topology equiprobably. In the second model, each tree is equally likely to be generated. The third model implies that every branching point is equally likely when growing a tree; interestingly, dendograms can be generated equiprobably under this so-called Markovian model (Lapointe and Legendre (1991)). As mentionned above, it is important to use a generation model that corresponds to the type of classifications in the profile **P** to test the correct null hypothesis.

4 Conclusion

It probably will be difficult to generate tables of statistical significance (or pertinence tables) because of the large number of different parameters to take into account in the model. A computer program that can perform the test in a wide range of cases will have to be developed. This method could be extended to supertrees and to multiple consensus classifications as well.

References

- ADAMS, E.N., III. (1972): Consensus Techniques and the Comparison of Taxonomic Trees. *Systematic Zoology*, 21, 390–397.
- BANDELT, H.-J. and DRESS, A.W.M. (1989): Weak Hierarchies Associated with Similarity Measures: an Additive Clustering Technique. *Bulletin of Mathematical Biology*, 51, 133–166.
- BARTHÉLEMY, J.-P. and McMORRIS, F.R. (1986): The Median Procedure for n-Trees. *Journal of Classification*, 3, 329–334.
- BAUM, B.R. (1992): Combining Trees as a Way of Combining Data for Phylogenetic Inference, and the Desirability of Combining Gene Trees. *Taxon*, 41, 3–10.
- BERTRAND, P. and DIDAY, E. (1985): A Visual Representation of the Compatibility between an Order and a Dissimilarity Index: the Pyramids. *Computational Statistics Quarterly*, 2, 31–42.

- BOSIBUD, H.M. and BOSIBUD, L.E. (1972): A Metric for Classifications. *Taxon*, 21, 607–613.
- BREMER, K. (1990): Combinable Component Consensus. *Cladistics*, 6, 369–372.
- BROSSIER, G. (1990): Piecewise Hierarchical Clustering. *Journal of Classification*, 7, 197–216.
- BUNEMAN, P. (1971): The Recovery of Trees From Measures of Dissimilarity. In: F.R. Hudson, D.G. Kendall and P. Tautu (Eds.): *Mathematics in Archeological and Historical Sciences*. Edinburgh University Press, Edinburgh, 387–395.
- CUCUMEL, G. and LAPOINTE, F.-J. (1998): Assessing the Pertinence of a Consensus with Permutations. *Short Papers of the VI Conference of the IFCS*. Istituto Nazionale di Statistica, Roma, 89–91.
- DAY, W.H.E. (1983): Distributions of Distances Between Pairs of Classifications. In: J. Felsenstein (Ed.): *Numerical Taxonomy*. Springer-Verlag, Berlin, 127–131.
- ESTABROOK, G.F., McMORRIS, F.R. and MEACHAM, C. (1985): Comparison of Undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology*, 34, 193–200.
- FINDEN, C.R. and GORDON, A.D. (1985): Obtaining Common Pruned Trees. *Journal of Classification*, 2, 225–276.
- FOWLKES, E.B. and MALLOWS, C.L. (1983): A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78, 553–569.
- FURNAS, G.W. (1984): The Generation of Random, Binary Unordered Trees. *Journal of Classification*, 1, 187–233.
- GORDON, A.D. (1986): Consensus Supertrees: The Synthesis of Rooted Trees Containing Overlapping Sets of Labeled Leaves. *Journal of Classification*, 3, 335–348.
- HARTIGAN, J.A. (1967): Representation of Similarity Matrices by Trees. *Journal of the American Statistical Association*, 62, 1140–1148.
- LAPOINTE, F.-J. (1998a): How to Validate Phylogenetic Trees? A Stepwise Procedure. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock and Y. Baba (Eds.): *Data Science, Classification, and Related Methods*. Springer-Verlag, Tokyo, 71–88.
- LAPOINTE, F.-J. (1998b): For Consensus (with Branch Lengths). In: A. Rizzi, M. Vichi and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer-Verlag, Berlin, 73–80.
- LAPOINTE, F.-J. and CUCUMEL, G. (1997): The Average Consensus Procedure: Combination of Weighted Trees Containing Identical or Overlapping Sets of Objects. *Systematic Biology*, 46, 306–312.
- LAPOINTE, F.-J. and LEGENDRE, P. (1990): A Statistical Framework to Test the Consensus of Two Nested Classifications. *Systematic Zoology*, 39, 1–13.
- LAPOINTE, F.-J. and LEGENDRE, P. (1991): The Generation of Random Ultrametric Matrices Representing Dendograms. *Journal of Classification*, 8, 177–200.
- LAPOINTE, F.-J. and LEGENDRE, P. (1995): Comparison Tests for Dendograms: A Comparative Evaluation. *Journal of Classification*, 12, 265–282.
- LECLERC, B. (1998): Consensus of Classifications: the Case of Trees. In: A. Rizzi, M. Vichi and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer-Verlag, Berlin, 81–90.

- LEFKOVITCH, L.P. (1985): Euclidean Consensus Dendograms and Other Classification Structures. *Mathematical Biosciences*, 74, 1–15.
- MARGUSH, T. (1982): Distances Between Trees. *Discrete Applied Mathematics*, 4, 281–290.
- MARGUSH, T. and McMORRIS, F.R. (1981): Consensus n-Trees. *Bulletin of Mathematical Biology*, 43, 239–244.
- McMORRIS, F.R., MERONK, D.B. and NEUMANN, D.A. (1983): A View of Some Consensus Methods for Trees. In: J. Felsenstein (Ed.): *Numerical Taxonomy*. Springer-Verlag, Berlin, 122–126.
- McMORRIS, F.R. and POWERS, R.C. (1991): Consensus Weak Hierarchies. *Bulletin of Mathematical Biology*, 53, 679–684.
- NEUMANN, D.A. (1983): Faithful Consensus Methods for n-Trees. *Mathematical Biosciences*, 63, 271–287.
- ODEN, N.L. and SHAO, K.T. (1984): An Algorithm to Equiprobably Generate all Directed Trees With k Labeled Terminal Nodes and Unlabeled Interior Nodes. *Bulletin of Mathematical Biology*, 46, 379–387.
- PENNY, D. and HENDY, M.D. (1985): The Use of Tree Comparison Metrics. *Systematic Zoology*, 34, 75–82.
- QUIROZ, A.J. (1989): Fast Random Generation of Binary, t-ary and Other Types of Trees. *Journal of Classification*, 6, 223–231.
- ROBINSON, D.F. (1971): Comparison of Labeled Trees With Valency Three. *Journal of Combinatorial Theory*, 11, 105–119.
- ROBINSON, D.F. and FOULDS, L.R. (1979): Comparison of Weighted Labeled Trees. In: C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock and Y. Baba (Eds.): *Lecture Notes in Mathematics*, Volume 748. Springer-Verlag, Berlin, 119–126.
- ROBINSON, D.F. and FOULDS, L.R. (1981): Comparison of Phylogenetic Trees. *Mathematical Biosciences*, 53, 131–147.
- ROHLF, F.J. (1982): Consensus Indices for Comparing Classifications. *Mathematical Biosciences*, 59, 131–144.
- SIMBERLOFF, D., HECK, K.L., MCCOY, E.D. and CONNOR, E.F. (1981): There Have Been no Statistical Tests of Cladistics Biogeographical Hypotheses. In: G. Nelson and D.E. Rosen (Eds.): *Vicariance Biogeography: A Critique*. Columbia University Press, New York, 40–63.
- SOKAL R.R. and ROHLF, F.J. (1962): The Comparison of Dendograms by Objective Methods. *Taxon*, 9, 33–40.
- SOKAL R.R. and ROHLF, F.J. (1981): Taxonomic Congruence in the Lepidoptera Re-examined. *Systematic Zoology*, 30, 309–325.
- STEEL, M.A. (1988): Distribution of the Symmetric Difference Metric on Phylogenetic Trees. *SIAM Journal of Discrete Mathematics*, 1, 541–555.
- STEEL, M.A. (1992): The Complexity of Reconstructing Trees From Qualitative Characters and Subtrees. *Journal of Classification*, 1, 91–116.
- STEEL, M.A. and PENNY, D. (1993): Distribution of Tree Comparison Metrics—Some New Results. *Systematic Biology*, 42, 126–141.
- STINEBRICKNER, R. (1984): An Extension of Intersection Methods From Trees to Dendograms. *Systematic Zoology*, 33, 381–386.
- WILKINSON, M. (1994): Common Cladistic Information and its Consensus Representation: Reduced Adams and Reduced Cladistic Consensus Trees and Profiles. *Systematic Biology*, 43, 343–368.

On a Class of Aggregation-invariant Dissimilarities Obeying the Weak Huygens' Principle

F. Bavaud

Section d'Informatique et de Méthodes Mathématiques,
Faculté des Lettres, Université de Lausanne
CH-1015 Lausanne-Dorigny
(e-mail: Francois.Bavaud@imaa.unil.ch)

Abstract. We propose a complete characterization of a certain class of aggregation-invariant dissimilarities between row (or column) profiles. This class (for which row and column dispersions coincide) contains the chi-square, ratio, Kullback-Leibler, Hellinger, Cressie-Read dissimilarities, as well as a presumably new “type *s*” class of dissimilarities. Distinguishing between two forms of Huygens’ principle from Classical Mechanics, we show “type *s*” dissimilarities to satisfy the weak Huygens’ principle; the strong Huygens’ principle however holds for a single member of the class, namely the chi-square dissimilarity. Extending the concept of dissimilarity to “type *s*” divergences restores the strong principle.¹

1 Introduction and notations

Let n_{jk} be an $(J \times K)$ contingency table, with relative frequency $f_{jk} := n_{jk}/n$, row profiles $w_{jk} := n_{jk}/n_{j\bullet}$, column profiles $w_{kj}^* := n_{jk}/n_{\bullet k}$ and marginal (strictly positive) profiles $\rho_j^* := n_{j\bullet}/n = f_{j\bullet}$ and $\rho_k := n_{\bullet k}/n = f_{\bullet k}$, where $n := n_{\bullet\bullet}$ is the grand total. By construction, $f_{jk} = \rho_j^* w_{jk} = \rho_k w_{kj}^*$; also, the row and column profiles transform as $w_{jk} = \rho_k w_{kj}^*/\rho_j^*$ and $w_{kj}^* = \rho_j^* w_{jk}/\rho_k$.

A dissimilarity D between J objects is a symmetric $(J \times J)$ non-negative matrix $D_{jj'}$ with a null diagonal². The so-called chi-square dissimilarity between rows j and j' , defined as $D_{jj'}^\chi := \sum_k \frac{1}{\rho_k} (w_{jk} - w_{j'k})^2$ is well known to be aggregation invariant (see definition 1). Furthermore, the chi-square measure of row/column dependence obtains as half the weighted average of the dissimilarity between each pairs of rows, or *equivalently*, as the weighted average of the dissimilarity between each row and the average profile (the latter constituting the canonical measure of *row dispersion*):

$$\frac{\chi^2}{n} = \frac{1}{2} \sum_{jj'} \rho_j^* \rho_{j'}^* D_{jj'}^\chi = \sum_j \rho_j^* D_{j\bullet}^\chi \quad \text{where } D_{j\bullet}^\chi := \sum_k \frac{1}{\rho_k} (w_{jk} - a_k)^2 \quad (1)$$

¹ The author gratefully acknowledges the detailed remarks and helpful suggestions of an anonymous referee.

² Throughout this paper, $D_{jj'}$ stands for what might be denoted as $D_{jj'}^2$ in more traditional expositions.

The equivalence emphasized above constitutes the *weak Huygens' principle*. In the present case, the *strong Huygens' principle* $\sum_j \rho_j^* D_{ja}^\chi = \sum_j \rho_j^* D_{j\rho}^\chi + D_{a\rho}^\chi$ also holds (see definition 3).

In general, the strong Huygens' principle enables to control the dependence of the total dissimilarity with respect to some reference object a , as exemplified by the determination of rotational inertia in classical mechanics; on the other hand, the weak Huygens' principle suffices in defining a local dissimilarity by splitting the double summation into a summation restricted on pairs satisfying a given relation (such as a contiguity relation in local variance formulations; see e.g. Lebart (1969)) and its complementary.

This contribution (see Bavaud (2000) for further emphasis upon Information Theory, hypothesis testing and factorial data analysis) extends some of the above properties to other dissimilarity functionals. We first define necessary and sufficient conditions for aggregation-invariance applied to a fairly broad family of dissimilarities, containing the chi-square, Kullback-Leibler, Cressie-Read, Neyman and Hellinger similarities as particular cases. Such dissimilarities possess identical row and column dispersion measures. We then turn to the issue of the Huygens' principles, and exhibit an apparently new "type s " class of dissimilarities satisfying the weak principle, but not the strong one (with the noticeable exception of D^χ). We finally extend the concept of dissimilarity to (possibly non symmetric and non positive) divergences, and show the strong Huygens' principle to hold for "type s " divergences. Theorems 1 to 4, as well as the concept of *type of a dissimilarity*, seem original.

2 Main results

Definition 1 A dissimilarity D is **aggregation invariant** if its value $D_{jj'}$ remains unchanged when two identical profiles $w_{k_1 j}^* = w_{k_2 j}^*$ associated to distinct columns k_1 and k_2 are further aggregated into a single column denoted $[k_1 \cup k_2]$ yielding the same profile $w_{[k_1 \cup k_2] j}^*$.

Theorem 1. For $|K| \geq 3$, any dissimilarity of the form

$$D_{jj'} := \sum_{k \in K} G(\rho_k) F(w_{jk}, w_{j'k}) \quad (2)$$

(where $F(x, x') \geq 0$ is defined for $x > 0$, $x' > 0$, continuous, symmetric with $F(x, x) = 0$ and $G(y) > 0$ is defined for $y > 0$, continuous) is aggregation invariant iff the scaling relation $G(\epsilon y) F(\epsilon x, \epsilon x') = \epsilon G(y) F(x, x')$ holds for any $y, x, x', \epsilon > 0$.

Proof: D of the form (2) is aggregation invariant iff

$$G(\rho_{k_1}) F(w_{jk_1}, w_{j'k_1}) + G(\rho_{k_2}) F(w_{jk_2}, w_{j'k_2}) = G(\rho_{[k_1 \cup k_2]}) F(w_{j[k_1 \cup k_2]}, w_{j'[k_1 \cup k_2]})$$

for all $j \neq j'$, whenever $w_{k_1 j}^* = w_{k_2 j}^* =: a_j$ for all j . Using $w_{jk_1} = \rho_{k_1} a_j / \rho_j^*$, $w_{jk_2} = \rho_{k_2} a_j / \rho_j^*$, $w_{j[k_1 \cup k_2]} = (\rho_{k_1} + \rho_{k_2}) a_j / \rho_j^*$ for all j as well as $\rho_{[k_1 \cup k_2]} = \rho_{k_1} + \rho_{k_2}$, and defining $b_j := a_j / \rho_j^*$ shows the previous condition to amount to

$$\begin{aligned} G(\rho_{k_1}) F(\rho_{k_1} b_j, \rho_{k_1} b_{j'}) + G(\rho_{k_2}) F(\rho_{k_2} b_j, \rho_{k_2} b_{j'}) = \\ G(\rho_{k_1} + \rho_{k_2}) F((\rho_{k_1} + \rho_{k_2}) b_j, (\rho_{k_1} + \rho_{k_2}) b_{j'}) \end{aligned}$$

If $|K| \geq 3$, the quantities $\rho_{k_1} > 0$ and $\rho_{k_2} > 0$ can be varied independently, and independently of the quantities b_j and $b_{j'}$ which can be kept constant. Defining $H(\rho) := G(\rho) F(\rho b_j, \rho b_{j'})$, aggregation invariance thus holds iff $H(\rho_{k_1}) + H(\rho_{k_2}) = H(\rho_{k_1} + \rho_{k_2})$, i.e. iff $H(\rho) = \alpha_{jj'} \rho$ by continuity, where $\alpha_{jj'} > 0$ still depends upon b_j and $b_{j'}$. Setting $y := \rho$, $x := \rho b_j$ and $x' := \rho b_{j'}$ yields $H(\rho) = G(y) F(x, x')$ and $H(\epsilon\rho) = G(\epsilon y) F(\epsilon x, \epsilon x')$. The identity $H(\epsilon\rho) = \alpha_{jj'} \epsilon \rho = \epsilon H(\rho)$ thus entails the scaling relation of theorem 1. Conversely, it is straightforward to show the latter (with $\epsilon =: \rho$, $y = 1$, $x =: b_j$, $x' =: b_{j'}$ and $\alpha_{jj'} := G(1)F(x, x')$) to imply $H(\rho) = \alpha_{jj'} \rho$.

The chi-square dissimilarity D^χ fits into (2) with $G(y) = y^{-1}$ and $F(x, x') = (x - x')^2$. Fichet (1978) proves $G(y) = c y^{-1}$ to be the only solution insuring aggregation invariance in the particular case $F(x, x') = (x - x')^2$, even when the continuity assumption on $G(y)$ is dropped and the values of x , x' and y are restricted to rational ones, as it is the case with empirical contingency tables.

More generally, the scaling condition $F(\epsilon x, \epsilon x') = \epsilon^{t+1} F(x, x')$ and $G(\epsilon y) = \epsilon^{-t} G(y)$ insures the aggregation invariance, where we shall refer to t as the *type* of the dissimilarity. Among members of this family, one finds the:

type -1 ratio dissimilarity with $F(x, x') = \frac{x}{x'} + \frac{x'}{x} - 2$ and $G(y) = y$.

type 0 symmetrized “generalized power divergence” dissimilarity (Cressie and Read (1984)) with $F^\lambda(x, x') = \frac{1}{2\lambda(\lambda+1)} [x^{\lambda+1} x'^{-\lambda} + x'^{\lambda+1} x^{-\lambda} - x - x']$ and $G^\lambda(y) = 1$; the invariance $\lambda \rightarrow -\lambda - 1$ permits to restrict to values $\lambda \geq -1/2$. In particular one recovers the Hellinger or “Freeman-Tuckey-like” dissimilarity (Escoffier (1978); Cressie and Read (1984)) with $F(x, x') = 2(\sqrt{x} - \sqrt{x'})^2$ (for $\lambda = -1/2$), the symmetrized Kullback-Leibler dissimilarity (Kullback (1959)) with $F(x, x') = \frac{1}{2}(x - x') \ln(x/x')$ (for $\lambda = 0$, using $\lim_{\lambda \rightarrow 0} (a^\lambda - 1)/\lambda = \ln a$ for $a > 0$), the dissimilarity $F(x, x') = \frac{2}{3}(\sqrt{x} - \sqrt{x'})(\frac{x}{\sqrt{x'}} - \frac{x'}{\sqrt{x}})$ (for $\lambda = 1/2$) and the “Neyman-like dissimilarity” $F(x, x') = \frac{1}{4}[\frac{(x-x')^2}{x'} + \frac{(x'-x)^2}{x}]$ (for $\lambda = 1$).

type 1 chi-square dissimilarity D^χ .

type s “type s ” dissimilarity with $F_s(x, x') = \frac{1}{s}(x^s - x'^s)(x - x')$ and $G_s(y) = y^{-s}$; in particular one recovers the ratio dissimilarity (for $s = -1$), (twice) the symmetrized Kullback-Leibler dissimilarity (i.e. $F_0(x, x') = (x - x') \ln(x/x')$ for $s \rightarrow 0$) and the chi-square dissimilarity (for $s = 1$).

In equation (1) we actually extended the concept of dissimilarities $D_{jj'}$ between rows $j, j' \in J$ of a contingency table $\{\rho_j^*, w_{jk}\}$ to the more general concept of dissimilarities between profiles, in the sense of

Definition 2 values $D_{aa'}$ between arbitrary normalized profiles a, a' ($a_k \geq 0$, $a'_k \geq 0$, $\sum_k a_k = \sum_k a'_k = 1$) such that $D_{aa'} = D_{jj'}$ whenever $a_k = w_{jk}$ and $a'_k = w_{j'k}$ for all $k \in K$.

Definition 3 The row dispersion is defined as the weighted average of the profile dissimilarity between each row and the average profile, namely $\sum_{j \in J} \rho_j^* D_{j\rho}$.

All dissimilarities of the form (2) are profile dissimilarities. Under aggregation invariance, row and column dispersions coincide:

Theorem 2. The row and column dispersions associated to aggregation invariant profile dissimilarities of the form (2) are identical:

$$\sum_{j \in J} \rho_j^* D_{j\rho} = \sum_{k \in K} \rho_k D_{k\rho^*}^*$$

where D^* is the column dissimilarity dual to the row dissimilarity D of (2), namely

$$D_{kk'}^* := \sum_{j \in J} G(\rho_j^*) F(w_{kj}^*, w_{k'j}^*)$$

Proof: the scaling relation $G(\epsilon y) F(\epsilon x, \epsilon x') = \epsilon G(y) F(x, x')$ of theorem 1 yields

$$\begin{aligned} & \sum_{k \in K} \rho_k D_{k\rho^*}^* = \sum_k \rho_k \sum_j G(\rho_j^*) F(w_{kj}^*, \rho_j^*) = \\ &= \sum_{j,k} \rho_k G\left(\frac{\rho_j^*}{\rho_k} \rho_k\right) F\left(\frac{\rho_j^*}{\rho_k} w_{jk}, \frac{\rho_j^*}{\rho_k} \rho_k\right) = \sum_{j,k} \rho_k \frac{\rho_j^*}{\rho_k} G(\rho_k) F(w_{jk}, \rho_k) = \sum_{j \in J} \rho_j^* D_{j\rho} \end{aligned}$$

Definition 4 Huygens' principles relative to a profile dissimilarity D are

$$\sum_j \rho_j^* D_{ja} = \sum_j \rho_j^* D_{j\rho} + D_{\rho a} \quad \text{strong Huygens' principle} \quad (3)$$

$$\sum_{jj'} \rho_j^* \rho_{j'}^* D_{jj'} = 2 \sum_j \rho_j^* D_{j\rho} \quad \text{weak Huygens' principle} \quad (4)$$

where a_k is any normalized profile, and $\rho_k = \sum_j \rho_j^* w_{jk}$ is the average profile.

Writing $a_k = w_{j'k}$ and applying the weighted average operator $\sum_{j'} \rho_{j'}^* \dots$ shows the strong formulation (3) to entail the weak one (4).

Theorem 3. “Type s ” dissimilarities satisfy the weak Huygens’ principle (4), for any s .

Proof: define $F_{jj';k} := F_s(w_{jk}, w_{j'k})$. Then, for $s \neq 0$ (the proof for the symmetrized Kullback-Leibler case $s = 0$ is analogous to the proof following theorem 4 below),

$$\begin{aligned} \sum_{jj'} \rho_j^* \rho_{j'}^* F_{jj';k} &= \frac{1}{s} \sum_{jj'} \rho_j^* \rho_{j'}^* (w_{jk}^s - w_{j'k}^s)(w_{jk} - w_{j'k}) = \\ &= \frac{2}{s} \sum_j \rho_j^* w_{jk}^{s+1} - \frac{2}{s} \sum_j \rho_j^* w_{jk}^s \rho_k = \frac{2}{s} \sum_j \rho_j^* [w_{jk}^{s+1} - w_{jk}^s \rho_k] = \\ &= \frac{2}{s} \sum_j \rho_j^* [w_{jk}^{s+1} - w_{jk}^s \rho_k + \rho_k^{s+1} - \rho_k^s w_{jk}] = 2 \sum_j \rho_j^* F_{j\rho;k} \end{aligned}$$

As a result, any dissimilarity of the form $D_{jj'} = \sum_k G(\rho_k) F_{jj';k}$ satisfies the weak Huygens’ principle. By theorem 1, $G(\rho_k)$ has to be proportional to ρ_k^{-s} if $D_{jj'}$ is further required to be aggregation invariant. Thus, as claimed, “type s ” dissimilarities constitute an aggregation-invariant class obeying the weak Huygens’ principle. By contrast, generalized power divergence dissimilarities do not obey the weak Huygens’ principle, unless $\lambda = 0$.

On the other hand, treating the cases $s \neq 0$ and $s = 0$ separately, direct substitution in (3) shows “type s ” dissimilarities to fail to satisfy the strong Huygens’ principle, unless $s = 1$, i.e. unless $D = D^\lambda$.

In the last part of this contribution we consider more general *divergences* \tilde{D} matrix with a null diagonal, and such that $D_{jj'} := \tilde{D}_{jj'} + \tilde{D}_{j'j}$ is non-negative, i.e. constitutes a dissimilarity. Profile divergences are divergences whose values $\tilde{D}_{aa'}$ are defined for arbitrary normalized profiles a, a' , and coincide with $\tilde{D}_{jj'}$ when $a_k = w_{jk}$ and $a'_k = w_{j'k}$. A family of profile divergences is given by the “type s ” divergences

$$\tilde{D}_{aa'} := \frac{1}{s} \sum_{k \in K} \rho_k^{-s} a_k (a_k^s - a_{k'}^s)$$

The limit $s \rightarrow 0$ yields the familiar Kullback-Leibler divergence $K(a||a') := \sum_k a_k \ln \frac{a_k}{a'_k} \geq 0$. By construction, $D_{aa'} := \tilde{D}_{aa'} + \tilde{D}_{a'a} \geq 0$ is the “type s ” dissimilarity.

Definition 5 Huygens’ principles relative to a profile divergence \tilde{D} are

$$\sum_j \rho_j^* \tilde{D}_{ja} = \sum_j \rho_j^* \tilde{D}_{j\rho} + \tilde{D}_{\rho a} \quad \text{strong Huygens’ principle} \quad (5)$$

$$\sum_{jj'} \rho_j^* \rho_{j'}^* \tilde{D}_{jj'} = 2 \sum_j \rho_j^* \tilde{D}_{j\rho} \quad \text{weak Huygens’ principle} \quad (6)$$

This time, the weak principle (6) for \tilde{D} does not follow from the strong principle (5) for \hat{D} . However, the latter entails the weak principle (4) for the associated dissimilarity $D_{jj'} := \tilde{D}_{jj'} + \tilde{D}_{j'j}$. As a consequence, theorem 3 can be proved alternatively as following from theorem 4 below.

Theorem 4. “Type s” divergences satisfy the strong Huygens’ principle (5), for any s .

Proof: the case $s \neq 0$ is treated similarly to the proof of theorem 3 above. For $s = 0$,

$$\begin{aligned} \sum_j \rho_j^* \sum_k w_{jk} \ln \frac{w_{jk}}{a_k} &= \sum_j \rho_j^* \sum_k w_{jk} \ln \frac{w_{jk}}{\rho_k} + \sum_j \rho_j^* \sum_k w_{jk} \ln \frac{\rho_k}{a_k} = \\ &= \sum_j \rho_j^* \sum_k w_{jk} \ln \frac{w_{jk}}{\rho_k} + \sum_k \rho_k \ln \frac{\rho_k}{a_k} \end{aligned}$$

The above identity (showing in particular the minimum of $\sum_j \rho_j^* K(j||a)$ to be attained for $a_k = \rho_k$) has been reported by Jardine and Sibson (1971, p.13) in a different context.

References

- BAVAUD, F. (2000): An Information Theoretical approach to Factorial Correspondence Analysis. To appear in the proceedings of the 5th International Conference on the Statistical Analysis of Textual Data (JADT 2000)
- CRESSIE, N. and READ, T.R.C. (1984): Multinomial goodness-of-fit tests. *J.R.Statist.Soc.B*, 46, 440–464
- ESCOFIER, B. (1978): Analyse factorielle et distances répondant au principe d'équivalence distributionnelle. *Revue de Statistique Appliquée*, 26, 29–37
- FICHET, B. (1978): Note sur la métrique de l'analyse des correspondances. *Statistique et Analyse de Données*, 2, 87–93
- JARDINE,N. and SIBSON,R. (1971): *Mathematical Taxonomy*. Wiley, New York.
- KULLBACK, S. (1959): *Information Theory and Statistics*. Wiley, New York.
- LEBART, L. (1969): L'analyse statistique de la contiguïté. *Publications de l'ISUP*, XVIII, 81–112

A Short Optimal Way for Constructing Quasi-ultrametrics From Some Particular Dissimilarities

B. Fichet

Laboratoire de Biomathématiques.
Aix-Marseille II University.
13385 Marseille, Cedex 5, France
(e-mail : Bernard.Fichet@medecine.univ-mrs.fr)

Abstract. Recently, Diatta has established a lower maximal quasi-ultrametric approximation of a dissimilarity fulfilling the inclusion condition. The approach is purely algorithmical, but incidentally the solution is characterised by a formula. From this formula, we give here two straightforward and short proofs of the result. One is based on the properties of the dissimilarities under consideration, and the second one derives from the bijection between quasi-ultrametrics and indexed quasi-hierarchies.

1 Introduction

Quasi-ultrametricity has been introduced by Diatta and Fichet (1994, 1998) and, via a four-point characterisation, by Bandelt in an unpublished paper (1992) and Bandelt and Dress (1994). It unifies two extensions of ultrametricity, given by strongly-Robinsonian dissimilarities and semi-distances of tree-type.

Let us precise some notations. For a dissimilarity d on a finite set I , $B^d(i, r)$, $i \in I$, $r \geq 0$, is the (closed) ball of centre i and radius r , and for every $i, j \in I$, the 2-ball B_{ij}^d is $B^d(i, d(i, j)) \cap B^d(j, d(i, j))$. Again, for $J \subseteq I$, $\text{diam}_d(J)$ is the diameter of J , i.e. $\max\{d(i, j) \mid i, j \in J\}$.

Then, we recall that d is quasi-ultrametric if and only if it obeys the following two conditions.

$$\forall i, j \in I, \forall k, l \in B_{ij}^d, B_{kl}^d \subseteq B_{ij}^d \text{ (inclusion condition).}$$

$$\forall i, j \in I, \text{diam}_d(B_{ij}^d) = d(i, j) \text{ (diameter condition)}$$

Every dissimilarity obeying the inclusion condition, hence every quasi-ultrametric, is easily seen to be even, i.e. satisfying $d(i, j) = 0 \implies d(i, k) = d(j, k)$ for every k in I .

Bandelt and Dress (1989) introduced the weak clusters. As shown by Diatta and Fichet (1994), a subset H of I is a weak cluster, sensu Bandelt and Dress, if and only if for every i, j in H , $B_{ij}^d \subseteq H$. Moreover H always is a 2-ball. So, a dissimilarity d fulfils the inclusion condition whenever the set of 2-balls coincides with the set of weak clusters.

Quasi-ultrametrics are in bijection with indexed quasi-hierarchies, which extend hierarchical classification. Of course, a dissimilarity coefficient extracted from some data, does not generally fulfil the conditions required by quasi-ultrametricity. So, as usually, we need an approximation. To our knowledge, no optimal fitting in any sense has been established in a general framework. However, in the particular case of a given dissimilarity obeying the inclusion condition, Diatta (1998) exhibited a lower maximal approximation. His procedure is algorithmical, in the spirit of the famous descent algorithm yielding a subdominant for ultrametrics or k -ultrametrics. See Jardine and Sibson (1971). When we encounter a 2-ball B_{ij} violating the diameter condition, the deficient quantities are shrunk to the value $d(i, j)$. Diatta shows that the 2-balls are preserved and incidentally establishes a formula characterising the quasi-ultrametric arising from the algorithm.

In this note, we directly define the approximation by the formula. That provides a simpler and shorter way to justify the optimality, and even to deduce a more efficient algorithm. Our approach is two-fold. In the next paragraph we only use the basic definition and some immediate properties of quasi-ultrametricity. Then a second proof is proposed. It is much shorter, but requires the knowledge of the one-to-one correspondence with index quasi-hierarchies.

2 The optimal approximation from a formula.

The following lemma will be very useful in the sequel. We give a slightly different version of Diatta (1998), by partly relaxing the inclusion condition.

Lemma 1. *Let d be a dissimilarity on I and B_{ij}^d and B_{kl}^d be two weak clusters. Suppose that $k, l \in B_{ij}^d$ and $d(k, l) \geq d(i, j)$.*

Then $B_{ij}^d = B_{kl}^d$.

Proof. Since B_{ij}^d is a weak cluster, $B_{kl}^d \subseteq B_{ij}^d$. With the hypothesis : $\max[d(i, k), d(j, k), d(i, l), d(j, l)] \leq d(i, j) \leq d(k, l)$. Therefore $i, j \in B_{kl}^d$, and since B_{kl}^d is a weak cluster, $B_{ij}^d \subseteq B_{kl}^d$.

Given a dissimilarity d on I fulfilling the inclusion condition, Diatta (1998) characterises the quasi-ultrametric approximation derived from his algorithm, by a formula. Then, he establishes the optimality.

The formula is the following.

$$\forall i, j \in I, d_*(i, j) = \min \{d(u, v) \mid u, v \in I, i, j \in B_{uv}^d\} \quad (1)$$

Note that the condition $i, j \in B_{uv}^d$ is equivalent to $B_{ij}^d \subseteq B_{uv}^d$.

For fixed i, j in I , let $x, y \in I$ realise :

$d_*(i, j) = d(x, y)$, $B_{ij}^d \subseteq B_{xy}^d$. Then, $d(x, y) \leq d(i, j)$, so that by the lemma $B_{xy}^d = B_{ij}^d$.

Thus, an equivalent definition of d_* may be given. It will be the one used in the following.

$$\forall i, j \in I, d_*(i, j) = \min \{d(u, v) \mid u, v \in I, B_{uv}^d = B_{ij}^d\}. \quad (2)$$

Now, we have the ingredients to give a straightforward proof of optimality.

Proposition 1. *Let d be a dissimilarity on I fulfilling the inclusion condition, and d_* defined by (2). Then :*

- i) d_* is a lower maximal quasi-ultrametric less than d .
- ii) $\forall i, j \in I, B_{ij}^d = B_{ij}^{d_*}$.

Proof. Fix $i, j \in I$ and define x, y as above : $d_*(i, j) = d(x, y), B_{xy}^d = B_{ij}^d$. Note that $d_*(x, y) = d(x, y)$ and $d_*(i, j) \leq d(i, j)$. Thus d_* is less than d . For every k, l in B_{ij}^d , we prove the following equivalences (3)

$$\begin{aligned} B_{kl}^d = B_{ij}^d &\iff d(k, l) \geq d(x, y) \iff d_*(k, l) = d_*(i, j) \\ B_{kl}^d \subset B_{ij}^d &\iff d(k, l) < d(x, y) \iff d_*(k, l) < d_*(i, j). \end{aligned}$$

Indeed, by definition of $d(x, y), B_{kl}^d = B_{ij}^d$ implies $d(k, l) \geq d(x, y)$. The converse derives from the lemma. Thus, the first two equivalences of the two series are proved. Moreover, $B_{kl}^d = B_{ij}^d$ clearly implies $d_*(k, l) = d_*(i, j)$ and $d(k, l) < d(x, y)$ implies $d_*(k, l) \leq d(k, l) < d(x, y) = d_*(i, j)$. The proof of (3) is complete.

Now, we show that d_* preserves the 2-balls. By (3), for every k in B_{ij}^d , $d_*(i, k) \leq d_*(i, j)$ and $d_*(j, k) \leq d_*(i, j)$. Thus, $B_{ij}^d \subseteq B_{ij}^{d_*}$. Conversely, let $w \notin B_{ij}^d$, and without loss of generality, $d(i, w) \geq d(j, w)$ and $d(i, w) > d(i, j)$. Then, $j \in B_{iw}^d$, so that $B_{ij}^d \subseteq B_{iw}^d$, and even $B_{ij}^d \subset B_{iw}^d$ since $w \notin B_{ij}^d$. By (3), $d_*(i, j) < d_*(i, w)$ and $w \notin B_{ij}^{d_*}$. Point ii) of the proposition is proved.

As an immediate consequence, d_* obeys the inclusion condition.

Again by (3), d_* obeys the diameter condition. Thus d_* is quasi-ultrametric. Only optimality remains to be shown. We refer to Diatta (1998). Let δ be a quasi-ultrametric such that : $d_* \leq \delta \leq d$. In particular, $d_*(x, y) = \delta(x, y) = d(x, y)$.

For every k in I , we have :

$$\max [d_*(x, k), d_*(y, k)] \leq \max [\delta(x, k), \delta(y, k)] \leq \max [d(x, k), d(y, k)].$$

Thus, $k \in B_{xy}^d$ implies $k \in B_{xy}^\delta$, and $k \in B_{xy}^\delta$ implies $k \in B_{xy}^{d_*}$. Since $B_{xy}^d = B_{xy}^{d_*}$, we deduce : $B_{xy}^{d_*} = B_{xy}^\delta = B_{xy}^d$. Thus $i, j \in B_{xy}^\delta$. Since δ is quasi-ultrametric, it follows : $\delta(i, j) \leq \delta(x, y)$, whence : $d_*(i, j) \leq \delta(i, j) \leq \delta(x, y) = d(x, y) = d_*(x, y)$.

Thus $\delta = d_*$ and d_* is lower maximal.

The complexity of exhibiting all the B_{ij}^d 's is $O(n^3)$, and the one of pairwise comparisons is at most $O(n^5)$. That is the complexity to check whether the inclusion condition holds and, at all events, the one of every step of the descent algorithm. So, although Diatta's algorithm is polynomial, it is more time consuming than an algorithm based upon the characteristic formula. Recall that the complexity for quasi-ultrametricity is at most $O(n^4)$ by Bandelt's condition.

3 The optimal approximation via the one-to-one correspondence.

In introducing weak hierarchies, Bandelt and Dress(1989) show that the weak clusters associated with any dissimilarity d (in fact any similarity) form a closed weak hierarchy. Recall that a weak hierarchy is a collection \mathcal{H} of nonvoid subsets of I , satisfying :

$\forall H_1, H_2, H_3 \in \mathcal{H}, H_1 \cap H_2 \cap H_3 \in \{H_1 \cap H_2, H_2 \cap H_3, H_3 \cap H_1\}$. This axiom has been introduced by several authors, such as Batbedat (1989, 1990), Bandelt and Dress (1989). A closed weak hierarchy is a weak hierarchy closed under finite nonempty intersection. Adding an evenness condition, Diatta and Fichet (1994) prove that the weak clusters form a quasi-hierarchy and exhibit a level index based upon the diameter of the weak clusters. Recall that a quasi-hierarchy is a closed weak hierarchy containing the whole set I as a cluster, and such that the minimal clusters partition I . A level index f is a strictly monotone increasing mapping from \mathcal{H} to \mathbb{R}_+ , vanishing on the minimal elements.

Actually, many level indices may be proposed. For a given weak cluster H , let us consider the family x_H of real numbers formed by the quantities $d(i, j)$, $i, j \in I$ such that $H = B_{ij}^d$. When $H \subset H'$, H, H' weak clusters, the lemma shows that $a < a'$ for every a in x_H and every a' in $x_{H'}$. Moreover the evenness condition gives $a = o$ for every a in x_H , whenever H is minimal (H is some B_{ii}^d). Let φ be any numerical function defined on the set of finite families x of nonnegative numbers, such that $\min_{a \in x} \leq \varphi(x) \leq \max_{a \in x}$ for every x . Then clearly, f defined by $f(H) = \varphi(x_H)$ is a level index. The mappings φ defined by the mean, the median, the middle range, or more generally every parameter deriving from an L_p -transformation ($1 \leq p \leq \infty$) provide some simple examples. The minimum and the maximum also are convenient. Observe that the maximum yields the diameter as level index.

Now, suppose that d fulfils the inclusion condition and choose the index based upon the minimum. The weak clusters are the 2-balls and for every i, j in I , the smallest cluster H_{ij} containing i and j is B_{ij}^d . Thus $f(H_{ij})$ is $d_*(i, j)$ defined by (2). By the one-to-one correspondence, d_* is just the quasi-ultrametric associated with the indexed quasi-hierarchy. It is less than d , and the smallest cluster containing i, j is $B_{ij}^{d_*}$, so that $B_{ij}^{d_*} = B_{ij}^d$. The 2-balls are preserved. Then the optimality may be shown as in the proposition.

References

- BANDELT, H.J. (1992): Four point characterization of the dissimilarity functions obtained from indexed closed weak hierarchies, *Mathematisches Seminar*, Universität Hamburg, Germany.
- BANDELT, H.J. and DRESS A.W.M. (1989): Weak hierarchies associated with similarity measures: an additive clustering technique. *Bull. Math. Biol.*, 51, 113-166.

- BANDELT, H.J. and DRESS A.W.M. (1994): An order theoretic framework for overlapping clustering. *Discrete Math.*, 136, 21-37.
- BATBEDAT, A. (1989): Les dissimilarités Médas et Arbas, *Stat. Anal. Données*, 14, 1-18.
- BATBEDAT, A. (1990): Les Approches Pyramidales Dans La Classification Arborée. Masson, Paris.
- DIATTA, J. (1998): Approximating dissimilarities by quasi-ultrametrics. *Discrete Math.* 192, 81-86.
- DIATTA, J. and FICHET B. (1994): From Apresjan hierarchies and Bandelt-Dress weak hierarchies to quasi-hierarchies. In: E. Diday et al. (Eds.): *New approaches in Classification and Data Analysis*, Springer, Berlin, 111-118.
- DIATTA, J. and FICHET B. (1998): Quasi-ultrametrics and their 2-ball hypergraphs. *Discrete Math.* 192, 87-102.
- JARDINE, N. and SIBSON R. (1971): *Mathematical Taxonomy*, Wiley, London.

Estimating Missing Values in a Tree Distance

A. Guénoche, S. Grandcolas

IML, LIM,
163 Av. de Luminy, 13009 Marseille, France
(e-mail: guenoche@iml.univ-mrs.fr)
(e-mail: grandcolas@lim.univ-mrs.fr)

Abstract. In phylogeny, one tries to approximate a given dissimilarity by a tree distance. In some cases, especially when comparing biological sequences, some dissimilarity values cannot be evaluated and a partial dissimilarity with undefined values is only available. In that case one can develop a sequential method to reconstruct a weighted tree or to evaluate the missing values using a tree model. In this paper we study the latter approach and measure the quality of the estimated values using simulated noisy tree distances.

1 Introduction

In phylogeny we try to represent a distance D over a set of taxa X with an X -tree (X is the set of leaves, the edges are weighted by non-negative values and the length of the path between two leaves x and y approximate $D(x, y)$ (Barthélemy & Guénoche 1991). The leaves (or external vertices) correspond to the living species and the nodes (or internal vertices) to their common ancestors. It is well known that this representation is correct iff D is a tree distance, that is D fulfills the famous Four Point Condition (Buneman 1971): For any $x, y, z, t \in X$,

$$D(x, y) + D(z, t) \leq \text{Max}\{D(x, z) + D(y, t), D(x, t) + D(y, z)\}.$$

It is the same to say that for any quadruple of distinct elements $\{x, y, z, t\}$, among the three sums $D(x, y) + D(z, t)$, $D(x, z) + D(y, t)$, $D(x, t) + D(y, z)$, the two greatest are equal.

Generally, distance values evaluated comparing homologous sequences do not constitute a tree distance; they may not satisfy the triangle inequality and one can have a simple dissimilarity. Furthermore, sometimes we just compare fragments, instead of complete sequences, poorly overlapping each other. In this situation we get a partial distance, denoted in the following Δ , to reconstruct a complete X -tree.

In this real situation, we have two possibilities. The first one consists in developing a method adapted to partial distances. In the second one we first estimate the missing values and we reconstruct the tree, when the distance is no more partial, using a classical method. The first approach leads to sequential algorithms, for which one element is added to the growing tree at each

iteration, using only known values. We no longer recall here this approach and refer to Hein (1989), Leclerc & Makarenkov (1998) and Guénoche & Grandcolas (1999).

Here, we deal with the problem of estimating missing values, under the hypothesis that we are close to a tree distance. We measure the quality of these estimations, realizing simulations. For that, we select at random an X -tree T , hence a tree distance D , and we erase some values and add some noise to obtain a partial dissimilarity Δ . Then we reconstruct a valued X-tree Θ and we compare it to T using metric and non-metric (topological) criteria.

2 Evaluating missing values

2.1 Historically speaking

Following the first works of De Soete (1984), the estimation of missing values has been largely studied in phylogeny along several articles of Landry, Lapointe and Kirsch; we refer to the final one in 1996. The authors compare two types of evaluations; the first one corresponds to the ultrametric distance model; the second one suits to the additive tree distance model (our X-trees). Let $\Delta(u, v)$ be an unknown value and Δ_{uv} its estimation to determine.

According to the first model, triangle $\{x, u, v\}$ satisfies the ultrametric property, the two greatest values are equal: $\Delta(u, v) \leq \text{Max}\{\Delta(x, u), \Delta(x, v)\}$. Consequently, they adopt the formula:

$$\Delta_{uv} := \text{Min}_{x \in X} \text{Max}\{\Delta(x, u), \Delta(x, v)\}. \quad (1)$$

According to the second model, for any quadruple $\{x, y, u, v\}$, they apply the Four Point Condition to conclude:

$$\Delta_{uv} := \text{Min}_{x \neq y \in X} [\text{Max}\{\Delta(x, u) + \Delta(y, v), \Delta(x, v) + \Delta(y, u)\} - \Delta(x, y)]. \quad (2)$$

They underline that formula (2) can only be applied if there exists a quadruple in which $\Delta(u, v)$ is the single undefined value. After many simulations on selected data, they observe that the evaluation according to the X -tree model fits better than the first one. But if there is a great number of missing values, it is not always feasible to complete distances according to the additive tree model. Finally they recommend to use iteratively formula (2) as long as it computes new values and to pass to formula (1) when it is necessary.

2.2 A more accurate analysis

The above formulae can be improved. We come back to the case where $\Delta(u, v)$ is the single missing value over $\{x, y, u, v\}$ and let A and B be the two computable sums: $A := \Delta(x, u) + \Delta(y, v)$ and $B := \Delta(x, v) + \Delta(y, u)$. In accordance with the Four Point Condition, the third unavailable sum

$\Delta(x, y) + \Delta(u, v)$ is equal to $\text{Max}\{A, B\}$ only if A and B are different. In that case, we get an *evaluation by quadruple* of $\Delta(u, v)$ with the same formula:

$$\Delta uv := \text{Max}\{A, B\} - \Delta(x, y).$$

But if $A = B$, one can say nothing about $\Delta(u, v)$ and applying systematically formula (2) leads to overestimate $\Delta(u, v)$, especially when u and v are two adjacent leaves - u and v are siblings. In that case, $\Delta(u, v)$ never appears in one of the two greatest sums. No pair $\{x, y\}$ can make A and B different and $\Delta(u, v)$ cannot be evaluated by quadruple. For the same reason, the ultrametric model and formula (1) are not correct; if u and v are siblings, for any x , $\Delta(u, v) < \Delta(u, x) = \Delta(v, x)$. So we will realize an *evaluation by path*.

Metric bounds To control these estimations, we first calculate lower and upper bounds for each missing value. As we try to evaluate a distance function, these values must satisfy the triangle inequality. For triangle $\{x, u, v\}$ we must have $|\Delta(x, u) - \Delta(x, v)| \leq \Delta uv \leq \Delta(x, u) + \Delta(x, v)$. As it is true for any triangle containing $\{u, v\}$, we have:

$$\text{Max}_{x \neq u, x \neq v} |\Delta(x, u) - \Delta(x, v)| \leq \Delta uv \leq \text{Min}_{x \neq u, x \neq v} \Delta(x, u) + \Delta(x, v)$$

We test this condition and, if it is not satisfied, we force the estimated value to be in this interval. That is, if $\Delta uv < \text{Max}_{x \neq u, x \neq v} |\Delta(x, u) - \Delta(x, v)|$ then we pose $\Delta uv = \text{Max}_{x \neq u, x \neq v} |\Delta(x, u) - \Delta(x, v)|$, and it is the same for the upper bound. These bounds are not updated along iterations; they are established at the beginning with the only given values.

Evaluation by quadruples We consider missing values $\Delta(u, v)$ according to the number of pairs $\{x, y\}$ allowing to estimate them, that is the number of quadruples $\{x, y, u, v\}$ in which $\Delta(u, v)$ is the single missing value. For each quadruple, the question is to decide if A and B are close enough to be the two greatest sums; this will be denoted $A \simeq B$. If the answer is yes, we continue with the next quadruple; else we add quantities $\text{Max}\{A, B\} - \Delta(x, y)$ to compute, at the end, the average value. So

$$\Delta uv = \text{Average}_{\{A, B\} | A \neq B} \{\text{Max}\{A, B\} - \Delta(x, y)\}$$

Given a distance, more or less similar to a tree distance, it can be delicate to decide if the two computable sums are sufficiently close to be considered as the two greatest ones. We first compute, for any quadruple $\{x, y, z, t\}$ in which there is no missing value - a full quadruple - the three sums clearly denoted S_{\min} , S_{med} and S_{\max} . Then we calculate a factor f that would make equal S_{med} and S_{\max} , verifying $(1 + f)S_{\text{med}} = (1 - f)S_{\max}$. We get

$$f = \frac{S_{\max} - S_{\text{med}}}{S_{\max} + S_{\text{med}}}.$$

The final factor *Fact* is evaluated as the maximum value of f , over the set of full quadruples. It is as large as we are far from a tree distance. If there is no full quadruple to compute *Fact*, we directly apply the evaluation by path.

Now for every pair (u, v) , we decide that A and B are the two greatest sums if and only if $(1 + \text{Fact})\text{Min}\{A, B\} \geq (1 - \text{Fact})\text{Max}\{A, B\}$. At each iteration, we evaluate some missing values that can be used later, as long as their amount decreases. When no new estimation is made, we go to evaluation by path.

Evaluation by path It is well known that, in a tree distance, the length of the link between vertex u and a path $[x, y]$ is equal to $\frac{1}{2} [D(u, x) + D(u, y) - D(x, y)]$. In a tree reconstruction process, if we want to place a new taxa u , we select (x, y) in such a way that the quantity Lu is minimum:

$$Lu := \frac{1}{2} \text{Min}_{x \neq y \in X} [\Delta(u, x) + \Delta(u, y) - \Delta(x, y)]. \quad (3)$$

If no other element is connected to this link, Lu will be the length of the edge coming from u . If a vertex v is later connected to it, u and v will be siblings. In that case one can compute Lu and Lv with formula (3). But these lengths will be overestimated, since a shortest path for u comes from v and, reciprocally. So we must have $\Delta(u, v) \leq Lu + Lv$. In fact, if u and v are siblings, we have the topology of the left tree in Figure 1, but we can only estimate lengths as in the right tree. In the quantity $Lu + Lv$, the edge that separates paths $[u, v]$ and $[x, y]$ is counted twice.

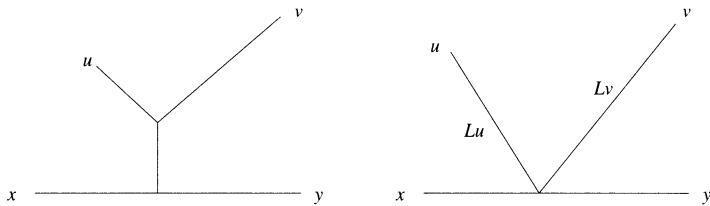


Fig. 1. Estimating the distance between u and v .

If we want to evaluate $Lu + Lv$ using the same path $[x, y]$, once again, there must be a single missing value over $\{x, y, u, v\}$. Consequently we compute Lu and Lv independently. So the estimation of Lu is always feasible if and only if there exist a full triangle $\{x, y, u\}$. To minimize the exceeding part of length between u and v , we apply formula

$$Duv := \frac{2}{3}(Lu + Lv).$$

in which $\frac{2}{3}$ is arbitrarily fixed, to fit a parameter used in the simulation process (in our random trees the ratio between internal and external edge lengths is equal to $\frac{1}{2}$).

3 Simulations

We begin calculating a dissimilarity close to a tree distance:

- We select at random an X -tree ($|X| = 20$) denoted T . We first define its topology, its edges list, that can be determined by a divisive process from class X to all the singletons. Then we give random weights to the edges with the following rule : the internal edges receive half the random value. This parameter permits to get reconstruction problems that are not too easy nor too difficult. After, we establish the tree distance D normalized to have an average value equal to 100.
- Then we add some noise to D according to a parameter τ . For each value $D(x, y)$, we select uniformly at random a value ε , such that $0 \leq \varepsilon \leq \tau$, and we apply formula $\Delta(x, y) := (1 \pm \varepsilon) * D(x, y)$, in which \pm is also the result of an equiprobable draw; half the values increase and half the values decrease with a percentage bounded by τ .
- Finally we consider a percentage of values ι as unknown. Pairs (u, v) corresponding to the missing values are also selected at random, keeping at least, for any element, 30% of its distance values.

From the partial distance or dissimilarity Δ we first estimate missing values by quadruple and by path. Then we apply the popular Neighbor Joining method (Saitou & Nei, 1987). We get an X -tree Θ and the corresponding D_θ distance that is compared to the initial tree distance D . For a long time, the metric point of view has been considered as essential; the best tree was the one minimizing the quadratic difference. Nowadays, considering trees as a representation of an evolutionary process, topological criteria become more important. They can be quantified comparing the shape of induced sub-trees on four taxa. In the following, we measure three criteria and give mean values corresponding to 100 trials.

- the quadratic difference Eq between D and D_θ .
- the percentage of quadruples well designed, denoted *quad*. Let $xy|zt$ be the topology of quadruple $\{x, y, z, t\}$ in T . If we get the same topology in Θ , we count one point; if not, but one topology is without internal edge, we count half a point.
- the percentage of bipartitions in T that are recovered in Θ denoted *RF*; we do not consider trivial splits, that are bipartition with a singleton. These non-trivial bipartitions correspond to internal common edges in both trees. As the random trees are binary, both have $n-3$ bipartitions of this type. This criterium is equivalent to the Robinson-Foulds distance (1981) between X -trees.

Results, given in the Table 1, prove that these estimations are efficient; 30% of unknown values permit to recover a satisfactory tree if the known values have distortion not greater than 15% of the initial values. If there are only 10% of missing values, then a noise corresponding to $\tau = .25$ can be accepted. To end we indicate that this strategy is more efficient than the

reconstruction of a tree with a sequential algorithm, using only given values, as illustrated in Guénoche and Grandcolas (1999).

	ι	0%	10%	20%	30%	40%	50%		ι	0%	10%	20%	30%	40%	50%
τ	Eq	0	0	2	4	11	19	τ	Eq	13	23	43	76	162	345
0%	$quad$	1	1	1	1	.99	.99	15%	$Quad$	1	.99	.98	.95	.89	.78
	RF	1	1	.99	.98	.95	.91		RF	.99	.95	.92	.81	.65	.47
τ	Eq	2	3	10	17	44	116	τ	Eq	25	38	71	140	270	478
5%	$quad$	1	1	.99	.99	.97	.91	20%	$quad$.99	.98	.96	.91	.84	.75
	RF	1	.99	.97	.94	.86	.70		Rf	.98	.92	.84	.71	.55	.40
τ	Eq	6	9	20	35	94	204	τ	Eq	40	65	111	202	372	629
10%	$Quad$	1	1	.99	.98	.94	.86	25%	$quad$.99	.97	.93	.88	.79	.69
	RF	1	.98	.95	.90	.77	.59		RF	.94	.88	.77	.65	.47	.34

Table 1 : Criterion values according to noise and uncertainty parametres

References

- BARTHÉLEMY J.P., GUÉNOCHE A., (1991): *Trees and Proximity Representations*, J. Wiley.
- BUNEMAN P., (1971): The recovery of trees from measures of dissimilarity: 387-395, in *Mathematics in Archaeological and Historical Sciences*, F.H. Hodson, D.G. Kendall, P. Tautu (Eds.), Edinburg University Press.
- DE SOETE G., (1984): Ultrametric tree representations of incomplete dissimilarity data, *J. of Classification*, 1: 235-242.
- GUÉNOCHE A., GRANDCOLAS S., (1999): Approximation par arbre d'une distance partielle, *Mathématiques, Informatique et Sciences humaines*, 146: 51-64.
- HEIN J.J., (1989): An optimal algorithm to reconstruct trees from additive distance data, *Bulletin of Mathematical Biology* 51 (5): 597-603.
- LAPOINTE F.J., KIRSCH J.A.W., (1996): Estimating phylogenies from lacunose distances matrices: Additive is superior to Ultrametric estimation, *Mol. Biol. Evol.*, 13(6): 266-284.
- LECLERC B., MAKARENKO V., (1998): On some relations between 2-trees and tree metrics, *Discrete Math.*, 192.
- ROBINSON D.R., FOULDS L.R., (1981): Comparision of phylogenetic trees, *Mathematical Biosciences*, 53: 131-147.
- SAITOU N., NEI M., (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, 4: 406-425.

Estimating Trees From Incomplete Distance Matrices: A Comparison of Two Methods

Claudine Levasseur, Pierre-Alexandre Landry
and François-Joseph Lapointe

Département de sciences biologiques, Université de Montréal,
C.P. 6128, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3J7
(e-mail: levassec@magellan.umontreal.ca)

Abstract. In the present paper, we compare two methods (TRIANGLE and MW) for estimating trees from incomplete distances matrices through simulations. Our results illustrate that MW performs better for recovering path-length distances whereas TRIANGLE is superior in terms of topological recovery. Recommendations are provided as to which method should be used with real experimental data.

1 Introduction

Numerous comparison methods used in experimental biology can lead to incomplete data sets that are not directly amenable to statistical analyses. Indeed, the vast majority of tree-building algorithms for distance data cannot handle missing values. As a solution to this problem, it has been proposed to estimate the missing cells in the lacunose (i.e. incomplete) distance matrices prior to tree reconstruction (De Soete, 1984; Landry and Lapointe, 1997). Such methods rely on basic tree-distance properties to fill incomplete matrices, using the available distances to estimate the missing ones. Recently, two tree-building algorithms allowing for missing cells in distance matrices were proposed by different authors, however; the TRIANGLE method of Guénoche and Grandcolas (1999) relies on a constructive approach whereas the MW (i.e. Method with Weights) procedure of Makarenkov and Leclerc (1999) is based on an optimization approach. This paper aims at comparing these two techniques through simulations, in order to determine their relative performance in terms of distance as well as topological recovery. We will show that MW generally provides better estimates of the path-length distances while TRIANGLE is better at recovering the correct topology of the tree.

2 Material and methods

2.1 Tree-building algorithms

The TRIANGLE method originally designed for complete distance matrices (Leclerc and Makarenkov 1998) can also be applied to reconstruct a tree from an incomplete set of distances (Guénoche and Leclerc, 1998). The algorithm

is based on a sequential procedure in which a support twotree (i.e., a bidimensional generalization of a tree; Harary and Palmer, 1968) is used to build an additive tree, adding a new vertex at each step with respect to those already placed on the tree (Guénoche et Grandcolas, 1999). The MW procedure, on the other hand, represents a weighted least-squares method (Makarenkov and Leclerc, 1999). In the case of complete distance matrices, all weights are identical and are set to one; in the case of incomplete matrices, null weights are assigned to missing distances while known distances are set to a unit weight. Consequently, the missing values do not contribute to the sum-of-squares criterion when lacunose matrices are used in the computations.

2.2 Distance matrices

In the present study, we used five complete DNA-hybridization matrices of various sizes ($n = 7$ to 15) for consistency with our previous work (Landry and Lapointe, 1997). Such matrices are of interest because they do not always satisfy all distance properties (DNA-hybridization matrices are never symmetrical and seldom metric). In addition to these experimental data, we also used tree-distance matrices satisfying the ultrametric property, as well as additive matrices meeting the four-point condition. The tree-distance matrices were computed from the corresponding trees obtained by applying ultrametric and additive least-squares algorithms to the original DNA-hybridization matrices (see Landry and Lapointe 1997). Thus, a total of fifteen matrices were considered for the simulations.

2.3 Simulation design

Each of the fifteen matrices was submitted to the same treatments: (1) we first generated lacunose matrices by removing distances at random from the complete matrices; (2) a tree was then obtained for each lacunose matrix with either TRIANGLE or MW; (3) the trees were finally compared to those derived from the original matrices with the corresponding algorithm (in the case of ultrametric and additive matrices, that step was not required). Increasing percentages P of missing distances (from $P = 10\%$ to 60%) were considered by deleting pairs of corresponding distances $d(i,j)$ and $d(j,i)$ from the complete matrices. In each case, 100 replicate matrices were generated. To assess whether the competing methods could recover the actual path-lengths of the trees derived from the complete data, we computed the correlations (r) and sum-of-squared differences (SoS) between the tree-distance matrices obtained from each of the 100 replicates and the original tree-distances. The average correlation was then calculated for every P and each type of matrices. Since it was shown that there is no relationship between the size of a matrix and the recovery values for a given P (see Landry and Lapointe, 1997), all correlations corresponding to the different matrices of the same type were pooled together for a global comparison of the methods. Moreover,

the SoS values were compared in a pairwise fashion to assess which of MW or TRIANGLE was more accurate in each simulation, and determine how many times the different methods produced identical results.

For topological comparisons, the Robinson and Foulds (1981) metric (RF) was computed to assess whether the original trees could be recovered from lacunose matrices. In order to combine the topological recovery for matrices of different sizes, each RF value was standardized by its maximal possible value ($\max = 2n-6$) for a matrix of size n . The complement of this standardized statistic ($1 - \text{RF}/\max$) was then taken as an index of topological similarity, with a maximal value of one for topologically identical trees, and a value of zero for the most different possible pairs of trees. As for metric recovery, we counted the number of times that a given method was superior or equal to the other in terms of RF, and reported those frequencies.

3 Results

The results of the path-length correlations are presented in Figure 1A. These simulations show that MW is usually superior to (or as good as) TRIANGLE in terms of average recovery values, except when DNA-hybridization matrices are considered and P is larger than 40%. This can be attributed to the fact that TRIANGLE is not always able to build a tree when too many distances are missing. Thus, when it succeeds, the results are often better than those obtained for MW in the same cases (this discrepancy will then be reflected in the reduced average correlations for MW). Interestingly, the results of topological comparisons (Figure 1B) are opposite to those obtained for metric recovery. In this case, there is no doubt about the superiority of TRIANGLE for additive and ultrametric matrices. On the other hand, MW does perform better for DNA-hybridization matrices, except when P is larger than 40%. Again, this result can be attributed to the poor performance of TRIANGLE when many distances are missing (and the data are not metric).

In Figure 2A, we present the results of the pairwise SoS comparisons. They show that both methods will recover the same path-length distances for small values of P . For larger P , TRIANGLE gets better and better; but at the same time, it becomes less likely to return solution (for $P = 60\%$, TRIANGLE was unable to compute a tree in half of the cases). Once again, these results are different when DNA-hybridization matrices are considered. With such experimental data, MW is by far superior to TRIANGLE (about 60% of the times) and the competing methods seldom return the same solution. The results of the topological comparisons (Figure 2B) are compatible with the metric recovery values. In the majority of cases, TRIANGLE makes less topological errors when building a tree from incomplete matrices. For DNA-hybridization matrices, MW performs better, however.

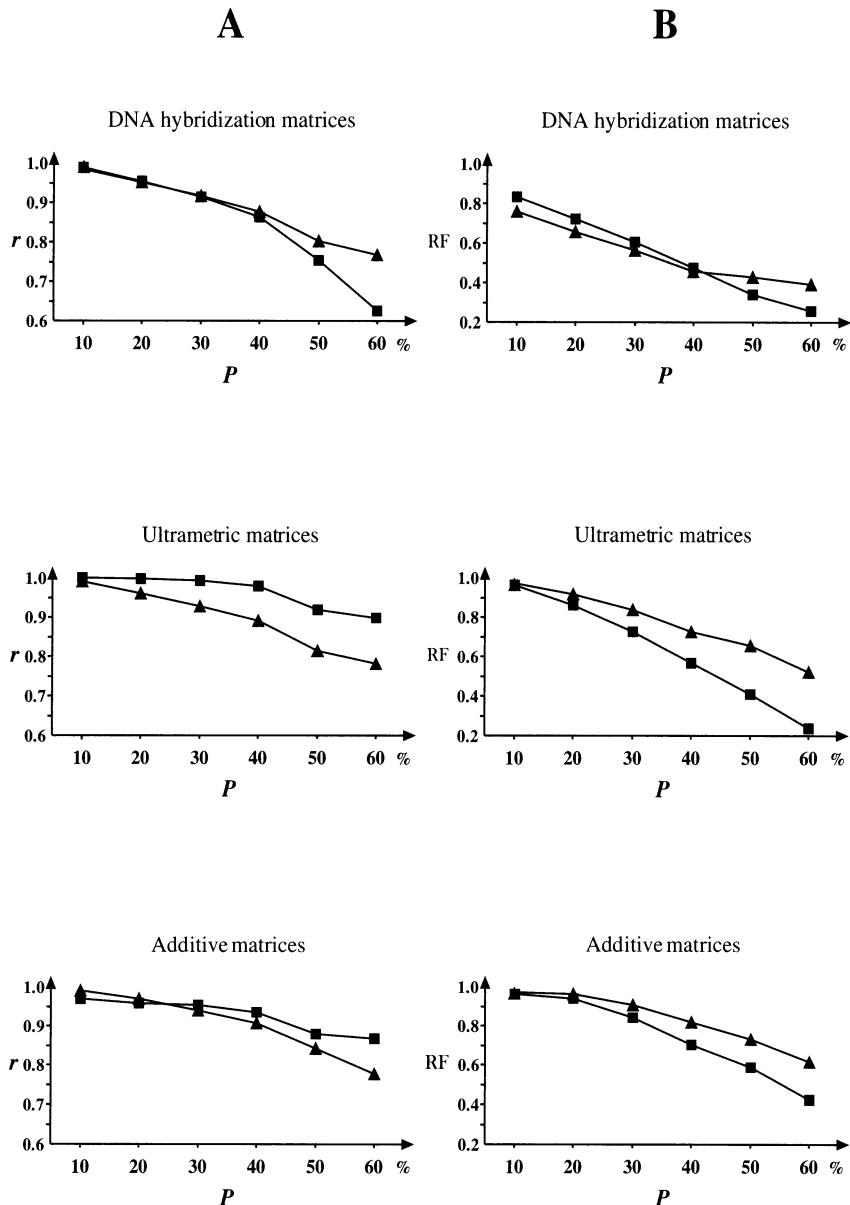


Figure 1: Metric and topological recovery values obtained with increasing percentages of missing cells P , for the two competing methods [MW (■), TRIANGLE (▲)]. (A) Path-length correlations (r). (B) Standardized Robinson and Foulds metric values (RF).

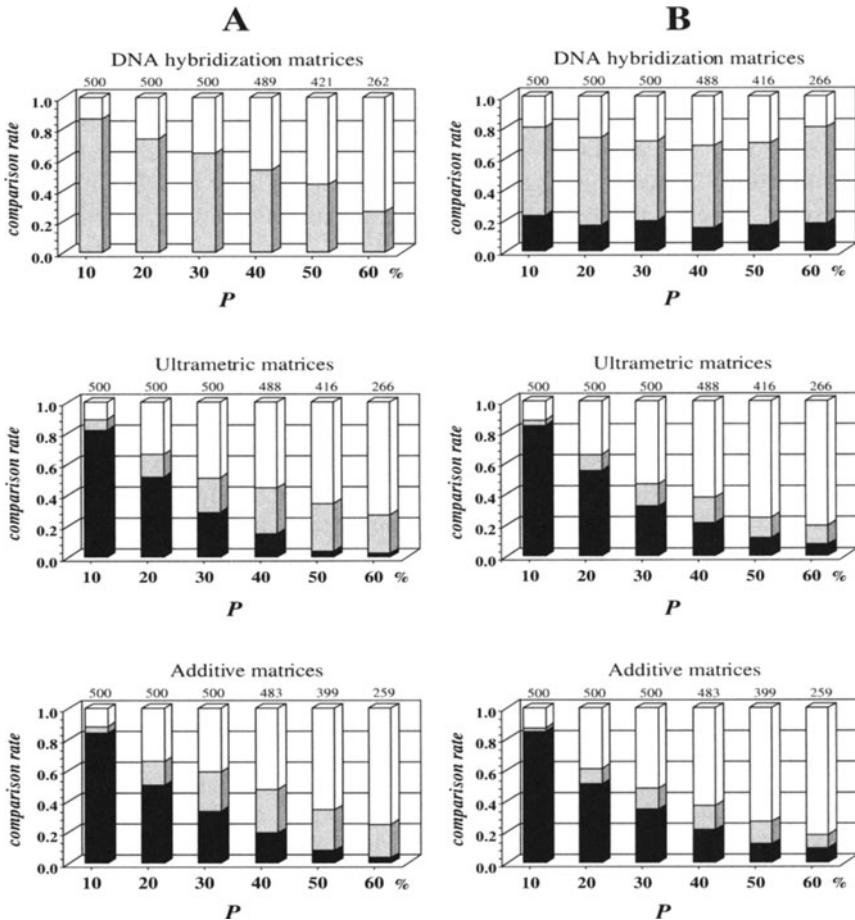


Figure 2: Relative comparison rates of the two competing methods for three types of distance matrices and increasing percentages of missing cells P . (A) Metric comparisons (SoS). (B) Topological comparisons (RF). The sample size on top of each bar represents the number of times for which TRIANGLE was able to return a solution. TRIANGLE better than MW (□); MW better than TRIANGLE (■); MW equal to TRIANGLE (■■).

4 Conclusion

In this paper, we have compared the performance of two different tree-building methods for incomplete distance matrices. We have shown that when TRIANGLE is able to return a solution, it usually provides better topological recovery than MW. On the other hand, MW should be preferred in terms of average metric recovery. Still, one has to realize that our results based

on tree-distance matrices are far from those obtained with "noisy" DNA-hybridization matrices. Experimental data are rarely ultrametric or additive, let alone metric. In such situations, MW must be used as it clearly outperforms TRIANGLE in terms of topological as well as metric recovery. In the general case, however, the choice of a given method seems to rely on whether one desires to recover the topology or the path-lengths of a tree. Ultimately, we would recommend to use both techniques in combination, increasing in this way the success of the reconstruction. We postulate that the recovery rate will be increased when MW and TRIANGLE return the same tree, as it is often the case for small P values.

5 Acknowledgments

The authors are grateful to A. Guénoche and V. Makarenkov for providing the TRIANGLE and MW programs. J.A.W. Kirsch kindly provided the DNA-hybridization matrices. This work was supported by a NSERC grant to FJL.

References

- DE SOETE, G. (1984): Additive-Tree Representations of Incomplete Dissimilarity Data. *Quality and Quantity*, 18, 387–393.
- GUÉNOCHE, A. and LECLERC, B. (1998): La méthode des Triangles pour Reconstruire un Arbre à Partir de Distances Incomplètes. In: *Actes des Journées de la Société Francophone de Classification*. Agro-Montpellier, 117–120.
- GUÉNOCHE, A. and GRANDCOLAS, S. (1999): Approximations par Arbre d'une Distance Partielle. *Mathématiques, Informatique et Sciences Humaines*, 146, 51–64.
- HARARY, F. and PALMER, E. M. (1968): On Acyclical Simplicial Complexes. *Matematika*, 15, 115–122.
- LANDRY, P.-A. and LAPONTE, F.-J. (1997): Estimation of Missing Distances in Path-Length Matrices: Problems and Solutions. In: B. Mirkin, F.R. McMorris, F.S. Roberts, and A. Rzhetsky (Eds.): *Mathematical hierarchies and biology*. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Providence, 209–224.
- LECLERC, B. and MAKARENKO, V. (1998): On Some Relations Between 2-trees and Tree Metrics. *Discrete Mathematics*, 192, 223–249.
- MAKARENKO, V. and LECLERC, B. (1999): The Fitting of a Tree Metric According to a Weighted Least-Squares Criterion. *Journal of Classification*, 16, 3–26.
- ROBINSON, D.F. and FOULDS, L.R. (1981): Comparison of Phylogenetic Trees. *Mathematical Biosciences*, 53, 131–147.

Zero Replacement in Compositional Data Sets

J. A. Martín-Fernández¹, C. Barceló-Vidal¹, and V. Pawlowsky-Glahn²

¹ Dept. Informàtica i Matemàtica Aplicada, Universitat de Girona, Campus Montilivi – Edifici PI, E-17071 Girona, Spain. (e-mail: jamf@ima.udg.es)

² Dept. Matemàtica Aplicada III, ETSECCPB, Universitat Politècnica de Catalunya, E-08034 Barcelona, Spain. (e-mail: vera.pawlowsky@upc.es)

Abstract. The sample space of compositional data is the open simplex. Therefore, zeros in a compositional data set are identified either with below detection limit values, or lead to a division of the data set into different subpopulations with the corresponding lower dimensional sample space. Most multivariate data analysis techniques require complete data matrices, thus calling for a strategy of imputation of zeros in the first case. Existing replacement methods of rounded zeros are reviewed, and a new method is proposed, who's properties are analyzed and illustrated. The method is applied in a hierarchical cluster analysis of compositional data.

1 Introduction

Compositional data are by definition proportions of some whole. Thus, their natural sample space is the open simplex and interest lies in the relative behavior of the components. The open simplex is defined as (Aitchison, 1986)

$$\mathcal{S}^D = \{(x_1, x_2, \dots, x_D)' : x_j > 0; j = 1, 2, \dots, D; x_1 + x_2 + \dots + x_D = 1\}.$$

Any vector of positive components, $\mathbf{y} \in \mathbb{R}_+^D$, can be projected into the simplex by the *closure operation* $\mathcal{C}(\mathbf{y}) = (y_1/\sum y_j, y_2/\sum y_j, \dots, y_D/\sum y_j)'$. The only operations known to induce a vector space structure on the simplex are the *perturbation operation*, $\mathbf{p} \circ \mathbf{x} = \mathcal{C}(p_1 x_1, p_2 x_2, \dots, p_D x_D)'$, defined on $\mathcal{S}^D \times \mathcal{S}^D$, and the *power transformation*, $\alpha \diamond \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha)'$, defined on $\mathbb{R} \times \mathcal{S}^D$. Perturbation can be proven to be equivalent to translation in \mathbb{R}^D and power transformation to the scalar product using the centered-logratio transformation (clr). This transformation has been defined by Aitchison (1986) as $\text{clr}(\mathbf{x}) = (\ln(x_1/g), \ln(x_2/g), \dots, \ln(x_D/g))'$, with $g = (\prod_{i=1}^D x_i)^{1/D}$. Another operation on the simplex, analogous to projection onto a smaller dimensional space in \mathbb{R}^D , is obtained through the concept of *subcomposition*. It is defined as $\mathbf{x}_S = \mathcal{C}(\mathbf{S}\mathbf{x})$, $\mathbf{x} \in \mathcal{S}^D$, where \mathbf{S} is a $(S \times D)$ selecting matrix with all elements zero except one in each row and at the most one in each column. The subcomposition \mathbf{x}_S belongs to the simplex \mathcal{S}^S . A distance, compatible with the previously defined operations, is Aitchison's distance $d_a(\mathbf{x}, \mathbf{x}^*) = d_e(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{x}^*))$, where $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$ and d_e is the Euclidean distance. Its properties have been discussed in Martín-Fernández et al. (1998) and in Aitchison et al. (1999).

The “Achilles heel” of $d_a(\mathbf{x}, \mathbf{x}^*)$ is the presence of zero values in the data, as it is not possible to take the logarithm of zero. Zero values are present in many applications as, for example, in a household budget spending nothing on the commodity group “tobacco and alcohol”, or in a rock specimen containing “no trace” of a particular mineral. In compositional data we distinguish two kinds of zeros: *essential zeros* and *rounded zeros*. The zero in the household spending pattern is essential. The zero in a particular mineral is usually a rounded zero, *i. e.* it indicates that no quantifiable proportion of the mineral can be recorded according to the accuracy of the measurement process. In hierarchical cluster analysis the presence of an essential zero in a component is an indication that the observation belongs to a different group and straightforward division of the sample is advisable. The principal problem refers to rounded zeros.

The purpose of this paper is to revise, from a theoretical point of view, the additive method of replacement suggested by Aitchison (1986), whose drawbacks have been described from an empirical point of view in Tauber (1999), and to propose a new method of replacement of zeros in compositional data. First, we present the solution proposed by Aitchison (1986). Next, we summarize the most usual non-parametric approaches for missing values with non-compositional data and then we propose a new method of zero replacement. Finally, we present an example where the proposed method is applied to a compositional data set.

2 Additive replacement strategy

Aitchison (1986) suggests that an observation $\mathbf{x} \in \mathcal{S}^D$ containing C rounded zeros can be replaced by a new observation $\mathbf{r} \in \mathcal{S}^D$ without zeros according to the following replacement rule:

$$r_j = \begin{cases} \frac{\delta(C+1)(D-C)}{D^2}, & \text{if } x_j = 0, \\ x_j - \frac{\delta(C+1)C}{D^2}, & \text{if } x_j > 0, \end{cases} \quad (1)$$

where δ is smaller than a given threshold derived from the measurement process.

Note that the constant-sum-constraint of compositional data forces to modify both the zero and the non-zero values. Moreover, the imputed value r_j depends not only on δ but also on the dimension D and the number C of zeros. Note also that a different δ_j could be considered for every component x_j leading to a slightly more complicated expression.

Due to the fact that the transformation (1) of non-zero values is additive, it holds that $r_k/r_l \neq x_k/x_l$, for x_k, x_l non-zero values, and the value of the new ratios r_k/r_l depends on δ . Therefore, Aitchison’s distance between two replaced observations is extremely sensitive to changes in δ as illustrated empirically by Tauber (1999).

3 Replacement strategies for non-compositional data

Let \mathbf{Y} be a data set with missing values in real space \mathbb{R}^D . If the goal is to perform a cluster analysis based on a hierarchical clustering method using the Euclidean distance, it is necessary to complete first the matrix of distances between observations. Several strategies have been suggested in the literature for that purpose. The one by Krzanowski (1988) can be synthesized as follows: (i) omit any variable that has a missing value when computing the distance between two observations and work only with those variables that have all values present for both the observations concerned; (ii) if the previous step means working with S variables instead of D , inflate the resulting distance by a factor D/S . To see that this strategy is not suitable for compositional data, consider the following example: given three compositional observations $\mathbf{x} = (0, 0.8, 0.2)$, $\mathbf{x}^* = (0.95, 0.04, 0.01)$, and $\mathbf{x}' = (0.06, 0.76, 0.18)$, the strategy of Krzanowski implies to consider the subcompositions formed by the second and third variables: $\mathbf{x}_S = (0.8, 0.2)$, $\mathbf{x}_S^* = (0.8, 0.2)$, and $\mathbf{x}_S' = (0.81, 0.19)$. Assuming that the zero in sample \mathbf{x} is actually a very small positive value, we expect \mathbf{x} and \mathbf{x}' to be more similar than \mathbf{x} and \mathbf{x}^* . Nevertheless, we obtain that $d_a(\mathbf{x}_S, \mathbf{x}_S^*) = 0$ and $d_a(\mathbf{x}_S, \mathbf{x}_S') = 0.07$.

The most common strategy to complete the matrix of distances is to employ “imputation”, *i. e.* the insertion of an estimate for each missing value, thereby completing the data set, and then calculate the matrix of distances. When the missing values are actually censored data, that is, when the values for some variables are reported as “less than” a given threshold value, a simple imputation can be considered. For a “small” proportion of “less than” values (not more than 10%) a simple-substitution method using 0.55 of the threshold value is suggested in Sandford et al. (1993). More general imputation methods are exposed in Little and Rubin (1987).

All these imputation methods have one thing in common: the canonical projection $\Pi(\mathbf{y})$ on the non-missing variables of observation \mathbf{y} is identical to the same projection $\Pi(\mathbf{z})$ of the replaced observation \mathbf{z} . Also, if \mathbf{y} and \mathbf{y}^* have “common” missing values, *i. e.* missing values on the same variables, it holds that $y_j - y_j^* = z_j - z_j^*$ for y_j, y_j^* non-missing values and z_j, z_j^* the corresponding replacement. Furthermore, if the imputation method assigns the same replacement value to every missing component y_j of the two observations, then $d_e(\mathbf{z}, \mathbf{z}^*)$ does not depend on the imputed values and it is identical to the Euclidean distance between the projections $d_e(\Pi(\mathbf{y}), \Pi(\mathbf{y}^*))$. With the above features in mind, let us proceed to define a suitable replacement method for zeros in compositional data.

4 Multiplicative replacement strategy

Let be $\mathbf{x} \in \mathcal{S}^D$ and assume it has C zeros. We propose to replace \mathbf{x} with an observation $\mathbf{r} \in \mathcal{S}^D$ without zeros using the expression

$$r_j = \begin{cases} \delta_j, & \text{if } x_j = 0, \\ x_j(1 - \sum_{k|x_k=0} \delta_k), & \text{if } x_j > 0, \end{cases} \quad (2)$$

where δ_j is the imputed value on the component x_j . Following Sandford et al. (1993), whenever δ_j is equal to 0.55 of the threshold determined from the measurement process corresponding to component x_j , a simple-substitution in the simplex is obtained.

The multiplicative modification of non-zero values in (2) has the following desirable properties not satisfied by (1):

1. It is “natural” in the sense that, if the imputed values δ_j in an observation \mathbf{x} are equal to the “true” censored values, then \mathbf{r} recovers the “true” observation.
2. It is coherent with the basic operations in the simplex, *i. e.* if a selecting matrix \mathbf{S} of non-zero components of observation \mathbf{x} is considered, and $\mathbf{x}_S = \mathcal{C}(\mathbf{S}\mathbf{x})$ is the subcomposition obtained, denoting by $\mathbf{r}_S = \mathcal{C}(\mathbf{S}\mathbf{r})$ the subcomposition derived from the replacement vector, the following properties hold:
 - (a) *perturbation invariance* — for all $\mathbf{p} \in \mathcal{S}^D$, $(\mathbf{p} \circ \mathbf{r})_S = (\mathbf{p} \circ \mathbf{x})_S$;
 - (b) *power transformation invariance* — for all $\alpha \in \mathbb{R}$, $(\alpha \diamond \mathbf{r})_S = (\alpha \diamond \mathbf{x})_S$;
 - (c) *subcomposition invariance* — $\mathbf{x}_S = \mathbf{r}_S$.
3. When \mathbf{x} and \mathbf{x}^* have “common” zero values, and the replaced observations \mathbf{r} and \mathbf{r}^* are obtained using identical imputation values $\delta_j = \delta_j^*$, then
 - (a) $r_j/r_j^* = x_j/x_j^*$ for all non-zero values x_j , x_j^* , and $d_a(\mathbf{r}, \mathbf{r}^*)$ does not depend on the imputed values;
 - (b) $d_a(\mathbf{r}, \mathbf{r}^*)$ is not equal to $d_a(\mathbf{x}_S, \mathbf{x}_S^*)$, but the following equality holds:

$$d_a^2(\mathbf{r}, \mathbf{r}^*) = d_a^2(\mathbf{x}_S, \mathbf{x}_S^*) + \frac{C}{D(D-C)} \left[\sum_{x_j>0} \log \left(\frac{x_j}{x_j^*} \right) \right]^2,$$

where C is the number of common zeros in \mathbf{x} and \mathbf{x}^* .

5 Example

Consider the Glacial data set included in Aitchison (1986). It has 92 samples of pebbles of glacial tills sorted into four categories: red sandstone, gray sandstone, crystalline, and miscellaneous. The components x_1 , x_2 , x_3 , and x_4 represent the corresponding percentages by weight of these four categories. Zeros appear in 41 out of the 92 observations either in component x_3 or in x_4 . We assume the zeros to be non-essential zeros, *i. e.* rounded zeros. Before applying a hierarchical clustering algorithm, the zeros have to be replaced. For comparison purposes, we consider the additive replacement approach proposed by Aitchison (1) and the multiplicative replacement (2)

proposed in this paper, combined with two different δ values $\delta_1 = 0.001$ and $\delta_2 = 0.0005$. As a consequence, four data sets without zeros are obtained: $\mathbf{R}_{1,1}$ using method (1) and δ_1 ; $\mathbf{R}_{1,2}$ using the same method but δ_2 ; $\mathbf{R}_{2,1}$ using method (2) and δ_1 ; and $\mathbf{R}_{2,2}$ using method (2) and δ_2 . The clustering algorithm used has been Ward's method adapted to compositional data (Martín-Fernández et al., 1998), resulting in two distinct groups in all four cases. Comparing the two groups obtained in each case, the following facts can be observed: classifications of $\mathbf{R}_{1,1}$ and $\mathbf{R}_{2,1}$ are extremely coincident, as only one observation is assigned to a different group; classifications of $\mathbf{R}_{2,1}$ and $\mathbf{R}_{2,2}$ are identical; and classification of $\mathbf{R}_{1,2}$ is appreciably different of the rest, as 17 observations are assigned to a different group when compared to $\mathbf{R}_{1,1}$. This indicates that with the multiplicative replacement (2) the matrix of distances is more stable with respect to changes of the imputed values δ_j . But, when the imputed values tend to zero the two replacement sets tend to give us the same results. If we take $\delta = 10^{-8}$ in the two cases and we apply Ward's method, we obtain 4 distinct groups (see Figure 1).

Group G1 corresponds to the observations without zeros, group G2 corresponds to observations with zero only in component x_3 , group G3 corresponds to observations with zero only in component x_4 , and group G4 corresponds to observations with zeros in both components. These groups could be obtained if initially we assume the zeros as essential zeros rather than rounded zeros.

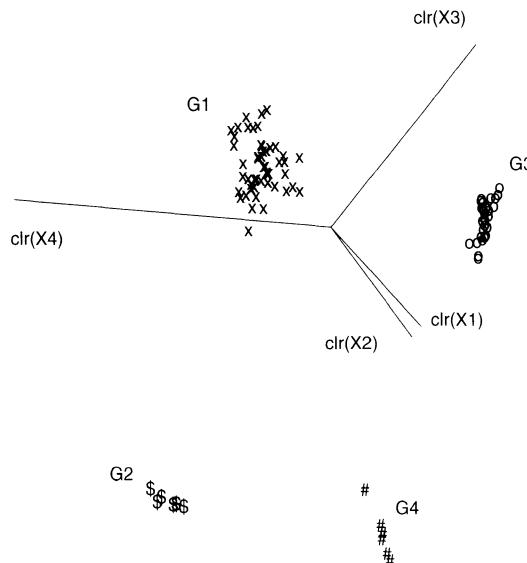


Fig. 1. Biplot of the compositional data set obtained by replacement (2) when $\delta = 10^{-8}$. Labels G1,G2,G3, and G4 represents the four groups.

6 Conclusions

In this paper, a multiplicative zero replacement method for compositional data is defined. This replacement is coherent with the basic operations which provide the simplex with a vector space structure. In particular, the multiplicative approach is “natural” in the sense that it recovers the “true” observation if replacement values are identical to the missing values.

Acknowledgments

This research has been partially supported by the University of Girona through the project UdG98/32, and by the Dirección General de Enseñanza Superior (DGES) of the Spanish Ministry for Education and Culture through the project PB96-0501-C02-01.

References

- AITCHISON, J. (1986): *The Statistical Analysis of Compositional Data*. Chapman and Hall, New York (USA), 416 p.
- AITCHISON, J., BARCELÓ-VIDAL, C., MARTÍN-FERNÁNDEZ, J.A., and PAWLOWSKY-GLAHN, V. (2000): Logratio analysis and compositional distance. *Mathematical Geology*, (in press).
- KRZANOWSKI, W.J. (1988): *Principles of Multivariate Analysis. A User's Perspective*, Clarendon Press, Oxford (GB), 563 p. (reprinted 1996).
- LITTLE, R.J.A and RUBIN, D.B. (1987): *Statistical Analysis with Missing Data*. John Wiley & Sons, New York (USA), 278 p.
- MARTÍN-FERNÁNDEZ, J.A., BARCELÓ-VIDAL, C., and PAWLOWSKY-GLAHN, V. (1998): A Critical Approach to Non-parametric Classification of Compositional Data. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer, Heidelberg, pp. 49-56.
- SANDFORD, R.F, PIERSON, C.T., and CROVELLI, R.A. (1993): An Objective Replacement Method for Censored Geochemical Data. *Mathematical Geology*, Vol. 25:1, pp 59-80.
- TAUBER, F. (1999): Spurious clusters in Granulometric Data Caused by Logratio Transformation. *Mathematical Geology*, Vol. 31:5, pp. 491-504.

EM Algorithm for Partially Known Labels

C. Ambroise and G. Govaert

UMR CNRS 6599, Centre de recherches de Royallieu
BP 20259 60205 Compiègne cedex France
(e-mails: ambroise@utc.fr and govaert@utc.fr)

Abstract. Mixture models are widely used for clustering or discrimination problems. Estimating the parameters of such models can be viewed as an incomplete data problem and has thus often been handled by the Expectation-Maximization (EM) algorithm. It has been shown that this method can integrate additional information such as the label of some observations. In this paper we propose a generalization of this approach which can take into account partial information about the observation labels. An example illustrates the relevance of the proposed method for mixture density estimation.

1 Introduction

Mixture models are a popular tool for density estimation; they are used in discrimination for modeling conditional class densities and in clustering for representing a mixture of different subpopulations with parametric cluster densities.

The EM algorithm (Dempster et al. (1977), McLachlan and Krishnan (1997)), which is usually used for the estimation of the mixture model parameters, leads to a very simple and efficient iterative estimation for such models.

In this paper, we focus on the estimation of mixture model from partially labeled observations. This problem has already been addressed by some authors (O'Neil (1978), Ganesalingam and McLachlan (1978)) when estimating on the basis of sample containing both labeled and unlabeled objects. They have shown that considering mixed samples for parameter estimation improves the quality of the estimators and that it is particularly suited for discrimination problems. Here, we propose an innovative application of the EM algorithm which generalizes this approach to the estimation from partially known labels. It allows to take into account the knowledge that some objects do belong to subsets of possible components of the mixture or conversely cannot originate from some components.

Considering partially known labels allows to incorporate additional knowledge in clustering or discrimination problems. In the case of discrimination, the parameters of the classifier are estimated using a learning set. This requires that an expert has labeled all the objects of a given set. But it often happens that the expert easily states that an object does not belong to some classes without knowing exactly what is the label of the considered object.

Usually this kind of knowledge is not taken into account in the framework of pattern recognition. For example, in the medical domain a doctor is able to determine that his patient has not these kind of diseases without being able to identify the exact problem of the patient.

Cluster analysis aims at partitioning a set of objects into subgroups where the objects inside a cluster show some degree of “similarity”. It can be very helpful for identifying the structure of the subgroups to know that two distinct objects do not belong to the same cluster. Most of the existing clustering algorithms do not handle such information.

The remainder of this paper proceeds as follows. We first present the EM algorithm for mixture distributions and describe how to take into account partially classified data. Then we develop a more general framework for handling partially known labels. In the final section we present an example which illustrates the relevance of the approach.

2 The EM algorithm for mixture models

In the mixture model setting (Titterington et al. (1985), McLachlan and Basford (1989)), data $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ are assumed to arise independently from a random vector with density

$$f(\mathbf{x}|\Phi) = \sum_{k=1}^g p_k f_k(\mathbf{x}|\theta_k), \quad (1)$$

where $\Phi = (p_1, \dots, p_g, \theta_1, \dots, \theta_g)$, p_k are the mixing proportions ($0 < p_k < 1$ for all $k = 1, \dots, g$ and $\sum_k p_k = 1$), $f_k(\mathbf{x}|\theta_k)$ denotes a probability distribution function parametrized by θ_k , and g is the number of components.

Generally, the maximum likelihood estimation (MLE) of this model cannot be obtained analytically. The classical approach to solve this problem is the EM algorithm which provides an iterative procedure for computing MLEs in situation where, but for the absence of some additional data, estimation would be straightforward.

In order to set this problem as an incomplete-data one, we introduce as missing data the vector $(\mathbf{z}_1, \dots, \mathbf{z}_n)$ which gives the component of each \mathbf{x}_i : more precisely, $\mathbf{z}_i = (z_{i1}, \dots, z_{ig})$ where $z_{ik} = 1$ means that \mathbf{x}_i arises from the k^{th} component. ($z_{ik} \in \{0, 1\}$ and $\sum_{k=1}^g z_{ik} = 1$). Then, $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ denotes the augmented data or so-called complete data.

Starting from an initial parameter Φ^0 , an iteration of the EM algorithm consists in computing the parameter Φ^{q+1} which maximizes the function $Q(\Phi|\Phi^q) = \mathbb{E}[L(\Phi;\mathbf{y})|\mathbf{x}, \Phi^q]$, where L is the log-likelihood of Φ . This q th iteration is defined as follows:

- **E Step:** Calculate the probabilities

$$t_k^q(\mathbf{x}_i) = \frac{p_k^q f_k(\mathbf{x}_i|\theta_k^q)}{f(\mathbf{x}_i)}$$

using $\Phi^q = (p_1^q, \dots, p_g^q, \theta_1^q, \dots, \theta_g^q)$.

- **M Step:** Calculate Φ^{q+1} which maximizes

$$Q(\Phi|\Phi^q) = \sum_{i=1}^n \sum_{k=1}^g t_k^q(\mathbf{x}_i) \log p_k f_k(\mathbf{x}_i|\theta_k).$$

For Gaussian mixture we have $\Phi^q = (p_1^q, \dots, p_g^q, \theta_1^q, \dots, \theta_g^q)$ with θ_k standing for the mean vector and the variance matrix (μ_k, Σ_k) , and the M step becomes $\mu_k^{q+1} = \frac{\sum_{i=1}^n t_k^q(\mathbf{x}_i)\mathbf{x}_i}{n_k^q}$, $\Sigma_k^{q+1} = \frac{1}{n_k^q} \sum_{i=1}^n t_k^q(\mathbf{x}_i)(\mathbf{x}_i - \mu_k^{q+1})(\mathbf{x}_i - \mu_k^{q+1})^t$ and $p_k^{q+1} = \frac{n_k^q}{n}$ where $n_k^q = \sum_{i=1}^n t_k^q(\mathbf{x}_i)$.

3 Sample with labeled data

We consider now the situation where the label of some observations is known. This situation has been studied by O'Neill (1978) and Ganesalingam and MacLachlan (1978). The observed data can be denoted

$$\mathbf{y} = ((\mathbf{x}_1, \mathbf{z}_1), (\mathbf{x}_1, \mathbf{z}_1), \dots, (\mathbf{x}_m, \mathbf{z}_m), \mathbf{x}_{m+1}, \dots, \mathbf{x}_n) \quad (2)$$

and in this case the missing data are $(\mathbf{z}_{m+1}, \dots, \mathbf{z}_n)$. The log-likelihood of the complete data can be written

$$L(\Phi; \mathbf{y}) = L_m(\Phi; \mathbf{y}) + L_{n-m}(\Phi; \mathbf{y})$$

where $L_m(\Phi; \mathbf{y})$ is the log-likelihood of Φ according to the labeled observations and $L_{n-m}(\Phi; \mathbf{y})$ is the log-likelihood of Φ according to the unlabeled observations.

The function Q becomes:

$$\begin{aligned} Q(\Phi|\Phi^q) &= \mathbb{E}[L(\Phi; \mathbf{y})|\mathbf{x} = ((\mathbf{x}_1, \mathbf{z}_1), (\mathbf{x}_1, \mathbf{z}_1), \dots, (\mathbf{x}_m, \mathbf{z}_m), \mathbf{x}_{m+1}, \dots, \mathbf{x}_n), \Phi^q] \\ &= \mathbb{E}[L_m(\Phi; \mathbf{y}) + L_{n-m}(\Phi; \mathbf{y})|\mathbf{x}, \Phi^q] \\ &= L_m(\Phi; \mathbf{y}) + \mathbb{E}[L_{n-m}(\Phi; \mathbf{y})|(\mathbf{x}_{m+1}, \dots, \mathbf{x}_n), \Phi^q] \\ &= \sum_{i=1}^m \log p_{\mathbf{z}_i} f_{\mathbf{z}_i}(\mathbf{x}_i|\theta_{\mathbf{z}_i}) + \sum_{j=m+1}^n \sum_{k=1}^g t_k^q(\mathbf{x}_j) \log p_k f_k(\mathbf{x}_j|\theta_k) \end{aligned}$$

In the Gaussian case, the computation of the mean vectors becomes

$$\mu_k^{q+1} = \frac{\sum_{i|\mathbf{z}_i=k} \mathbf{x}_i + \sum_{j=m+1}^n t_k^q(\mathbf{x}_j) \cdot \mathbf{x}_j}{n_k^m + \sum_{j=m+1}^n t_k^q(\mathbf{x}_j)}$$

where n_k^m is the number of observations whose component is known and equal to k .

4 Handling partially known labels with the EM algorithm

4.1 A general framework

In this section, we present a framework generalizing the approach of section 3. The presentation of Equation 2 is clumsy and we propose to leave the presentation based on label indicator \mathbf{z}_i to adopt a more general form using the concept of authorized label set: we consider that the data has the following form $(\mathbf{x}, Z) = (\mathbf{x}_1, Z_1), \dots, (\mathbf{x}_n, Z_n)$ where Z_i corresponds to the set of authorized components for the observation \mathbf{x}_i . Formally, each Z_i is a subset of the possible label indicator $\mathcal{Z} = \{(1, 0, \dots, 0), (0, 1, \dots, 0), \dots, (0, \dots, 0, 1)\}$.

Estimating θ is then performed by maximizing the likelihood $L(\theta; \mathbf{x}, Z)$. It is easy to show that situation of paragraph 2 results from the assumption that $Z_i = \mathcal{Z}$ for all observations i , and that situation described in paragraph 3 assumes that $Z_i = \mathcal{Z}$ for all the observations with unknown label and $Z_i = \{\mathbf{z}_i\}$ is a singleton for the labelled observations.

The EM algorithm, described in section 2, can be applied with a single modification concerning the posterior probabilities computation:

$$t_k^q(\mathbf{x}_i) = \begin{cases} \frac{p_k^q f_k(\mathbf{x}_i | \theta_k^q)}{\sum_{k' \in Z_i} p_k^q f_{k'}(\mathbf{x}_i | \theta_{k'}^q)} & \text{if } k \in Z_i \\ 1 & \text{otherwise.} \end{cases} \quad (3)$$

4.2 Partially known labels

The preceding presentation allows to encompass the case where an observation is known not to belong to a given subset of the possible components. This is easily taken into account by associating to each observation \mathbf{x}_i whose label is partially known, subset Z_i defining the authorized components of the mixture for these observations. For example, if we consider a mixture of three components, the proposed framework allows to specify that some of the observations do not originate from the first component, without labeling completely these observations. Actually three possible cases can arise: $\#Z_i = g$ for unlabeled observations, $\#Z_i = 1$ for the labeled observations, and $1 < \#Z_i < g$ for the observations whose label is partially known. Taking into account this additional knowledge is straightforward and the formulation of the EM algorithm remains unchanged using the equation 3. From a practical point of view, the implementation of this approach using the formulation based on label indicators is also very simple: it is done by forcing the posterior probabilities of the partially labeled observations to zero when a component is not authorized.

5 Illustrative examples

In the presented example, 200 observations have been generated according to a four component Gaussian mixture (Figure 1 (a)). Two components of the mixture have the same mean vector but different variance matrices. With so few available observations and starting with 30 different random initializations the EM algorithm is unable to clearly identify the variances of the two nested components. Figure 1 (b) displays the best result obtained by the EM algorithm.

Figure 1 (c) shows the estimation of the parameter with the EM algorithm taking into account additional knowledge about the origin of the observations. No observation is precisely labeled but we have indicated that some of the observations originating from the component with the biggest volume do not belong to the component with the tiny volume and vice versa. This can be achieved by defining two complementary sets of authorized components for some of the observations coming from the two nested Gaussian distributions. In the example, we have four components. Some of the observations of the tiny volume component are constrained to belong to the two first components ($Z_i = \{1, 2\}$) while some of the observations drawn from the big volume component are only authorized to belong to the two last components. ($Z_i = \{3, 4\}$). Practically, this is done by forcing some of the posterior probabilities to zero. This additional knowledge dramatically improves the quality of the estimation (Figure 1 (c)).

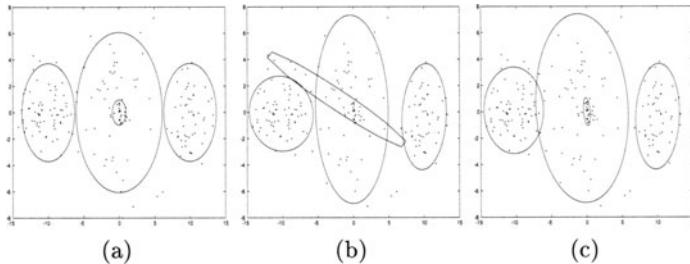


Fig. 1. Clustering with partially known labels

6 Conclusion

In conclusion, we advocate introducing additional knowledge to estimate the parameters of mixture models using the EM algorithm. In this paper, we have proposed a generalization of the mixture model estimation, which allows to handle simultaneously different kinds of information about the origin of the observations:

- unknown labels,
- completely known labels,
- or partially known labels.

The variety of possible configurations shows that mixture models are highly flexible and allow to consider some kind of *prior knowledge* in the parameter estimation. From our experience, taking additional knowledge into account in this framework has two main advantages: it accelerates the convergence of the EM algorithm and it allows to compute more reliable estimates of the mixture parameters.

To confirm the first promising results, a study based on Monte Carlo simulations and real data will be necessary.

References

- DEMPSSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977): Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, B*, 39, 1–38.
- GANESALINGAM, S. and MCLACHLAN, G.J. (1978): The Efficiency of Linear Discriminant Function Based on Unclassified Initial Samples, *Biometrika*, 65, 658–662.
- JORDAN, M.I. and JACOBS, R.A. (1994): Hierarchical Mixtures of Experts and the EM Algorithm, *Neural Computation*, 6, 181–214.
- MCLACHLAN, G.J. and BASFORD, K.E. (1989): *Mixture Models. Inference and Applications to Clustering*. Marcel Dekker, New York.
- MCLACHLAN, G. and KRISHNAN, T. (1997): *The EM Algorithm and Extensions*. Wiley, New York.
- O'NEIL, T.J. (1978): Normal Discrimination with Unclassified Observations, *Journal of the American Statistical Association*, 73, 821–826.
- TITTERINGTON, D.M., SMITH, A.F. and MAKOV, U.E. (1985): *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York.

Part II

Discrimination, Regression Trees, and Data Mining

Detection of Company Failure and Global Risk Forecasting

Mireille Bardos

Banque de France
Direction des entreprises
Observatoire des entreprises - 44 1356
31 rue Croix des Petits Champs, Paris 75001
(e-mail: mireille.BARDOS@banque-france.fr)

Abstract. Work on detecting and monitoring company risk has greatly increased over the last ten years both in credit reference agencies or associations of credit managers and in banks or public bodies which monitor the national and international economic situation. The Banque de France's work on risk is part of its credit supervision and banking system refinancing responsibilities.

Since 1982, the Banque de France's research into operational credit scoring has intensified with the use of increasingly vast and reliable databases. The aim is to describe risk in statistical terms and create medium-term forecasts for firms in France. The tools created provide both an individual risk diagnosis for each firm and an overview of company risk as a whole. This paper gives a progress report on the work carried out and some indications of the prospects for the future.

1 How credit scoring is constructed

The data files used for the most recent research on manufacturing firms are extensive. Their representativity is confirmed over a long time period where sampling bias has been corrected. Failure is defined as a legal procedure declared by the Commercial Court. To ensure the maximum amount of information the samples are composed of firms present in the Banque de France files and whose balance sheet data are available (180 000 companies each year).

The variables are quantitative data calculated with the balance sheet data: economical and financial ratios. The preselection of the discriminant ratios which provide advance warning of company failure (Bardos 1995) is based on non-parametric methods which determine confidence intervals for the quantiles on each category (failing and non-failing companies).

The classic Fisher linear discriminant analysis technique was chosen as a result of comparative studies because of its numerous advantages, including: robustness over time, interpretability, simple probabilistic utilization, easy maintenance (Gnanadesikian et al., 1989, Mc Lachlan, 1992, Bardos & Zhu, 1998).

The principle of Fisher linear discriminant analysis consists of finding the optimum frontier between failed and non-failed companies, which in this

case is a hyperplane separator. On the preselected ratios, it calculates the coefficients in the scoring function. It is given by the following formula:

$$f(a) = (\mu^N - \mu^D)'T^{-1}(a - \frac{\mu^N + \mu^D}{2}) = \sum_{j=1}^{j=p} \alpha_j(a_j - p_j)$$

where N is the group of non-failing companies, D is the group of failing companies,

$a = (a_1, a_2, \dots, a_p)$ is the vector of the firm's p ratios,

T is either the total variance-covariance matrix or the within variance covariance matrix,

$\alpha = (\mu^N - \mu^D)'T^{-1}$ is the p coefficients vector ($\alpha_1, \alpha_2, \dots, \alpha_p$),

$p_j = \frac{\mu_j^N + \mu_j^D}{2}$ is the pivot value for the ratio j ,

$\alpha_j(a_j - p_j)$ is the contribution of ratio j to the score $f(a)$.

The score $f(a)$ can be calculated as the sum of ratio contributions which highlight a firm's strong and weak points. This therefore provides a risk interpretation guide when examining a firm's balance sheet.

The decision rules are based on metric criteria, or Bayes' decision criteria.

An important phase of this work is the validation of the scoring function by testing it on a large number of test samples. In particular, care is taken to ensure high percentages of correct allocation rates that are stable over time for both categories of firms. Since the construction has been done, quality control tests have been carried out each year. This score has been constructed in 1997, it is called BDFI, it controlled each year and used for individual diagnosis and global portfolio analysis.

2 Posterior probability of failure

The full value of a diagnosis of a firm is only apparent when it is compared with other firms. This is provided by calculating a firm's posterior probability of failure. The term "posterior" here means that the firm's score is known.

The calculation, based on Bayes' theorem, requires knowledge of prior probability, π_D estimated by the annual failure rate in a manufacturing sector. The horizon of the forecast is three years and the prior probability of failure is taken as $3\pi_D$, under the strong hypothesis of stability of industrial environment.

The Banque de France, unlike commercial banks, aims to describe the risk of failure and not to take a decision on whether to grant credit. There is therefore no point in defining decision-making thresholds, or unequal error costs. Conversely, if posterior probabilities per score interval are known, risk classes can be defined, corresponding to rather homogeneous regions.

A firm is considered at risk when its score belongs to an interval where the posterior probability of failure is much higher than the prior probability of failure. The score value intervals can then be classified into at-risk intervals, sound intervals and neutral intervals. For the score, seven classes have been defined and a posterior probability of failure is assigned to each class. The risk

scale goes from 1 to 40 and it gives informations on very diverse situations (Bardos, 1998a, 1998b).

The accuracy of the estimation of the probability of failure is an important issue. Two methods have been used: means and standard error over several years, and bootstrap.

3 Individual risk diagnosis

Individual risk diagnosis is carried out on the basis of historical monitoring of scores and the associated probability of failure. For a firm, if the contribution of a ratio is *negative*, it indicates a *weakness*; if it is *positive* it indicates a *strength*. In addition, the firm is situated in comparison with the quartiles of the firm's sector of activity. Scores and trends in scores can be explained by the contribution of the ratios for which sectoral quartiles are also available.

Scoring thus provides an analysis of the firm's risk on the basis of its accounts. This information can and should then be supplemented by other information, available to the Banque de France: ratings, defaults, risks, information from the branch, etc. The combination of all these complementary approaches gives a diagnosis that is both reliable and detailed.

The advantage of this arrangement is to permit dialogue with the firm. To support this aim, Banque de France branches have computer-generated screens which display complete risk analysis for several years and serve as a guide when examining a firm's balance sheet.

4 Use of credit scoring for credit supervision

One very important application of the scores derived from the probabilistic approach is the analysis of a portofolio of commitments. The probability of failure assigned to each firm allows an overall risk assessment to be made for a given population. This type of analysis opens up a range of options, since it makes it possible to analyse the risk of a population of firms. It can be used to describe not only the situation of a sector, or the clients of a commercial bank, but also commitments to firms in the banking system as a whole.

A very simple case can give an idea of the possibilities inherent in such an approach. Several items of information are required in order to assess the potential loss for a commercial bank at a given time in the future. This assessment may be made for a given type of exposure.

If E_i is the current debt of the borrower i , p_i its probability of failure at the time horizon studied, γ_i its non-repayment coefficient at the future time studied (which depends on the collateral and rate of recovery), and n the number of borrowers, it is possible to calculate the average loss and the average part of the commitments at risk.

The average loss at the future time is:

$$\mu = \sum_{i=1}^{i=n} p_i \gamma_i E_i$$

The average part of the loans at risk is:

$$r = \frac{\sum_{i=1}^{i=n} p_i \gamma_i E_i}{\sum_{i=1}^{i=n} E_i}$$

From the standpoint of the RAROC method (Bessis, 1995) these two indicators are important. The average loss is known as economic provision. It estimates a statistical loss.

However, actual payment defaults may be different. The loss may exceed the upper bound: since commitments are highly concentrated, failure may affect some of the biggest commitments. To assess what happens in a worst case scenario, risk classes offer the possibility of an instructive estimation.

For lenders, it is important to relate the potential risk represented by the firm to the amount of the loan. This link with the probability of failure is also of interest for other variables V such as the number of employees or value added.

For all firms and for each variable V , the at-risk part is defined by the equation:

$$P = \frac{\sum V_i p_i}{\sum V_i}$$

In this equation, P is the average of p_i weighted by V_i . The p_i are random because they depend on the year and the sample. The average of p_i , \bar{p}_i , has been estimated over several years in order to ensure its robustness. The standard deviation of \bar{p}_i , σ_i , has also been estimated using data for several years and will be used to estimate a confidence interval.

It is extremely important to know the maximum at-risk part in order to assess a lender's risk because it represents what may happen in the worst-case scenario. This is an important research scheme.

4.1 Measurement of posterior probability taking size into account

Analysis shows that the score, called BDFI, discriminates firms in the same way regardless of size. The same indicators are used and the distribution of scores does not differ according to size, whether failing firms or non-failing firms are considered. Therefore the probability of failure according to size can be estimated using the same risk classes.

In order to estimate posterior probabilities by size category, it is necessary to evaluate prior probabilities by size, and Bayes' theorem is applied to the size category t .

$$P(D/BDFI \in r, t) = \frac{P(BDFI \in r/D, t) \pi_D(t)}{P(BDFI \in r/D, t) \pi_D(t) + P(BDFI \in r/N, t) \pi_N(t)}$$

where r is the score interval defining the risk class, t is the size category, D denotes failure, N non-failure, and $\pi_N(t) = 1 - \pi_D(t)$.

Each firm belongs to a risk class according to its BDFI score and to a size category according to its turnover. The posterior probability of failure corresponding to that case is therefore assigned to the firm.

4.2 Comparative forecasts of potential losses

It is possible to compare the at-risk parts estimated using both methods, i.e., one method which does not take size into account (Table 1) and one which does (Table 2). The estimated average at-risk part using posterior probabilities according to size for the four size categories and for all firms.

The comparison with the real failure rate in the three years (Table 3) confirms the quality of the measurement of failure risk given by the BDFI score and the probability of failure, provided that size is taken into account (Bardos 1998a, Bardos & Plihon, 1999).

Table 1 : Average portion at risk of the different variables examined for a three-year forecast horizon without taking the firm size into account

Size class in FRF millions of sales	1 to 5	5 to 50	50 to 100	over 100	Aggregate
Number of firms	12.3	9.5	8.4	7.1	9.3
Number of employees	12.5	9.9	9.0	6.6	7.6
Added value	11.2	8.7	7.7	5.5	6.3
Borrowings and liabilities	17.2	12.2	11.3	8.3	8.9

Source and production : Banque de France – Fiben - Companies Observatory

Table 2 : Average portion at risk of the different variables examined for a three-year forecast horizon taking the firm size into account

Size class in FRF millions of sales	1 to 5	5 to 50	50 to 100	over 100	Aggregate
Number of firms	10.2	7.3	4.6	2.2	6.7
Number of employees	10.5	7.5	5.0	2.0	3.6
Added value	9.4	6.6	4.3	1.6	2.8
Borrowings and liabilities	14.0	9.4	6.3	2.5	3.6

Source : Banque de France – Fiben

Production : Banque de France – Companies Observatory

Last update February 1998

Table 3: Value at risk

Size class in FRF millions of sales	1 to 5	5 to 50	50 to 100	over 100	Aggregate
Number of firms	10.7	7.5	4.9	1.8	6.8
Number of employees	12.7	8.4	5.9	0.8	3.2
Added value	10.6	6.8	4.8	0.5	2.1
Borrowings and liabilities	10.9	7.7	5.4	0.6	1.7

Source : Banque de France – Fiben
 Production : Banque de France – Companies Observatory
 Last update March 1998

5 Conclusion

The tools developed by the Banque de France currently cover a very large proportion of credit risk. Future plans include the development of a comprehensive detection procedure based on the files of the Companies Division at the Banque de France. It will be used both for individual diagnosis and for a general approach to the risk of company failure, and will thus fully contribute to the ways in which the Banque de France fulfils its responsibilities of prudential control and credit supervision.

References

- M. BARDOS, D. PLIHON (1999): Detection of risk sectors. The IRISK method, Banque de France, Bulletin n° 69, septembre 1999.
- M. BARDOS (1998a): Risque et taille des entreprises industrielles (Risk and size of industrial firms), Banque de France, direction des Entreprises, PECON 832-9805.
- M. BARDOS (1998b): Detecting the risk of company failure at Banque de France, Journal of Banking and Finance, n° 22, 1998.
- M. BARDOS (1998c): Le score BDFI : du diagnostic individuel à l'analyse de portefeuille (The BDFI Score: from individual diagnosis to portfolio analysis), Les études de l'Observatoire des entreprises, Banque de France, 1998.
- M. BARDOS, W.H.ZHU (1998): Comparison of linear discriminant analysis and neural networks, application for the detection of company failures,in Biomimetic approaches in management science, Kluwer Academic Publishers, 1998.
- M. BARDOS (1995): Les défaillances d'entreprises dans l'industrie : ratios significatifs, processus de défaillances, détection précoce (Company failures in industry: significant ratios, failure processes, early detection), Collection Entreprise B 95/03, Banque de France.
- J. BESSIS (1995): Gestion des risques et gestion actif-passif des banques .
- R. GNANADESIKAN and panel of authors (1989): Discriminant Analysis and Clustering, Statistical Science, vol. 4, N° 1, p.34-69.
- G.J. Mc LACHLAN (1992): Discriminant Analysis and Statistical Pattern Recognition, Wiley, New-York.

Discriminant Analysis by Hierarchical Coupling in EDDA Context

Isabel Brito^{1,2} and Gilles Celeux²

¹ Department of Mathematics, Instituto Superior de Economia e Gestão,
Rua do Quelhas 6, 1200-781 Lisboa, Portugal

² Inria Rhône-Alpes,
655 av. de l'Europe, 38330 Montbonnot St. Martin, France
(e-mail: isabel.brito@inrialpes.fr)

Abstract. Friedman(1996) proposed a strategy for the classification multigroup problem. He build independently a classifier for each pair of classes and then combined all the pairwise decisions to form the final decision. We suggest an alternate approach in the context of EDDA models. Our technique, the hierarchical coupling, is also based on pairwise decisions but we abandon the independence and work on nested pairs of classes. We evaluate the performance of hierarchical coupling on simulated and real datasets.

1 Introduction

Discriminant analysis through eigenvalue decomposition EDDA (Bensmail and Celeux, 1996) has been defined for Gaussian populations. Let the population be partitioned in K classes, each of them with a priori probability π_k . Each population entity is characterized by a vector $\mathbf{x} = (x_1, x_2, \dots, x_d)$ of d quantitative feature variables. Each class has density $f_k(\mathbf{x})$, ($k = 1, \dots, K$) which is supposed to be Gaussian with vector mean μ_k and variance matrix Σ_k . The vector parameter $\theta_k = (\mu_k, \Sigma_k)$ is estimated from the training set \mathbf{z} of n observations. Each observation is denoted $\mathbf{z}_i = (\mathbf{x}_i, y_i)$, ($i = 1, \dots, n$) where \mathbf{x}_i are the feature measurements for entity i and $y_i \in \{1, \dots, K\}$ denotes its class origin.

EDDA asks for the variance matrix Σ_k eigenvalue decomposition $\Sigma_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k$ where $\lambda_k = |\Sigma_k|^{1/d}$, \mathbf{D}_k is the eigenvectors matrix of Σ_k and \mathbf{A}_k is a diagonal matrix such that $|\mathbf{A}_k| = 1$, with the normalized eigenvalues of Σ_k on the diagonal in a decreasing order. This decomposition is interesting because the discrimination is based on meaningful parameters. Actually, λ_k denotes the volume of the k th class, \mathbf{A}_k its shape and \mathbf{D}_k its orientation. Different assumptions on these parameters lead to eight elliptical models: $[\lambda \mathbf{D} \mathbf{A} \mathbf{D}']$, $[\lambda_k \mathbf{D} \mathbf{A} \mathbf{D}']$, $[\lambda \mathbf{D} \mathbf{A}_k \mathbf{D}']$, $[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$, $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$, $[\lambda_k \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$, $[\lambda \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$ and $[\lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k]$. The absence of subscript k means that the parameter in question has the same value for all classes and its presence that the parameter is free for the K classes. For a diagonal variance matrix only

the volume and the orientation are of interest. Denoting \mathbf{B} the diagonal normalized variance matrix, four diagonal models appear $[\lambda\mathbf{B}]$, $[\lambda_k\mathbf{B}]$, $[\lambda\mathbf{B}_k]$ and $[\lambda_k\mathbf{B}_k]$. Assuming models with spherical shapes there are $[\lambda\mathbf{I}]$ and $[\lambda_k\mathbf{I}]$ where \mathbf{I} denotes the identity matrix. EDDA gives raise to a total of fourteen models and selects the model that minimizes the sample-based estimate of classification error by cross-validation(leaving-one-out method).

We discuss a method for polychotomous classification which inspiration comes from Friedman(1996) and is taken back by Hastie and Tibshirani(1996). Friedman emphasizes that for a K class problem the classifiers are easier to interpret and even to estimate in the case $K = 2$ than in the case $K > 2$. So, he proposed the following approach:

1. On the partition set of the classes built all the $(K \choose 2)$ possible dichotomous problems.
2. Solve each of the two-class problems.
3. Combine all the pairwise decision to form a K-class decision. The combination rule is to assign a test observation to the class that wins the most pairwise comparisons.

Our strategy is somewhat different. We also decompose a polychotomous problem in several dichotomous problems but the way of building both these two-class problems and the classification rule are different. In our procedure the dichotomous problems are nested and we structure them in binary trees. It is why we call it hierarchical coupling. The classification rule determines a branch tree which corresponds to the class the observation test is assigned to.

The motivation for that strategy joins simplicity and stability. The simplicity comes from the dichotomous decomposition; the stability comes from the models combination because the rule does not depend only on one model.

We will apply hierarchical coupling in EDDA context on both real and simulated datasets.

2 Hierarchical coupling

Let $\mathcal{C} = \{C_1, C_2, \dots, C_K\}$ be the set of classes. Consider a partition of \mathcal{C} in two elements which is called the first level partition $P_1 = \{G_1^1, G_1^2\}$. We denote group an element of this binary partition G_i^l , l indicating the group ($l = 1, 2$) and i the hierarchical level. There are groups containing only one class which we name simple groups and groups containing several classes named composed groups. For instance, in a four-class problem some of the possible partitions are $\{C_1, \{C_2, C_3, C_4\}\}, \{C_2, \{C_1, C_3, C_4\}\}$ or $\{\{C_1, C_2\}, \{C_3, C_4\}\}$.

At the first level there are $2^{K-1} - 1$ possible partitions. A decision rule is to be applied at this first level in order to choose the best partition P_1^* . For the composed groups of P_1^* another binary partition P_2 is designed in a

similar way. Again the decision rule is to be applied to choose the best partition. The procedure is repeated until all the groups of the actual partition are simple groups. A binary tree represents the procedure. At the last level the tree has K branches.

Until now we described the way of building the tree. To move around we need at each level to solve a two-class discriminant problem where the classes are the groups G_1 and G_2 . In that purpose we assume a two component mixture of Gaussian data. The composed groups are supposed arising from the same mixture component, and we make use of EDDA to separate them.

So, hierarchical coupling asks for a two-fold decision rule $r(\mathbf{x})$ at each level. A choice is to be made for the best partition and the best model. The decision rule we propose is as follows:

1. For all possible partitions all the fourteen EDDA models are estimated.
2. At the first level we have $14(2^{K-1} - 1)$ pairs (model, partition) and we choose the one for which the cross-validated (leaving-one-out) model classification error is minimum. For tied pairs (with the same cross-validated model classification error) the criterion is to choose the pair for which the model has the smallest number of parameters.
3. Once the first level partition is chosen we go to the second level and again the choice of step 2 is to be made.

The procedure is repeated for all levels of the tree. A hierarchical combined model results from this procedure. This model is also evaluated with the cross-validated classification error.

For example, suppose that the hierarchical combined model for a four class problem is the one represented in Figure 1.

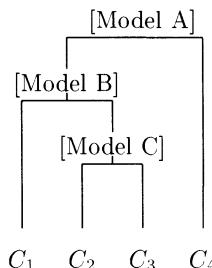


Fig. 1. Example of Hierarchical coupling.

When a new observation arrives it passes through model A which classifies in C_4 or else. If model A classifies the observation in C_4 the analysis stops.

Otherwise the observation passes through model B and the decision is C_1 or not. If model B does not classify the observation in C_1 it passes finally through model C and it is allocated to C_2 or C_3 .

3 Examples on real datasets

We used hierarchical coupling on two benchmarks datasets: the Iris and the Crabs datasets (<http://www.stats.ox.ac.uk/pub/MASS2>). Just remember that the famous Iris dataset is concerned with $K = 3$ classes and $d = 4$ features. The Crabs dataset is available on specimens of each sex of each of two species blue and orange ($K = 4$ classes) crabs. Each specimen has measurements on $d = 5$ body items.

3.1 Iris dataset

For the Iris dataset the hierarchical coupling combined model is visualized in Figure 2. At the first level the model $[\lambda_k \mathbf{I}]$ - a spherical one with different

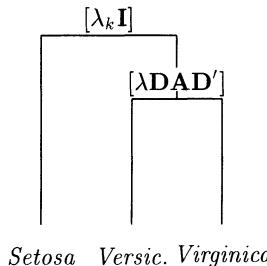


Fig. 2. Hierarchical coupling for Iris Dataset. The chosen model is indicated at each node.

volumes - best discriminates *Setosa* versus the group $\{\text{Versicolor}, \text{Virginica}\}$ and at the second level the linear model $[\lambda \mathbf{DAD}']$ best discriminates between *Versicolor* and *Virginica*. The hierarchical coupling cross-validated classification error is 0.02. It is the same error of the linear model for three classes. The linear model is the best model according to EDDA method.

Note that the hierarchical procedure leads to a quite simple decision rule for separating *Setosa* and $\{\text{Versicolor}, \text{Virginica}\}$.

3.2 Crabs dataset

The dataset is depicted in Figure 3 where the axes are the two first canonical variates.

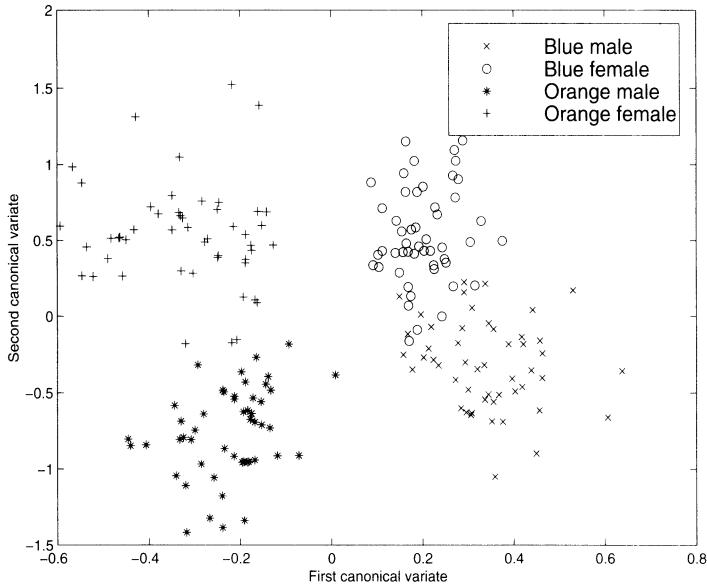


Fig. 3. Plot of Crabs Dataset on the two first canonical variates.

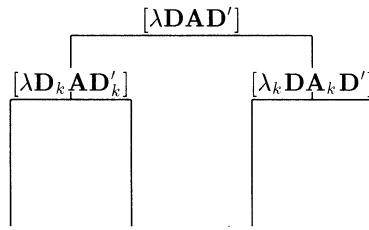


Fig. 4. Hierarchical coupling for Crabs Dataset. The chosen model is indicated at each node.

The hierarchical coupling tree is presented in Figure 4. At the first level hierarchical coupling discriminates between species and chooses the linear model. At the second level the model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ - elliptical model with different orientations - separates blue males and females. While at the same level is the model $[\lambda_k \mathbf{D} \mathbf{A}_k \mathbf{D}']$ - elliptical one with the same orientation - which discriminates between orange males and females. The hierarchical coupling cross-validated classification error rate is 0.045. For the four class problem EDDA method chooses the model $[\lambda \mathbf{D}_k \mathbf{A} \mathbf{D}'_k]$ with the same cross-validated error rate. But here again, the hierarchical procedure is meaningful and throws an interesting light on the crabs classification.

4 Results on simulated data

We present here an illustrative simulated situation where hierarchical coupling improves over the selection of EDDA model for K classes. We generated artificial data from a two-dimensional normal population. The training sample size is 150 divided into 3 equal sized classes. All the classes have the same volume. One of the classes is spherical with center in (0,0) and the other two classes are elliptical ones with centers in (4,0) and (6,4). These two classes have the same shape but different orientations. For 150 replications of the experiment the cross-validated mean classification error rate resulting from hierarchical coupling is 0.0915 while the best EDDA model for three classes leads to a cross-validated mean classification error rate of 0.0974. The improvement does not seem to be really sensitive.

5 Discussion

In the Crabs example, for instance, and looking at Figure 3 the proposed combined model (Figure 4) is quite reasonable. The two species are discriminated with a linear model. Inside the blue crabs group is an elliptical model with different orientations that separates males and females. Inside the orange crabs group is also an elliptical model but with different shapes and volumes.

It seems from the presented examples that hierarchical coupling is a promising method. It performs well and its easily interpretable representation is appealing. In fact, other experiments not reported here show that the main interest of the hierarchical procedure is not improved error rates but rather more readable decision rules. In particular, situations where hierarchical coupling can be fruitful are the following: large K values and some classes well-separated and some other poorly separated.

References

- BENSMAIL, H. and CELEUX, G. (1996): Regularized Gaussian Discriminant Analysis through Eigenvalue decomposition. *Journal of the American Statistical Association*, 91, 1743–48.
- FRIEDMAN, J. (1996): Another approach to polychotomous classification. *Technical Report*. Stanford University.
- HASTIE, T. and TIBSHIRANI, R. (1996): Classification by pairwise coupling. *Technical Report*. University of Toronto.

Discrete Discriminant Analysis: The Performance of Combining Models by a Hierarchical Coupling Approach*

Ana Sousa Ferreira¹, Gilles Celeux² and Helena Bacelar-Nicolau¹

¹ LEAD, Faculdade de Psicologia e Ciências da Educação,
Alameda da Universidade, 1649-013 Lisboa, Portugal

² INRIA-Rhône Alpes, Grenoble, France

Abstract. We are concerned with combining models in discrete discriminant analysis in the multiclass ($K > 2$) case. Our approach consists of decomposing the multiclass problem in several biclass problems embedded in a binary tree. The affinity coefficient (Matusita (1955); Bacelar-Nicolau (1981,1985)) is proposed for the choice of the hierarchical couples, at each level of the tree, among all possible forms of merging. For the combination of models we consider a single coefficient: a measure of the relative performance of models - the integrated likelihood coefficient (Ferreira et al., 1999)) and we evaluate its performance.

1 Introduction

Let $X = (x_1, \dots, x_n)$ denote a n -dimensional training sample of multivariate observations, associated with p discrete variables, for which each object is assumed to become from one of K exclusive groups G_1, G_2, \dots, G_K with prior probabilities $\pi_1, \pi_2, \dots, \pi_K$ ($\sum_{l=1}^K \pi_l = 1$). The Bayes classification rule assigns an individual vector x into G_k if $\pi_k P(x | G_k) \geq \pi_l P(x | G_l)$ for $l = 1, \dots, K, l \neq k$, where $P(x | G_l)$ denotes the conditional probability function for the l -th group. Usually, the conditional probability functions are unknown and are estimated on the basis of the training sample.

In this paper, we are concerned with discrete discriminant analysis in the small sample, multiclass ($K > 2$) setting. Most of the discrete discrimination methods perform poorly due to the high-dimensionality problem, which becomes even more complex in the case of $K > 2$ groups. We proposed a method, inspired from Friedman's approach (Friedman (1996)), of reducing the multiclass problem in several biclass problems in discrete discriminant analysis (Ferreira et al., 1999):

- We decompose the multiclass problem in several biclass problems using a structure of a binary tree (Hierarchical Coupling). The individual vector

* This work has been partially supported by the Franco-Portuguese Programme VECEMH (Embassy of France and Portuguese Ministry of Science and Technology - ICCTI) and the CEAUL/FCUL.

x is assigned to the group associated to the terminal node of the tree to which x is belonging.

- For each biclass problem, the classification rule is based on a combining model. This method has an intermediate position, between the full multinomial model and the first order independence model hereunder described. The main aim of this approach is to obtain a better prediction performance and more stability of the results.

We call this model the Hierarchical Model (HIERM). For the sake of simplicity, we focus on binary data.

2 Discrete discriminant analysis

For discrete data, the most natural model is to assume that the conditional probability function $P(x | G_k)$, $k = 1, \dots, K$, where $x \in \{0, 1\}^p$ are multinomial probabilities. In this case, the conditional probabilities are estimated by the observed frequencies. This model is called the full multinomial model (FMM) and involves $2^p - 1$ parameters in each group. Hence, even for moderate p , not all of the parameters are identifiable.

One way to deal with the curse of dimensionality consists of reducing the number of parameters to be estimated. The first-order independence model (FOIM) assumes that the p binary variables are independent in each group G_k , $k = 1, \dots, K$. Then, the number of parameters to be estimated for each group is reduced from $2^p - 1$ to p . This method is simple but may be unrealistic in some situations.

Since we are mainly concerned with small or very small sample sizes, we encounter a problem of sparseness in which some of the multinomial cells may have no data in the training sets.

Hand (1982) noticed that the choice of the smoothing method is not very important so that computationally less demanding methods may be used. Then, we suggest smoothing the observed frequencies as follows:

$$\hat{P}(x | \lambda, B) = \frac{1}{n} \sum_{i=1}^n \lambda^{p - \|x - x_i\|} (1 - \lambda)^{\|x - x_i\|},$$

where $\lambda = 1.00$ (no smoothing) or $\lambda = .99, .95$ or $.90$ (smoothing according to the training sample size).

3 Hierarchical coupling

The hierarchical coupling needs two decisions at each level:

1. The choice of the decomposition in a biclass problem among the $2^{K-1} - 1$ possible group couples.

2. The choice of the model that gives the best classification rule for the chosen couple.

In the beginning we dispose of K training subsamples and we want to reorganise these K classes into two classes. So, we propose to select the two new classes that are best separable.

The affinity coefficient (Matusita, 1955; Bacelar-Nicolau 1981,1985) is used to choose the decomposition of each level of the tree: Suppose $F_1 = \{p_j\}$ and $F_2 = \{q_j\}$, $j = 1, 2, \dots, p$ are two discrete distributions defined on the same space, the affinity coefficient between F_1 and F_2 is given by $\rho(F_1, F_2) = \sum_{j=1}^p \sqrt{p_j} \sqrt{q_j}$.

After choosing the first level of the tree, we derive the decision rule for these two classes. Then, we repeat the choice of the second level of the tree among the composed classes (formed with 2 or more initial groups). The process stops when a decomposition of classes leads to simple groups.

4 Combining models in discrete discriminant analysis

The idea of combining models is nowadays present in several papers (Celeux and Mkhadri (1992); Breiman (1995); LeBlanc and Tibshirani (1996); Raftery (1996)). The aim of this strategy is to obtain more robust and stable models.

In this article, we investigate the performance of an approach inspired from the ideas of Leblanc and Tibshirani (1996). We consider a simple linear combination $\sum_m P_k^m(x)\beta_m$ where $P_k^m(x)$ is the conditional probability function for the k -group and the model m . Thus, the problem is to estimate the coefficients of the combination. Different strategies to obtain the estimates of β_1, \dots, β_M can be proposed: the relative fit estimator, the least squares estimator obtained by cross-validation, etc.

An intuitive combination method is to propose a single coefficient producing an intermediate model between the full multinomial model and the first order independence model:

$$\hat{P}_k(x | \beta) = (1 - \beta)\hat{P}_M(x | G_k) + \beta\hat{P}_I(x | G_k)$$

We proposed β ($0 < \beta < 1$) (Ferreira et al. (1999)) as a measure of the relative performance of the FOIM model that takes account of model uncertainty (Raftery (1996)): $\beta = \frac{VFOIM}{VFOIM+VFMM}$, where VFOIM and VFMM are the integrated likelihood for the FOIM model and for the FMM model.

Using non-informative Jeffreys prior distributions, we obtained the expressions for VFOIM and VFMM (Ferreira et al. (1999)). We are now in position to evaluate the efficiency of this method.

5 Numerical experiments

The efficiency of this model has been investigated on both real and simulated binary data. However, for the sake of simplicity, we only present, in this

article, the application to real data, that is concerned with our professional applied field: Psychology and Education Sciences.

We investigate the performance of the present approach of hierarchical coupling, compared with others models: FOIM, FMM and KER (Kernel Discriminant Analysis), where the smoothing parameter has been $\lambda = .99$.

5.1 Psychological data

The Psychological Data set consists of 34 dermatology's patient evaluated by a Psychological Test set. The whole sample was divided into three groups (G_1 : Nonalexithymics Group, G_2 : Alexithymics Group, G_3 : Intermediate Group), according to the value obtained in the psychological test TAS-20 (Twenty Item Toronto Alexithymia Scale), which is supposed to evaluate the presence of alexithymia ("alexithymia" means having no words to express emotions). For each subject, the values of six binary variables of another psychological test (the Rorschach test) were available. These data have been analysed by a psychological team of the Faculty of Psychology and Education Sciences, University of Lisbon (Nina Prazeres and Professor Danilo Silva) in a context of a master's degree project (Prazeres (1996)).

Since the sample is very small, we use the whole sample for choosing of the hierarchical coupling and we estimate the misclassification risk by half-sampling. Table 1 summarises the results of the choice of the hierarchical coupling by the affinity coefficient and Table 2 summarises the results of the four methods for this data set.

For each method, we give the misclassification risk, estimated by half-sampling. The prior probabilities were taken to be equal, $\pi_k = .5(k = 1, 2)$.

Table 1: Results of the hierarchical coupling

1 st choice	G_3 v.s. $G_1 + G_2$
2 nd choice	G_1 v.s. G_2

Table 2: Estimated set misclassification risk and parameters values for the psychological data

	FOIM	FMM	KER	HIERM	HIERM
Half-sampling	53%	71%	65%	35%	35%
λ			.99	1.00	.99
β	1 st ch.			.1780	.1780
	2 nd ch.			.4415	.4415

The first decomposition chosen by the hierarchical coupling, suggest that the union of the extremes groups forms a well-separated class from the class composed by the intermediate group, since these subjects obtained balanced scores.

Since the data set is very sparse ($2^6=64$ states and only 17 observations) the HIERM method provides the lowest estimated misclassification risk.

5.2 Psychological counselling career data

The Psychological Counselling Career data set consists of 600 students of the 1st and 2nd forms of four licenciature's degree: Biology (G_1), Psychology (G_2), Language and Literature (G_3) and Engineering (G_4).

The Psychological Questionnaire (*My Vocational Situation*) is organised in three scales: Vocational Identity (VI) with 18 items; Occupational Information (OI) with 4 items; Barriers (B) with 4 items.

For each student were available the values of: six variables of the VI scale; the four variables of OI scale; the four variables of B scale. These data have been analysed by a psychological team of the Faculty of Psychology and Education Sciences, University of Lisbon (Rosário Lima and Professor Ferreira Marques) in the context of a doctor's degree project (Lima, 1998).

We drew at random a training sample of 200 students and the rest constituted the test sample. Table 3 summarises the results of the choice of the hierarchical coupling by the affinity coefficient and Table 4 summarises the results of the four methods for this data set.

For each method, we give the misclassification risk, estimated on the test sample. The prior probabilities were taken to be equal, $\pi_k = .25$ ($k = 1, 2, 3, 4$).

Table 3: Results of the hierarchical coupling

	1 st choice	2 nd choice	3 rd choice
VI scale	G_1 v.s. $G_2 + G_3 + G_4$	G_4 v.s. $G_2 + G_3$	G_2 v.s. G_3
OI scale	G_4 v.s. $G_1 + G_2 + G_3$	G_2 v.s. $G_1 + G_3$	G_1 v.s. G_3
B scale	$G_1 + G_3$ v.s. $G_2 + G_4$	G_1 v.s. G_3	G_2 v.s. G_4

Table 4: Test set misclassification risk and parameters values for the psychological counselling career data

		FOIM	FMM	KER	HIERM	HIERM
VI scale	Test	69%	75%	73%	39%	39%
	λ			.99	1.00	.99
	β	1 st			.9857	.9857
		2 nd			.9998	.9998
		3 rd			.9999	.9999
	Test	66%	67%	65%	42%	40%
	λ			.99	1.00	.99
OI scale	β	1 st			.0000	.0000
		2 nd			.0002	.0002
		3 rd			.0006	.0006
	Test	66%	67%	65%	53%	53%
	λ			.99	1.00	.99
	β	1 st			.9999	.9999
		2 nd			.9999	.9999
B scale		3 rd			1.0000	1.0000
	Test	66%	67%	65%	53%	53%
	λ			.99	1.00	.99
	β	1 st			.9999	.9999
		2 nd			.9999	.9999
		3 rd			1.0000	1.0000

The first decomposition chosen by the hierarchical coupling for the several scales, suggests that: Biology' students are different from the others students in what concerns the definition of a clear and stable picture of their goals and interests; Engineering students reveal a distinct need for vocational information from the others students; the students of odd groups show individual's perceived external obstacles or limitations in pursuing occupational goals different from the students of even groups.

Remark that this data set is not very sparse ($2^6=64$ or $2^4 =16$ states and 200 observations), but again the HIERM method provides markedly the lowest test estimates of the misclassification risk.

6 Concluding

We have presented a method for the multiclass case in discrete discriminant analysis. As we know, sparseness is the main problem of discrete discriminant analysis, particularly in the multiclass case. The numerical experiments showed that HIERM could be one promising tool to improve dramatically the power of discrimination in the multiclass, high-dimensional setting.

References

- BACELAR-NICOLAU, H. (1981): Contributions to the Study of Comparison Coefficients in Cluster Analysis. *PhD Th. (in Portuguese)*, Univ. Lisbon.
- BACELAR-NICOLAU, H. (1985): The Affinity Coefficient in Cluster Analysis. *Meth. Oper. Res.* , 53, 507-512.
- BREIMAN, L. (1995): Stacked Regression. *Machine Learning*, 24, 49-64.
- CELEUX, G.; MKHADRI, A.(1992): Discrete Regularized Discriminant Analysis. *Statistics and Computing*, 2, 143-151.
- FERREIRA, A.; CELEUX, G. and BACELAR-NICOLAU, H. (1999): Combining Models in Discrete Discriminant Analysis by an Hierarchical Coupling Approach. *Proceedings of the IX International Symposium of ASMDA*, 159-164.
- FRIEDMAN, J. H. (1996): Another Approach to Polychotomous Classification. *Technical Report, Stanford University*.
- HAND, D.J. (1982): *Kernel Discriminant Analysis*. Research Studies Press. Wiley, Chichester.
- KASS, R. E., RAFTERY, A. E. (1995): Bayes Factor. *Journal of the American Statistical Association*, 90, 773-795.
- LEBLANC, M., TIBSHIRANI, R. (1996): Combining Estimates in Regression and Classification. *Journal of the American Statistical Association*, 91, 1641-1650.
- LIMA, M. R. (1998): Orientação e Desenvolvimento da Carreira em Estudantes Universitários. *Tese de Doutoramento*, Univ. Lisbon.
- MATUSITA, K. (1955): Decision Rules Based on Distance for Problems of Fit, Two Samples and Estimation., *Ann. Inst. Stat. Math.*, 26, no4, 631-640.
- PRAZERES, N.L. (1996): Ensaio de um Estudo sobre Alexitimia com o Rorschach e a Escala de Alexitimia de Toronto (TAS-20). *Tese de Mestrado*, Univ. Lisbon.
- RAFTERY, A. E. (1996): Approximate Bayes Factors and Accounting for Model Uncertainty in Generalised Linear Models. *Biometrika*, 83, 251-266.

Discrimination Based on the Atypicality Index versus Density Function Ratio

H. Chamlal and S. Slaoui Chah

Département de Mathématiques et Informatique, Faculté des Sciences,
B.P : 1014, Rue Ibn Batouta, Rabat-Maroc
(e-mail: chamlal@yahoo.com or c.hasna@caramail.com)

Abstract. We propose a method of discrimination, based on the atypicality index and the density function. After a short survey of the atypicality index, we show that the presence of "critical regions", when we apply the bayesian quadratic discrimination, under some hypotheses, leads to misclassifications. The performance of the proposed method versus quadratic and linear discrimination is assessed via simulation. It is generally shown that the discrimination based on the ratio (atypicality index/density function) consistently yields noticeably higher percentage of well classified individuals relative to the traditional methods. The method is illustrated with a numerical example and is compared to quadratic discrimination.

1 Atypicality index

Consider an n-sample E , partitioned into m clusters : G_1, G_2, \dots, G_m . Suppose that the explanatory vector is $X = (X_1, X_2, \dots, X_p)^t$ with density function $f(x/r)$ in the cluster G_r , ($r = 1, 2, \dots, m$).

Definition 1. (Nakache (1980)) *Given two individuals i and j , i is more typical of cluster G_r than j if and only if $f(i/r) > f(j/r)$.*

Definition 2. (Nakache (1980)) *We define the atypicality index of the cluster G_l , assigned to the individual i by : $ind(i, l) = \sum_{j \in E, f(j/l) > f(i/l)} f(j/l)$.*

Remark 1. $ind(i, l) = 0 \iff \forall j \in E, f(j/l) \leq f(i/l), i \in E, l \in \{1, \dots, m\}$
 $(ind(i, l) \text{ reaches the maximum}) \Rightarrow \forall j \in E, f(j/l) > f(i/l)$.

Indeed, suppose that : $ind(i, l) = 0$

$$ind(i, l) = 0 \iff \sum_{j \in E, f(j/l) > f(i/l)} f(j/l) = 0$$

If $f(i/l) = 0$, then $f(j/l) = 0$ for all individuals j , because otherwise, it would exist an individual j_0 such that : $f(j_0/l) > 0 = f(i/l)$. Which contrasts the assumption $ind(i, l) = 0$ (since $ind(i, l) \geq f(j_0/l) > 0$).

$$\sum_{j \in E, f(j/l) > f(i/l)} f(j/l) = 0 \Rightarrow f(i/l) \neq 0, \forall j \in E, f(j/l) \leq f(i/l)$$

if we suppose the opposite, $ind(i, l)$ cannot take the value 0 ($f(i, l) \neq 0$).

This shows that, if the atypicality index of some cluster, affected to an individual to be classified, reaches its extreme values, the affectation (if the

index is equal to 0) or the non-affectation (if the index is maximum) of the individual to the cluster at issue, is justified. We decide to base on this index the following rule : a new observation i_s with unknown cluster membership is assigned to group G_r if $\text{ind}(i_s, r)$ is the minimum of the quantities $\text{ind}(i_s, l)$, ($l = 1, \dots, m$). The comparison of the different indexes normally leads one to prefer the cluster where the density function values are low.

2 Bayesian quadratic rule

Discriminant analysis is concerned with the construction of a statistical decision rule that allows, based on the observed data $X = (X_1, X_2, \dots, X_p)^t \in \mathbb{R}^p$, the identification of the cluster membership of X . Using information obtained from classified observations, the sample of the population P is partitioned into m clusters G_1, \dots, G_m . Given a new vector i_s to be classified, the bayesian quadratic rule (BQR) consists of searching a partition into m regions $R_1^*, R_2^*, \dots, R_m^*$, and in assigning i_s to cluster G_r , if it lies in the corresponding subset R_r^* , based on a rule that minimizes the error rate (Nakache (1980)). For a cluster G_r , let $f(\cdot/r)$, π_r and $c(j/r)$ be respectively the density function, the prior probability and the cost of misclassification resulting to an affectation of an individual from this cluster to an other cluster G_j ($(j, r) \in (\{1, 2, \dots, m\})^2$).

If we suppose that the costs of misclassification and the prior probabilities are equal for all clusters, then it follows that

$$(i_s \text{ is assigned to } G_r) \iff f(i_s/r) = \max_{l=1..m} f(i_s/l).$$

The problem leads to a comparison of the density functions in clusters. This comparison privileges the cluster for with the density function takes the highest values.

We constrain the discussion in this section to the normal unidimensional case (as will be justified in the last section). We suppose also that we have two groups induced by the dependant variable ($m = 2$). Assume that the explanatory variable is distributed as $\mathcal{N}_p(\mu_i, \sigma_i)$ in cluster G_i , $\mu_1 = 0, \mu_2 = \mu, \sigma_1 = 1, \sigma_2 = \sigma, i = 1, 2$. For the assignment of an individual, we study the ratio $R = \frac{f(x/2)}{f(x/1)}$ or the sign of $\ln(\frac{f(x/2)}{f(x/1)}) = (\sigma^2 - 1)x^2 + 2\mu x - 2\sigma^2 \ln(\sigma) - \mu^2$. The study of the variations of this quantity is summarized in Table 2.1.

X		$X_1^* = \frac{-\mu - \sqrt{(\delta)}}{\sigma^2 - 1}$		$X_2^* = \frac{-\mu + \sqrt{(\delta)}}{\sigma^2 - 1}$	
$\ln(R)$	+	0	-	0	+
R	$f(\cdot/2) > f(\cdot/1)$	$f(\cdot/2) = f(\cdot/1)$	$f(\cdot/2) < f(\cdot/1)$	$f(\cdot/2) = f(\cdot/1)$	$f(\cdot/2) > f(\cdot/1)$

Table 2.1: Study of variations of the quantity $\ln(\frac{f(x/2)}{f(x/1)})$, where $\delta = \mu^2 + (\sigma^2 - 1)(\mu^2 + 2\sigma^2 \ln(\sigma))$.

If we consider the case that the most probable observations in the first cluster are elements of the interval $[-2, 2]$, those in the second group are

elements of $[\mu - 2\sigma, \mu + 2\sigma]$. Then the risk of the cluster G_2 observations to get incorrect affectations increases with the probability of the event $([\mu - 2\sigma, \mu + 2\sigma] \cap [x_1^*, x_2^*])$, that one of G_1 increases with the probability of the event $([-2, 2] \cap [-\infty, x_1^*]) \cup ([-2, 2] \cap [x_2^*, +\infty])$.

We consider that $[\mu - 2\sigma, \mu + 2\sigma] \cap [x_1^*, x_2^*]$ is the G_2 critical region, similarly $([-2, 2] \cap [-\infty, x_1^*]) \cup ([-2, 2] \cap [x_2^*, +\infty])$ is the G_1 critical region.

3 New method of affectation

The proposed approach consists of integrating the notion of atypicality into the method based on the density function. The resulting criterion thus presents the virtues of the two methods and allows to obtain better results than those obtained, if we would apply each method separately. The new method consists of assigning i_s to the cluster G_r if and only if

$$\frac{ind(i_s, r)}{f(i_s/r)} = \min_{l=1, \dots, m} \frac{ind(i_s, l)}{f(i_s/l)}.$$

Consider the quotient $\frac{ind(i_s, l)}{f(i_s/l)}$ (i_s individual to be classified) and assume that the explanatory vector is distributed as $\mathcal{N}_p(\mu_l, \Sigma_l)$, $\mu_l \in \mathbb{R}^p$ in the cluster G_l , with density function $f(\cdot/l)$, then

$$\frac{ind(i_s, l)}{f(i_s/l)} = \frac{\sum_{f_0(i^l) > f_0(i_s^l)} f_0(i^l)}{f_0(i_s^l)}$$

where, f_0 is the density function of $\mathcal{N}_p(0, I_p)$ and i^l denotes the standardized observed data ($i^l = \Sigma_l^{-1/2}(i - \mu_l)$). The new criterion favours neither the cluster where the variables are the most dispersed nor the one, where the variables are the least dispersed, because the observations are standardized. In fact, assume that the population is partitioned into m subpopulations $P_i, i = 1, \dots, m$, we could consider that after normalization of each subpopulation, the resulting observations are those of a same centred reduced normal population, noted P_0 thus $\frac{ind(i_s, l)}{f(i_s/l)} = \frac{\sum_{i \in A_l} f_0(i)}{f_0(i_s^l)}$ where $A_l = \{i \in P_0 / f_0(i) > f_0(i_s^l)\}$. The quantity $\frac{\sum_{i \in A_l} f_0(i)}{f_0(i_s^l)}$, calculated for each cluster, depends only on $f_0(i_s^l)$, and to minimize it is equivalent to maximize $f_0(i_s^l), l = 1, \dots, m$. Therefore, in the normal case, the criterion $\frac{ind(i_s, l)}{f(i_s/l)}$ amounts to apply the BQR under the equal prior probabilities hypothesis, with normalization.

When we apply the BQR, we could solve the problem by standardizing the observations. Since in this paper we are interested in the most general case, where the density functions families are different, we could consider that the proposed rule generalizes the BQR with normalization.

The proposed criterion establishes an **equilibrium** between density functions in the different clusters without favouring neither the cluster where

the density function takes the highest values, nor the lower values, since we study for each cluster, the gaps between density function of i_s and those of the more typical individuals. The affectation of i_s is essentially bound to its inherent similarity with these individuals, as measured with the density function.

4 Comparison of classification rules

4.1 Simulation results

Assume that we have $m = 2$ populations. $X = (X_1, \dots, X_p)^t$ designates the explanatory vector. The comparison of classical rules is based on a model studied by O'Neill (1992). This model assumes that the distribution of X is $\mathcal{N}_p(\omega_i, \Omega_i)$ ($i = 1, 2$) in the two populations, where $\omega_1 = 0$, $\Omega_1 = I_p$ in population P_1 , $\omega_2 = \delta e_1$, $\Omega_2 = I_p + \tau e_1 e_1^t$ in population P_2 , e_1 is the first unit vector in \mathbb{R}^p , $\tau > -1$ and $\delta \in \mathbb{R}$ are parameters. Thus the two populations differ only on the first variable, this justifies the study in section 2. The O'Neill model has been generalized for the detection of shift in means and variances in several directions (Flury, 1996):

Definition 3. Let $X \hookrightarrow \mathcal{N}_p(\omega_i, \Omega_i)$ in P_i ($i = 1, 2$). Then X satisfies a q -dimensional O'Neill model ($q < p$), if $\omega_1 = 0$, $\Omega_1 = I_p$, $\omega_2 = \sum_{j=1}^q \delta_j e_j$, and $\Omega_2 = I_p + \sum_{j=1}^q \tau_j e_j e_j^t$.

Thus the two populations differ in the first q coordinate directions, whereas no change has occurred in the remaining $p - q$ directions.

To assess the performance of discrimination based on the ratio (atypicality index/density function), we ran a simulation study with: $\pi_1 = \pi_2$, $\mu_1 = 0 \in \mathbb{R}^p$, $\psi_1 = I_p$, $\mu_2 = \delta e_1 \in \mathbb{R}^p$, and $\psi_2 = \psi_1 + \tau e_1 e_1^t$, $p = 4$, $\delta \in \{3, 4, 5, 6\}$, $\tau \in \{3, 8, 15, 24, 35\}$. For six cases resulting from some parameters combination, $k = 1000$ training samples of size $n_1 = n_2$ were generated from the $\mathcal{N}_p(\mu_i, \psi_i)$ distributions, $i = 1, 2$. The sample sizes n_1 and n_2 were always chosen to be equal. We present here just two cases, because they exhibit those where the critical regions are probable and also they show that in the similar cases, the proposed method performs better than all classified rules. For each pair of samples, the expected well classified percentages were computed for the following methods of discrimination : the bayesian quadratic (BQR), the linear (LDA) and the method based on the ratio (atypicality index/density function) (RAD). The well classified percentage was approximated as the average of the k well classified percentages calculated for each simulation by the resubstitution method. For the six studied simulations, and particularly for the two selected simulations presented here, the proposed method performs much better than all rules. The lower well classified percentages for the (BQR) are reached because the critical regions probabilities are important (Table 4.2).

Case	Probability of the G_2 critical region	Probability of the G_1 critical region
$\delta = 3 \tau = 8$	0.277	0.0366
$\delta = 4 \tau = 15$	0.2306	0.02

Table 4.2: Critical regions probabilities.

There is a clear separation between the well classified percentages calculated for the RAD method, and those calculated for the others, as seen in Figure 1. For $(\delta = 3, \tau = 8)$ (Figure 1a), note that the linear rule be-

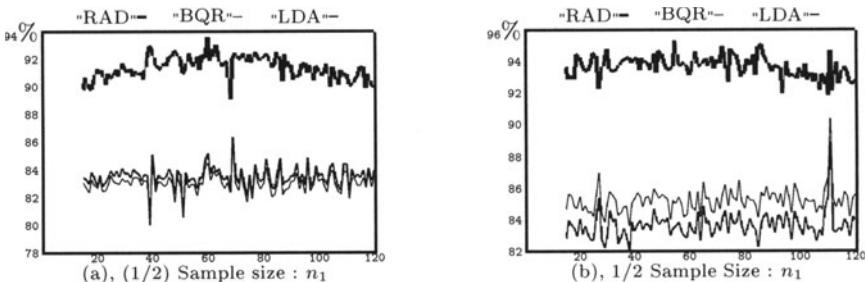


Fig. 1. Average of the k well classified %, $p = 4$, (a) $\delta = 3, \tau = 8$; (b) $\delta = 4, \tau = 15$

haves similarly to the quadratic, the linear performing slightly better than the quadratic. The separation between the well classified percentages calculated for the RAD method and those calculated for all others is better expressed in Figure 1b. Now the quadratic rule performs slightly better than the linear one.

In summary, the separation between the well classified percentages calculated for the RAD, and those calculated for the BQR, increases with the critical region probabilities.

We further illustrate this advantage with a numerical example. We are particularly interested in showing that, in the multinormal case, the proposed rule results are identical to those of the BQR under the equal prior probabilities hypothesis, with normalization.

4.2 Numerical example

For the example named **Normal4**, the equal covariance matrices hypothesis is realized, so the application of the BQR on the initial data is equivalent to its application on the standardized data. **Normal4** is a sample of $n = 800$ points consisting of 200 points each from the four components of a mixture of $p = 4$ -variate normals. The population mean vector and covariance matrix for each component of the normal mixture were

$\mu_i = 3e_i, \Sigma_i = I_4, i = 1, 2, 3, 4, e_i = (0, \dots, \underbrace{1}_{\text{---}}, \dots, 0)$. It is helpful to picture

the geometric structure of **Normal4**, which is (a sample of) 200 points each centered at three units from the origin of the four directions. By applying the resubstitution method, the well classified percentage calculated for the RAD method is equal to 96.125%, that calculated for the quadratic rule is equal to 96%. The two percentages are almost identical.

5 Conclusion

As the simulation and example shows, the success of our method is rather convincing. When the critical regions probabilities are important, there is a clear separation between the well classified percentages calculated for the RAD method and those calculated for the classical methods. In this article, we assume that the prior probabilities are equal, because, generally, they are estimated by the frequencies, but in most cases, the estimator of the prior probability is equal to $1/m$ (m is the number of Q modalities) (Nakache (1980)). The RAD method establishes an **equilibrium** between density functions in different clusters. In the simulation, we have concentrated on the O'Neill model (case $q = 1$), which we believe to be the most important case for practical applications. This is motivated both by practical experience and by the inherent similarity with linear discrimination.

References

- FLURY B., NEL D. G. and PIENAAR I. : Simultaneous Detection of Shift in Means and Variances : *Journal of the American Statistical Association* 91, pp. 1743–1748 (1996).
- NAKACHE J. P. : Méthodes de discrimination pour variables de nature quelconque, théorie et pratique : Thèse présentée pour obtenir le grade de Docteur-es sciences, Université Pierre et Marie Curie (1980).
- O'NEILL T. J. : Error Rates of Non-Bayesian Classification Rules and Robustness of Fisher's Linear Discriminant Function : *Biometrika*, 79, pp. 177–184 (1992).
- RIOUX P. : Quadratic discriminant analysis : *EDV in Medizin und Biologie*, 6, 112 (1975).
- ROHLFF J. : Adaptative Hierarchical clustering shemes : *Syst. Zool.*, 19, p. 58–82 (1970).
- SEBESTYEN G. S. : Decision Making Process in Pattern Recognition : Vol 19, *Mac Millan Company*.

A Third Stage in Regression Tree Growing: Searching for Statistical Reliability

Carmela Cappelli¹, Francesco Mola², and Roberta Siciliano¹

¹ Department of Mathematics and Statistics, Naples University Federico II
Monte S. Angelo, via Cinthia, 80126 Napoli, Italy
(e-mails: cappelli@dms.unina.it; roberta@unina.it)

² Department of Economics, University of Cagliari
Viale Fraignazio, 09127 Cagliari, Italy (e-mail: mola@unica.it)

Abstract. ¹ This paper suggests the introduction of a third stage in regression tree growing approach. To this aim, a statistical testing procedure based on the F statistics, is proposed. In particular, the testing procedure is applied to the CART sequence of pruned subtrees, resulting in a single final tree structured prediction rule, which is statistically reliable and might not coincide with any tree in the sequence itself.

1 Introduction

Recursive partitioning procedures, or tree based methods are a useful non parametric tool to generate a prediction rule in the form of a binary tree. Regression trees induction can be summarised as follows: let (Y, \mathbf{X}) be a multivariate random variable where (\mathbf{X}) is the vector of K predictors and Y is the numerical response variable. The so called totally expanded tree is obtained by recursively splitting a learning sample $S = \{(y_n \mathbf{x}_n), n = 1, \dots, N\}$ of N cases, taken from the distribution of (Y, \mathbf{X}) , into two subsets (the nodes of the tree) according to a splitting criterion which allows to select at each node the best predictor and cut point to split the node. Actually, the totally expanded tree is in general very large and complex and, above all, it fits the learning data in the sense that it appears very accurate respect to the learning cases but it is likely to perform poorly when it will be used to assess the response value of new cases. This is a typical inferential task which requires some tree processing, to give statistical reliability. Removing some of the branches, by means of the so called *pruning*, reduces the size of the tree resulting in improving its understandability as well as its accuracy (on a separate test set), but regardless of statistical reliability. In other words, a tree based procedure should ideally consist of three stages: 1) creating the totally expanded tree; 2) pruning the tree to improve understandability as well as accuracy; 3) validating the tree from the point of view of reliability.

As a matter of fact the third stage is quite neglected in the literature, in the sense that usually it is seen as the concluding step of the pruning

¹ Research supported by MURST funds 1999 (prot. 9913182289)

process, aimed simply to minimize some estimated error measure. In this paper, in the framework of the well known CART methodology (Breiman et al, 1984), a third stage, based on a statistical testing procedure, in regression tree construction is proposed. In order to show how the procedure works, an example on a real data set will be also presented.

2 CART regression tree pruning

CART pruning approach is based on the so called *error-complexity measure* which takes into account either the aspect of accuracy (measured by the same error measure used in the growing phase) or that of complexity (given by the number of terminal nodes of a tree). The error-complexity measure is defined for any node t and its branch T_t as:

$$R_\alpha(t) = R(t) + \alpha \quad (1)$$

$$R_\alpha(T_t) = R(T_t) + \alpha |\bar{T}_t| \quad (2)$$

where $R(T_t) = \sum_{h \in H_t} R(h)$, H_t is the set of terminal nodes h of the branch T_t whose cardinality is $|\bar{T}_t|$ (i.e., the number of leaves of the branch) and α is a sort of penalty for complexity. The idea is to prune the branch T_t if its error-complexity measure is not lower than the error-complexity measure of its root t ; actually, in this case it is useless to retain the subtree because it does not decrease the error while increasing the complexity. In particular, the two error measures become equal for a critical value of α given by:

$$\alpha_t = \frac{R(t) - R(T_t)}{|\bar{T}_t| - 1}, \quad (3)$$

in other words, α , named *complexity parameter*, represents the reduction in error per terminal node. The method is in two phases: first a sequence of nested pruned subtrees, $T_{max} \supset T_{(1)} \supset \dots \supset T_{(k)} \supset \dots \supset \{t_1\}$, is created cutting at each step the subtree branching from the node with the minimum value of α (node named *weakest link*), then a single tree, i.e., a final prediction rule, is selected on the basis of the accuracy evaluated on a separate test set. The main criticism to this method as said before, concerns this final selection strategy, which takes into account only the accuracy aspect regardless of the statistical reliability of the trees in the created sequence. In regression, the sequence of pruned trees tends to be larger than in classification because usually only two terminal nodes are cut off at a time. As a consequence, plotting on a graph the test sample estimates of the error measure against the number of terminal nodes of the trees in the sequence the resulting curve appears very flat and wide so that the selection of the particular tree which produces the smallest error measure estimate is somewhat arbitrary.

3 A statistical testing procedure as a third stage in regression tree growing

The error measure used in the pruning stage is given for any node t by the sum of squares divided by a constant factor N . This measure coincides with the impurity measure used in the growing stage to split the nodes, where that split is chosen that maximises the decrease in impurity. Siciliano and Mola (1996) have proved that this decrease (multiplied by the constant factor N), denoted by $\Delta R(s^*, t)$, can be viewed for any node t as the between groups sum of squares:

$$N \Delta R(s^*, t) = TSS_Y(t) - WSS_{Y|s}(t) = BSS_{Y|s}(t), \quad (4)$$

where $TSS_Y(t)$ denotes the total sum of squares at node t , $WSS_{Y|s}(t)$ the within groups (i.e. nodes) sum of squares and $BSS_{Y|s}(t)$ the between groups sum of squares (induced by splitting the node into its left and right descendants).

Since Cappelli et al. (1998) have shown that the complexity parameter can be expressed, for any internal node t , as the average of the decrease in impurity arising from splitting the node and its non terminal descendants, it follows that in regression (3) can be written:

$$\alpha_t = \frac{1}{|\bar{L}_t|N} \sum_{l \in L_t} BSS_{Y|s^*}(l). \quad (5)$$

As a consequence, each complexity parameter, i.e. each pruning operation, can be validated by applying the analysis of variance testing procedure which compares for any node t and its branch, the variance between the groups resulting from splitting the node (and eventually its non terminal descendants) with the variance within the nodes:

$$F = \frac{\sum_{l \in L_t} BSS_{Y|s}(l)}{WSS_{Y|s}(t)} \times \frac{N(t) - |\bar{T}_t|}{|\bar{L}_t|} \quad (6)$$

which has, under the null hypothesis, a Snedecor-Fisher distribution with $|\bar{L}_t|$ and $N(t) - |\bar{T}_t|$ degrees of freedom (note that it holds $|\bar{T}_t| = |\bar{L}_t| + 1$, being, in binary trees, the number of terminal nodes equal to the number of internal nodes plus one). This result means that when datasets are large (so that the underlying assumption of multinormality maybe assumed to be satisfied), by fixing a significance level it is possible to verify at each step, whether the branch to be pruned produces a significant increase in the variance between (it should be kept) or not (it should be removed). Moreover, since the effect of a split (i.e. predictor and cut point) depends on the prior splits selected in the tree, to avoid the dependence problem, the value to be tested are computed considering the cases of a separate test set, opposed to the pruning stage which is based on the same learning set employed to induce the tree.

4 How the procedure works

In order to show how the proposed procedure works, we have applied it to a real data set. This consists of 455 cases randomly sampled from a larger data set concerning a Bank of Italy survey on family budgets. The response variable is the mean monthly number of payments by credit card; the predictors and their modalities are: $X_1 = \text{age}$ (numerical); $X_2 = \text{sex}$ (1=male, 2=female); $X_3 = \text{education}$ (1=elementary, 2=middle school, 3=high school, 4= degree); $X_4 = \text{geographic area}$ (1=N-W, 2=N-E, 3=Center, 4=S and Isles); $X_5 = \text{income per year}$ (1=< 35, 2=35 – 50, 3=> 50); $X_6 = \text{house ownership}$ (1=yes, 2=not); $X_7 = \text{monthly payment by cash}$ (numerical); $x_8 = \text{importance attached to payments delay}$ (1= none, 2= a fair deal, 3= a lot, 4= a great deal); $X_9 = \text{propensity to consume}$ (numerical). The totally expanded tree, grown using 70% of cases (the remaining has been employed as test set), with 43 terminal nodes, has been pruned according to the CART pruning procedure, generating a sequence of pruned subtrees which is described in Table 1. For each subtree are reported: the number of terminal nodes, the value of the smallest complexity parameter α , the corresponding weakest linked node with the associated empirical F value (in bold are those significant at 0.01), and the test sample error measure estimate. The starred trees in the first column correspond to the CART best choices i.e., the tree with lowest error rate estimate (so called $0 - SE$ rule) and the tree with error rate within one standard error of the minimum (so called $1 - SE$ rule), respectively. As expected the sequence is large and the error rate estimates so close to each other to make CART selection rules unsuitable. A further stage is required to identify the final tree which is statistically reliable according to the proposed F testing procedure. This testing follows the path of subtrees produced by the CART pruning process either cutting off or retaining those weakest links whose branches are significant. In other words, the testing takes at each step account of the previous outcomes and since these sometimes disagree with the CART choices, *the resulting tree does not necessarily coincide with any subtree in the sequence*. Figure 1 shows the final tree with 14 terminal nodes which is moreover, characterized by a test sample error measure estimate $\hat{R}(T)$ equal to 8.32, lower than any other belonging to the sequence.

5 Conclusions

In this paper a third stage in regression tree construction, based on the employment of a statistical testing procedure, has been proposed. Actually, regression trees are quite neglected by literature where the attention focuses on classification trees. In this framework, statistical testing has been proposed by Mingers (1989) and Zhang (1999). Here, the F statistic has been employed to validate retrospectively the CART pruning process, i.e., to assess whether to cut or not *any depth branches*, using the test set to tackle the problem of

dependence. In other words a distinction has been made between the mere simplification of a tree and the assessment of its statistical reliability by validating *a posteriori*, the values of the error measure used to grow the tree.

Subtrees	Number of terminal nodes	α_{min}	weakest linked node	F_e	$\hat{R}(T)$
T_1	43	0.0009	597	25.7333	9.5749
T_2	42	0.0011	2807	2.7499	9.5740
T_3^*	41	0.0023	23	53.5423	9.5502
T_4	39	0.0033	298	6.0122	9.5700
T_5	38	0.0035	1403	64.7907	9.5666
T_6	37	0.0038	163	3.2374	9.5693
T_7	36	0.0039	158	0.0000	9.5720
T_8	35	0.0050	149	0.0000	9.5775
T_9	34	0.0072	701	4.0119	9.5748
T_{10}	33	0.0078	81	7.3546	9.5857
T_{11}	32	0.0132	11	18.3860	9.5911
T_{12}	31	0.0146	40	9.3080	9.6170
T_{13}	30	0.0160	74	9.9438	9.6307
T_{14}	29	0.0191	79	0.0000	9.6539
		0.0191	346	1.3822	9.6539
T_{15}	27	0.0199	37	4.6612	9.6661
T_{16}	26	0.0232	350	1.1205	9.6621
T_{17}	25	0.0240	20	7.4223	9.6498
T_{18}	24	0.0293	351	8.1342	9.6702
T_{19}	23	0.0314	173	2.1952	9.7234
T_{20}	22	0.0453	87	0.1549	9.7480
T_{21}	20	0.0775	312	0.0000	9.6716
T_{22}	19	0.0931	33	191.2516	9.6839
T_{23}	18	0.1196	156	24.0626	9.8285
T_{24}	17	0.1284	78	4.4835	9.9731
T_{25}	16	0.1396	43	3.2148	10.0754
		0.1396	18	14.9300	10.0754
T_{26}	12	0.1503	5	3.4256	10.3427
T_{27}	9	0.2029	39	1.7169	10.3973
T_{28}	8	0.5173	9	9.6669	10.2100
T_{29}	6	0.5334	16	72.8133	11.4409
T_{30}^*	5	1.0769	1	7.5250	11.4149

Table 1. Sequence of pruned subtrees with the values of the F statistic associated to the corresponding weakest link

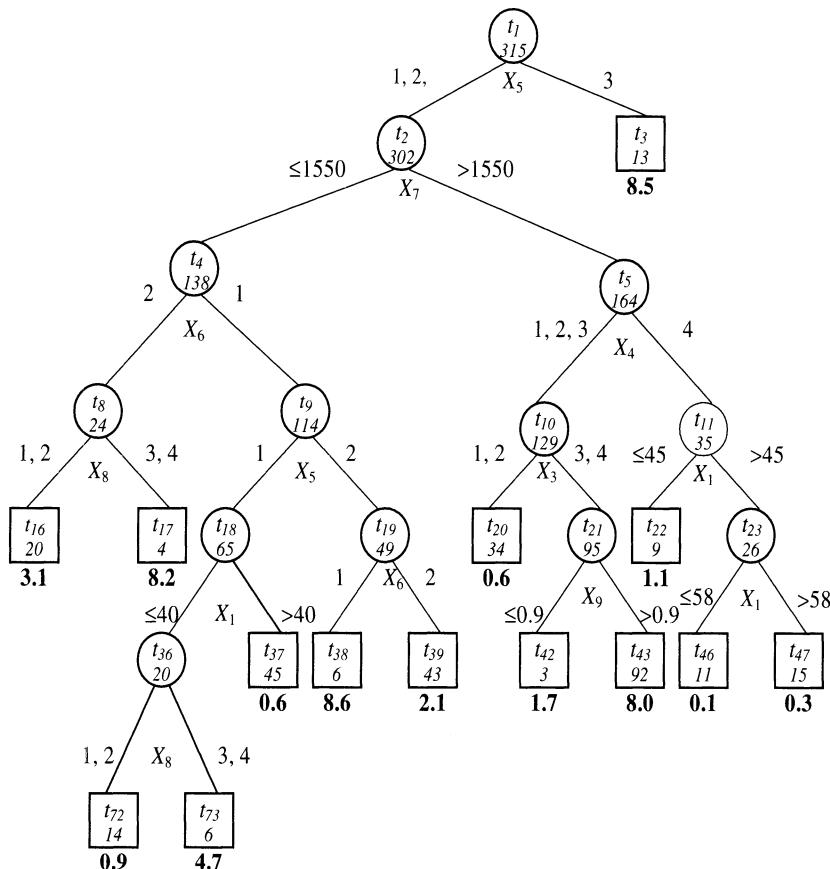


Fig. 1. The final tree resulting from the statistical testing procedure

References

- BREIMAN L., FRIEDMAN J. H., OLSHEN R. A. and STONE C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont CA.
- CAPPELLI, C., MOLA, F. and SICILIANO, R. (1998): An Alternative Pruning Method Based on the Impurity-Complexity Measure. In: R. Payne and P. Green (Eds): *Proceedings in Computational Statistics*, Physica-Verlag, Heidelberg, 221–226.
- MINGERS, J. (1987): Expert Systems- Rule Induction with Statistical data. *Journal of the Operational Research Society*, 38, 39–47.
- SICILIANO, R. and MOLA, F (1996): A Fast Regression Tree Procedure. In: A. Forcina et al. (Eds): *Proceedings of the 11th International Workshop on Statistical Modelling*, Graphos, Perugia, 332–340.
- ZHANG, H. and SINGER, B. (1999). *Recursive partitioning in the health science*, Springer.

A New Sampling Strategy for Building Decision Trees from Large Databases

J.H. Chauchat and R. Rakotomalala

Université Lumière Lyon 2
5 avenue Pierre Mendès France, C.P.11 69676 Bron Cedex, France
e-mail : chauchat.rakotoma@univ-lyon2.fr

Abstract. We propose a fast and efficient sampling strategy to build decision trees from a very large database, even when there are many numerical attributes which must be discretized at each step. Successive samples are used, one on each tree node. Applying the method to a simulated database (virtually infinite size) confirms that when the database is large and contains many numerical attributes, our strategy of fast sampling on each node (with sample size about $n = 300$ or 500) speeds up the mining process while maintaining the accuracy of the classifier.

1 Introduction

In this paper we propose a fast and efficient sampling strategy to build decision trees from a very large database, even when there are many numerical attributes which must be discretized at each step.

Decision trees, and more generally speaking decision graphs, are efficient and simple methods for supervised learning. Their "step by step" characteristic allows us to propose a strategy using successive samples, one on each tree node. In that way, one of the most limiting aspects of the decision tree method is overcome (analyzed data set reduction as the algorithm goes forward, successively dividing the set of training cases).

Working on samples is especially useful in order to analyze very large databases, in particular when these include a number of numerical attributes which must be discretized at each step. Since each discretization requires to sort the data set, this is very time consuming. Section 2 outlines the general decision tree method, the numerical attributes discretization problem and our new sampling strategy at each step.

In section 3, we apply the whole method to a simulated database (virtually infinite size). The results confirm that when the database is large and contains many numerical attributes, our strategy of fast sampling on each node (with sample size about $n = 300$ or 500) reduces drastically learning time while maintaining the accuracy in generalization.

2 Decision trees and induction graphs

2.1 Induction with graphs

Decision trees (Breiman et al, 1984), and more generally speaking decision graphs (Zighed and Rakotomalala, 2000), are efficient, step by step, and simple methods for supervised classification. *Supervised* classification means that a classification pre-exists and is known for each record in the (training) database we are working on: the patient has been cured, or not; the client has accepted a certain offer, or not; the machine breaks down, or not. Those situations have two values; sometimes there are three or more. The final objective is to learn how to assign a new record to its true class, knowing the available attributes (age, sex, examination results, etc.).

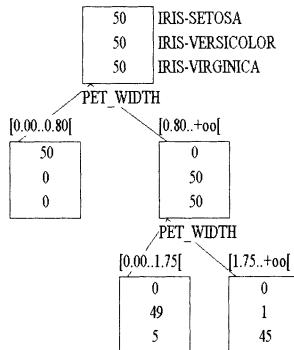


Fig. 1. A decision tree built on Fischer's Iris dataset. Iris classification learning, using petale, and sepale, length and width.

The wide utilization of decision tree methods is based on its simplicity and ease of use. One is looking for a dataset partition represented by a lattice graph (Figure 1). This partition must minimize a certain criterion. Generic algorithms (Breiman et al., 1984) (Quinlan, 1986) make local optimization. In spite of their simplicity, decision trees have a very good predictive power compared with more complex method such as Neural Network (Quinlan, 1993). Nowadays, as Knowledge Discovery in Databases (KDD) is growing fast (Fayyad et al., 1996), one can note a growing number of studies on decision trees and induction graphs, as well as broad software diffusion.

2.2 Using continuous attributes in decision trees

Most training-by-examples symbolic induction methods (Cohen, 1995) have been designed for categorical attributes, with finite value sets. For instance

"sex" has two values: male or female. However, when we want to use continuous attributes (income, age, blood pressure, etc.), we must divide the value set in intervals so as to convert the continuous variable into a discrete one. This process is named "discretization". The importance of this research area has recently become apparent.

Ever-growing data, due to the extensive use of computers, the ease of data collection with them and the advance in computer technology, drive dataminingers into handling databases comprising varied type and non pre-processed attributes.

The first methods for discretization were relatively simple, and few papers have been published to evaluate their effects on machine learning results. From the beginning of the '90s much theoretical research has been done on this issue. The general problem has been clearly formulated (Lechevallier, 1990) and several discretization methods are now in use (Zighed et al., 1998). Initial algorithms processed discretization during the pre-processing stage: each continuous attribute was converted to a discrete one; after which, a regular symbolic learning method was used.

Within the particular framework of the decision graphs, it is possible to simplify the discretization of a continuous attribute by carrying out a binary local cutting. The process is as follows: on each node of the tree, each continuous variable is first of all sorted, then all the possible cutting points are tested so as to find the binary cut which optimizes a criterion such as information gain or mutual information measure (Shannon and Weaver, 1949).

This strategy thus makes it possible to compare the predictive capacity of all the attributes, whether continuous or not. In spite of this simplicity, we are facing here one of the principal bottlenecks in the development of graphs. Cross tabulation, on which the criteria of the partitions are calculated, is a relatively inexpensive phase: it is of $O(n)$. On the other hand the processing of the continuous variables requires initially a sorting of the values in ascending order which, in the best case, is of $O(n \log n)$. This is why we propose to use sampling by reducing the number n of records to which these calculations apply.

2.3 Using successive samples, one on each tree node

To each node of the graph corresponds a subpopulation; it can be described by the conjunction of the attribute values located on the outgoing path from the root to the considered node. Thus, to split a node, the following general framework reduces dramatically the computing time :

1. draw a sample from the subpopulation corresponding to the node (see Figure 2);
2. use this sample to discretize continuous attributes and to determine the best splitting attribute.

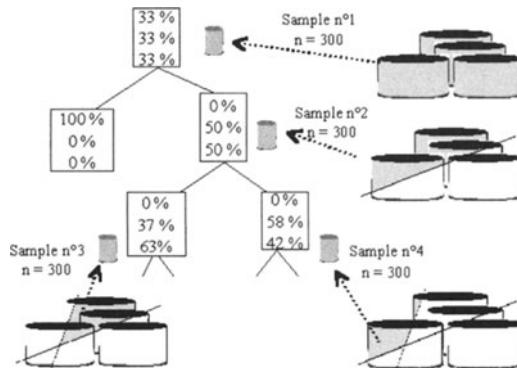


Fig. 2. Sampling and re-sampling on each node to build a decision tree.

Sampling should save time during data processing, but the sampling operation itself should not be time consuming. One can use very fast sampling methods (Vitter, 1987).

In order to determine the sample size, elements of statistical theory (using statistical test power function and non central chi-squared distribution) are presented in (Chauchat and Rakotomalala, 1999). Samples of a few hundred are usually enough to determine interesting predictive attributes.

3 Implementation on simulated databases

We will now apply the whole method (with sampling and binary discretization on each node of the tree) to a well known artificial problem: the "Breiman's waves" (Breiman et al., 1984). In § 2.6 of his book, Breiman poses this problem, now traditional: each of three classes is characterized by a specific weighting combination of 21 pseudo-random standardized normal variables.

We generated 100 times two files, one of 500,000 records for the training, the other of 50,000 records for the validation. Binary discretization was pre-processed on each node for each attribute. Sample size drawn from the file on each node varies from $n = 100$ to $n = 500$. ID3 method has been used because it is fast; it uses pre-pruning with the χ^2 -test.

The learning time is quasi null for $n = 100$ because the tree stops very quickly, even immediately: the pruning χ^2 -test has a low power (if n is too small, the *observed- χ^2* is small too, even if an attribute is useful). From $n = 200$, the run time increases a little quicker than linearly with n , in accordance with theory ($n \log(n)$).

Figure 3 shows how the error in generalization decreases as the sample size n increases. Even for this problem considered as a difficult one, the marginal profit becomes weak starting from sample sizes of $n = 300$ records; one approaches then 19%, the minimum of error in generalization obtained with trees using the entire database, (with its $N = 500,000$ records).

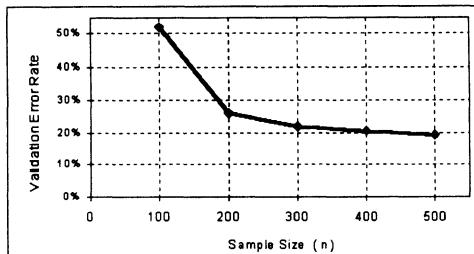


Fig. 3. Average ERROR RATE according to the sample size drawn on each node (Breiman's Waves Dataset; Optimal discretization of 21 continuous attributes at each step).

These results have been confirmed by several empirical studies on real databases. From a sample of approximately 300 records on each node, we obtain trees error rate close to that obtained using the whole database.

4 Conclusions

Decision trees, and more generally speaking decision graphs, are efficient, step by step, and simple methods for supervised classification. However, mining on very large databases, in particular when these include a number of numerical attributes which must be discretized at each step, is very time consuming. In these cases, working on samples is especially useful. The decision tree "step by step" characteristic allows us to propose a strategy using successive samples, one on each tree node. Empirical evidences show that our strategy of fast sampling on each node (with samples size about $n = 300$ or 500) reduces considerably learning time while preserving the accuracy. Sampling in data mining is not a new approach. (Toivonen, 1996) for instance is looking for the minimum sample size to get all the useful association rules. Our goal is different : we are not looking for the same decision tree as that built on the whole databases, what is impossible using a sample, but one which have roughly the same error rate.

This work raises some open questions. Optimal sampling methods (stratified random sampling, selection with unequal probabilities, etc.) may be used. However, those methods were developed for surveys on economic or sociological fields, when the cost of the information collected is high compared to calculation time. They must be adapted for data mining: in our situation the information is already known, it is in the database. We have to check if the gain in accuracy obtained by these methods, the sample size being fixed, may not be supplied by sample size enlargement, for the same learning time. An interesting way may be the balanced sampling strategy on each node (Chauchat et al., 1998).

An other question is the implementation of sampling in the core of the queries in databases.

References

- BREIMAN, L., FRIEDMAN, J., OLSHEN, R., and STONE, C. (1984) *Classification and Regression Trees*. California : Wadsworth International.
- CHAUCHAT, J., BOUSSAID, O., and AMOURA, L. (1998) Optimization sampling in a large database for induction trees. In *Proceedings of the JCIS'98-Association for Intelligent Machinery*, 28–31.
- CHAUCHAT, J., and RAKOTOMALALA, R. (1999) Détermination statistique de la taille d'échantillon dans la construction des graphes d'induction. In *Actes des 7^{èmes} journées de la Société Francophone de Classification*, 93–99.
- COHEN,W. (1995) Fast effective rule induction. In *Proc. 12th International Conference on Machine Learning*, 115–123. Morgan Kaufmann.
- FAYYAD, U., PIATETSKY-SHAPIRO, G., and SMYTH, P. (1996) Knowledge discovey and data mining : Towards an unifying framework. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*.
- LECHEVALLIER, Y. (1990) Recherche d'une partition optimale sous contrainte d'ordre totale. Technical Report 1247, INRIA.
- QUINLAN, J. (1986) Induction of decision trees. *Machine Learning* 1:81–106.
- QUINLAN, J. (1993) Comparing connectionist and symbolic learning methods. In Hanson, S.; Drastal, G.; and Rivest, R., eds., *Computational Learning Theory and Natural Learning Systems : Constraints and Prospects*. MIT Press.
- SHANNON, C. E., and WEAVER, W. (1949) *The mathematical theory of communication*. University of Illinois Press.
- TOIVONEN, H. (1996) Sampling large databases for association rules. In *Proceedings of 2nd VLDB Conference*, 134–145.
- VITTER, J. (1987) An efficient algorithm for sequential random sampling. *ACM Transactions on Mathematical Software* 13(1):58–67.
- ZIGHED, D. and RAKOTOMALALA, R. (2000) *Graphes d'Induction : Apprentissage et Data Mining*. Hermes.
- ZIGHED, D., RABASEDA, S., and RAKOTOMALALA, R. (1998) Fusinter : a method for discretization of continuous attributes for supervised learning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 6(33):307–326.

Generalized Additive Multi-Model for Classification and Prediction

Claudio Conversano¹, Francesco Mola² and Roberta Siciliano¹

¹ Dipartimento di Matematica e Statistica, Universita di Napoli Federico II, I-80126, Via Cintia, Napoli, Italia, roberta@unina.it, conversan@dms.unina.it

² Dipartimento di Economia, Universita di Cagliari, I-09100, Viale Fra Ignazio, Cagliari, Italia, mola@unica.it

Abstract. ¹ In this paper we introduce a methodology based on a combination of classification/prediction procedures derived from different approaches. In particular, starting from a general definition of a classification/prediction model named Generalized Additive Multi-Model (GAM-M) we will demonstrate how it is possible to obtain different types of statistical models based on parametric, semiparametric and nonparametric methods. In our methodology the estimation procedure is based on a variant of the backfitting algorithm used for Generalized Additive Models (GAM). The benchmarking of our methodology will be shown and the results will be compared with those derived from the applications of GAM and Tree procedures.

1 Introduction

In recent years there has been a great interest of the part of the statistical community in the use of combinations of models to improve the accuracy of the estimation procedures. In this paper, we introduce the Generalized Additive Multi-Model (GAM-M) which consists of a combination of parametric, nonparametric and semiparametric models. The reason of this growing interest derives from the need to minimize the uncertainty and causality elements characterizing statistical models and, at the same time, to encourage the use of statistical models in modern applied data analysis. However an incorrect methodological choice seriously compromises the final result of the estimation (even if the chosen method is applied in a rigorous manner), as it derives from inappropriate initial hypotheses concerning the data. The proposed approach might help to prevent from an incorrect choice of the model and is particularly suitable for complex data sets.

¹ Authors wish to thank referees and Prof. Jaromir Antoch for helpful remarks on a previous draft of the paper. Research was supported by MURST funds 1999 (prot. 9913182289).

2 A general definition of combined model integration

In this section we first deal with semiparametric models and then we propose a more general class of models. Semiparametric models consist of a parametric and a nonparametric component (besides the error term). The response variable depends on some covariates in a parametric (e.g. linear) fashion and on an additional (or several) covariate(s), not full-filling a parametric assumption(s). For the unparametrized relationships smooth functions belonging to the family of linear scatterplot smoothers of the spline or kernel type are assumed. A link function evaluates the effect of the linear combination of the predictors on the dependent variable, in the case the latter derives from a probability distribution. An example of a semiparametric model is Generalized Additive Model (GAM) of Hastie and Tibshirani (1990). This model can be defined as follows:

$$E[Y|\mathbf{X}] = G \left(\alpha + \sum_{j=1}^d \beta_j f_j(X_j) \right) \quad (1)$$

where G is a fixed link function, $\mathbf{X} = (X_1, \dots, X_d)$ is the vector of predictors, the parameters $\beta = (\beta_1, \dots, \beta_d)$ and Y is the dependent variable (which distribution is assumed to belong to the exponential family). The effect of each predictor on the dependent variable is evaluated through a smoothing function f_j . The backfitting algorithm allows to estimate the model considering each predictor separately by working recursively on partial residuals.

This type of estimation procedure is the starting point for our definition of a general formulation of *combined model integration*. The idea is to combine not only and not necessarily estimations derived from smoothing functions, one for each predictor, but estimations provided by either parametric or nonparametric or semiparametric models, *one for each predictor*. A somewhat similar idea has been recently worked out by Mertens and Hand (1999), but they have combined estimations of models fitted to *the whole set of predictors*. In the following, we introduce the *Generalized Multi-Model (GAM-M)* which will be estimated by means of a variant of backfitting algorithm. Let $(M_1, \dots, M_i, \dots, M_K)$ be a set of models of parametric, semiparametric or nonparametric type, where, without loss of generality, $M_1 = 1$ denotes the unspecified option. The effect of the j -th predictor on the response variable Y can be evaluated *either* through a smoothing function f_j (thus $M_1 = 1$) *or* any of the specified models. For that we introduce the following generalization of (1):

$$E[Y|\mathbf{X} = x] = G \left(\alpha + \sum_{j=1}^d \sum_{i=1}^K \delta_{ij} \beta_{ij} f_j(X_j|M_i) \right) \quad (2)$$

where β_{ij} are the parameters, δ_{ij} is a dummy variable such that $\sum_i \delta_{ij} = 1$, i.e., only one option between smoother (when $\delta_{1j} = 1$) or model (when $\delta_{ij} = 1$

for $i \neq 1$) is assigned to each predictor. In the latter case the smoothing function f_j reduces to the identity function. By definition there cannot be more than d terms into the additive part of the model (2) so that $\sum_j \delta_{ij} \leq d$. (i.e., a model can be assigned between zero and d times). In the particular case that $K = 1$ model (2) is equivalent to (1), namely we obtain simply an additive combination of smoothing functions like in GAM. For $K > 1$ we define more general models. At the same time, starting from (2) it is possible to obtain as special cases different types of models depending on the type of smoothing function f_j or statistical model M_i associated to each predictor as well as the specification of the link function G . Some possible outcomes are briefly summarized in what follows:

- For $K = 1$, $f_j \equiv \text{Identity}$ and $G \equiv \text{Identity}$:

$$E[Y|\mathbf{X}] = \alpha + \mathbf{X}\beta \quad \text{classical linear regression.}$$

If G is the exponential function and the response is binary, we obtain the logistic regression model.

- For $K = 1$, $\{f_j\} = m(\dots)$ where a predictor combines more predictors (multivariate smoothing function), $\beta_j = 1$ for each j and $G \equiv \text{Identity}$:

$$E[Y|\mathbf{X}] = m(\mathbf{X}) \quad \text{nonparametric regression.}$$

- For $K = 1$, $\{f_j\} = a_l B_l(X_j)$ (Basis functions), $\beta_j = 1$ for each j and $G \equiv \text{Identity}$:

$$E[Y|\mathbf{X}] = \sum_{j=1}^d \sum_{l=1}^L a_l B_l(X_j) \quad \text{CART (Breiman et al. (1984))}$$

Here L is the number of bases corresponding to the number of terminal nodes of the tree. The smoothing functions correspond to the basis functions defined on the hyper-rectangles which are derived from the tree fitting algorithm (Friedman, 1991). An alternative and computationally more efficient splitting criterion for tree based models has been introduced by Mola and Siciliano (1997).

- For $K = 1$, $\{f_j\} \equiv$ smoothing functions, $\beta_j = 1$ for each j and $G \equiv \text{Identity}$:

$$E[Y|\mathbf{X}] = \alpha + \sum_{j=1}^d \beta_j f_j(X_j) \quad \text{additive model.}$$

- For $K = 1$, $\{f_j\} \equiv$ smoothing functions, $\beta_j = 1$ for each j and $G \equiv \text{Exponential}$:

$$E[Y|\mathbf{X}] = G \left(\alpha + \sum_{j=1}^d \beta_j f_j(X_j) \right) \quad \text{generalized additive model.}$$

- For $K = 1$ there exists only one model type, either parametric or semi-parametric, $\{f_j\} = a_l B_l(X_j)$ (Basis functions), $\beta_j = 1$ for each j and $G \equiv$ Identity:

$$E[Y|\mathbf{X}] = \alpha + \sum_{j=1}^d \sum_{l=1}^L a_l B_l(X_j|M).$$

Here M is a semiparametric model. We obtain a procedure which also includes the model based recursive partitioning procedure of Siciliano and Mola (1994), resulting in a non- and semiparametric integration.

- For $K = 1$ there exists only one model: $M \equiv$ Regression Tree, $\{f_j\} \equiv$ smoothing functions, $\beta_j = 1$ for each j and $G \equiv$ Exponential:

$$E[Y|\mathbf{X}] = G \left(\alpha + \sum_{j=1}^d f_j(X_j|M) \right)$$

This model, considered by Conversano (1998), Mola (1998) and Conversano and Siciliano (1999), is based on the use of a regression tree procedure for the identification of the optimal smoothing parameters in GAM. In this way, we obtained an integration between nonparametric and semiparametric models.

3 The proposed algorithm

In this section we describe the estimation procedure for the GAM-M model (2). Additional details about the algorithms implemented in *S+* can be found in Conversano (1999).

The models used in this procedure depend on the type (categorical or numerical) of the variables. If both the response variable and the predictors are *numeric*, then we consider models such as linear regression, local regression, smoothing spline and regression tree. If the response variable is *categorical*, we consider models such as linear discriminant analysis, classification trees and nonparametric logistic regression.

We describe the structure of the algorithm in the case of a numerical response variable. We set $M_1 \equiv 1$ as the unspecified model, $M_2 \equiv LR$ as the linear regression model and $M_3 \equiv Tree$ as the regression tree procedure. Then $f(X_j|M_1)$ is the estimation coming from the use of a smoothing function f_j for the j -th predictor. Alternatively, $f(X_j|LR)$ is the estimation produced by the linear regression model and $f(X_j|Tree)$ is the result of a regression tree. The algorithm is characterized as follows.

The algorithm for GAM-M

1. For each predictor fit the models M_1 , M_2 and M_3 , choose the best model/smooth that minimizes the residual sum of squares and fix the corresponding dummy variable $\delta_{ij} = 1$.

2. Initialize. Put iteration counter $r = 0$.

Set the assignments $f_j^{(0)}(\mathbf{x}_j|LR)$, $f_j^{(0)}(\mathbf{x}_j|Tree)$ and $f_j^{(0)}(\mathbf{x}_j|M_1)$ to constant. Center both the predictors $\mathbf{x} = (x_1^T, \dots, x_d^T)^T$ and the predictand \mathbf{y} , and save the means.

3. Update. Put iteration counter $r = r + 1$.

Fit the model $f_j^{(r)}(\mathbf{x}_j|LR)$ to the residuals

$$\{\mathbf{y} - [f_j^{(r-1)}(\mathbf{x}_j|Tree) + f_j^{(r-1)}(\mathbf{x}_j|M_1)]\}.$$

Grow the tree $f_j^{(r)}(\mathbf{x}_j|Tree)$ for the residuals

$$\{\mathbf{y} - [f_j^{(r-1)}(\mathbf{x}_j|LR) + f_j^{(r-1)}(\mathbf{x}_j|M_1)]\},$$

and derive the smoothed regression model for $f_j^{(r)}(\mathbf{x}_j|M_1)$ from the residuals

$$\{\mathbf{y} - [f_j^{(r-1)}(\mathbf{x}_j|LR) + f_j^{(r-1)}(\mathbf{x}_j|Tree)]\}.$$

4. Stopping rule. Cycle step 3 until convergence.

4 The benchmarking of the proposed methodology

We compare in terms of 10-fold cross-validated predictive accuracy the Mean Squared Error of Prediction ($MSEP$) obtained from our GAM-M model with that coming from some classical statistical models, such as multiple linear regression (LR) using stepwise variable selection criteria, regression tree ($Tree$) using CART pruning with $SE = 0$ selection rule and GAM. We used the data set *employers* from the library SPSS concerning a sample of 474 employers and the aim is to predict the actual salary (Y) with respect to the initial salary (X_1), number of working months in previous employment (X_2), number of working months during the actual employment (X_3), and education level (X_4). The results are summarized as follows.

- Model LR : $iter = 4$, $MSEP = 542.821$;
- Model $Tree$: $iter = 5$, $MSEP = 401.694$;
- Model GAM : $iter = 3$, $\{f_1, f_3\} \equiv Linear$, $f_2 \equiv Loess$, $f_4 \equiv Spline$, $MSEP = 445.647$;
- Model $GAM-M$: $iter = 3$, $f_1(X_1|LR)$, $f_2(X_2|M_1) \equiv Loess$, $f_3(X_3|Tree)$, $f_4(X_4|M_1) \equiv Spline$, $MSEP = 301.699$;

It is clear that, with a lower number of iterations (r) respect to LR and $Tree$ models, we reduce considerably the $MSEP$ in our GAM-M approach. Other applications on real and simulated data sets, not shown here for a lack of space, demonstrate similar results.

5 Concluding remarks

Like the GAM approach, the GAM-M model stays additive in nature and the choice of either the model or the smoother is done for each predictor separately. Nevertheless, our GAM-M approach substantially improves the quality of the estimation coming from a combination of models and smoothers. In particular, it removes for each predictor the observations not influencing the estimation provided by a certain model assigning their residuals to the other models. The proposed approach is totally different from the methods based on the use of some combination of functions and/or variables, such as the bagging procedure of Breiman (1996). These procedures work with models which are previously defined, whereas our procedure adapts simultaneously the estimations coming from different models in the successive iterations, in similar way as it is used in GAM.

References

- BREIMAN, L. (1996): Bagging predictors. *Machine Learning*, 26, 46–59.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984): *Classification and Regression Trees*, Belmont C.A. Wadsworth.
- CONVERSANO, C. (1998): A Regression Tree Procedure for Smoothing in Generalized Additive Models. In M. Huskova et al. (eds.): *Prague Stochastics'98 Abstracts*, 13-14, Union of Czech Mathematicians and Physicists.
- CONVERSANO, C. (1999): *Semiparametric Models for Supervised Classification and Prediction. Some Proposals for Model Integration and Estimation Procedures* (in Italian), Ph.D Thesis in Computational Statistics and Data Analysis, Universitá di Napoli Federico II.
- CONVERSANO, C., and SICILIANO, R. (1998): A regression tree procedure for smoothing and variable selection in generalized additive models, submitted.
- FRIEDMAN, J.H. (1991): Multivariate adaptive regression splines. *The Annals of Statistics*, 19, 1-141.
- HASTIE, T.J., and TIBSHIRANI, R.J. (1990): *Generalized Additive Models*. Chapman and Hall, London.
- MERTENS, B.J., HAND, D.J. (1999): Adjusted estimation for the combination of classifiers. In D.Hand et al. (eds.): *Intelligent Data Analysis IDA99 Proceedings*. 317–330, Springer, Berlin.
- MOLA, F. (1998): Selection of cut points in generalized additive models. In M. Vichi and O. Optiz (eds.): *Classification and Data Analysis: Theory and Application*, Springer Verlag, Berlin, 121–128.
- MOLA, F., SICILIANO, R. (1997). A fast splitting procedure for classification trees, *Statistics and Computing*, 7, 208–216.
- SICILIANO, R., and MOLA, F. (1994): Modelling for recursive partitioning and variable selection. In: R. Dutter and R. Grossman (eds.): *Compstat'94 Proceedings*. Phisycal-Verlag, Heidelberg, 172–177.

Radial Basis Function Networks and Decision Trees in the Determination of a Classifier

Rossella Miglio and Marilena Pillati

Dipartimento di Scienze statistiche, Universitá degli Studi di Bologna,
Via Belle Arti, 41, 40126 Bologna, Italy
(e-mail: miglio@stat.unibo.it - pillati@stat.unibo.it)

Abstract. In this paper a nonparametric classifier which combines radial basis function networks and binary classification trees is proposed. The joint use of the two methods may be preferable not only with respect to radial basis function networks, but also to recursive partitioning techniques, as it may help to integrate the knowledge acquired by the single classifiers. A simulation study, based on a two-class problem, shows that this method represents a valid solution, particularly in the presence of noise variables.

1 Introduction

Artificial neural networks represent a rich class of methods often used in supervised classification problems (for a review, see among the others Bishop (1995) and Ripley (1996)).

In the class of neural network models, multilayer perceptrons and radial basis function networks can be represented as multilayer networks, in which the activations of the hidden layer units are nonlinear basis functions. The parameters of a multilayer perceptron are determined simultaneously by computational intensive numerical optimisation techniques. Radial basis function network parameters, on the other hand, are typically estimated in two steps. The main advantage of these networks is that they need much lower training time in comparison with multilayer perceptrons. However, a critical aspect in radial basis function networks is the determination of the parameters of the first step, *i.e.*, the centre and the width of each basis function, which determine in part the quality of the results given by the networks.

Several strategies for selecting the parameters of the basis functions have been proposed, most of which are based on the predictor (or input) variable values ignoring any information about the response variable. *i.e.*, the class membership.

The main purpose of this paper is to show how a classification tree could represent a supervised solution for the first stage of the estimation problem in radial basis function networks. Section 2 introduces radial basis function networks. Section 3 discusses the usefulness of a classifier based on combining classification trees and radial basis function networks. In Section 4, the results of the simulation study show that improvements are obtained both

with respect to the use of decision trees and with respect to other methods for determining the radial basis function parameters, particularly in presence of irrelevant variables.

2 Radial basis function neural networks

In the last ten years radial basis function networks have received considerable attention as an alternative to the traditional and widely used multilayer perceptrons. The often used architecture of radial basis function networks consists of three layers, in which the outputs are generally represented as linear combinations of h nonlinear functions of the vector \mathbf{x} . Unlike the multilayer perceptrons, in which these nonlinear functions are sigmoidal, the hidden unit activations of radial basis function networks are radially symmetric, often Gaussian, and depend only on the distance between the input vector \mathbf{x} and the centre of each hidden unit. The radial basis function network mapping computed by the k -th output units takes the following form

$$g_k(\mathbf{x}) = \sum_{j=0}^h \alpha_{kj} \phi_j(\mathbf{x}), \quad (1)$$

where ϕ_j is a nonlinear "activation" function and α_{kj} is the weight of the connection between the j -th hidden unit and the k -th output unit (the weight α_{k0} is related to a fixed $\phi_0 = 1$). In the Gaussian case and for Euclidean distance (1) can be written as

$$g_k(\mathbf{x}) = \sum_{j=0}^h \alpha_{kj} \exp\left(\frac{\|\mathbf{x} - \mu_j\|^2}{2\sigma_j^2}\right),$$

where μ_j is the vector determining the centre of the j -th basis function and the standard deviation σ_j is the width of the basis function. Unlike the multilayer perceptrons, in which all of the parameters are estimated simultaneously, radial basis function network parameters are typically estimated in two steps, due to the different roles of the first and second layer weights.

The parameters governing the nonlinear mapping from the input layer to the hidden layer, *i.e.*, μ_j and σ_j for the Gaussian basis function considered above, are generally first determined by relatively fast unsupervised methods, which use only the input data set \mathbf{x}_i ($i = 1, \dots, n$). Furthermore, the second stage implements a linear transformation from the hidden to the output layer and the estimation of the parameters α_{kj} ($k = 1, \dots, K$; $j = 1, \dots, h$) is also very fast. Therefore, one of the main reasons of the increasing interest in radial basis function networks is their computational efficiency, even if classification results can be slightly worse than with multilayer perceptrons. As a result, the key aspect in radial basis function networks becomes the determination of the parameters of each basis function. In particular, of crucial concern is

the choice of the number of the centres, which determines the complexity of the classifier.

Several strategies for selecting the parameters of the basis functions have been proposed, most of which are based on the predictor variable values and ignored any information about the class labels.

One simple procedure for selecting the basis function centres is to set them equal to a random (or stratified) sample of observations. As an improvement, clustering techniques can be used to find centres that more accurately represent the distribution of the data over the predictor space. Moody and Darken (1989) proposed a *k*-means clustering algorithm and this has often been used.

The width σ_j can be obtained by different methods. One simple approach is to choose a uniform width for all the basis functions given by some multiple of the maximum (or average) distance between the basis function centres. Such a choice is generally associated with a randomly selection of the centres. As an alternative, each basis function j may have different width σ_j . Furthermore, it is possible to associate different standard deviations to each component of the predictor space.

3 The use of classification trees to determine the complexity of radial basis function networks

The free parameters of the first layer are derived by an unsupervised clustering technique using only the distribution of the training data in the predictor space. As an alternative, the class membership and nonlinear optimisation techniques can be used to estimate all the neural network parameters. It can be shown that for this approach the classification performance of a radial basis function network is similar to that of a multilayer perceptron, but the computational advantages disappear (Tarassenko, 1994).

The nonparametric classifier we present in this paper is based on the intuitive requirement to have the parameters of the first layer depend in some way on the available class information, without increasing training times. Instead of using an unsupervised clustering technique, we show how a binary recursive partitioning method (see, among the others, Breiman, Friedman, Olshen and Stone, 1984) could represent a supervised technique to solve the first stage of the estimation process.

With binary classification trees it is possible to find a collection of nonoverlapping subregions, given by the terminal nodes in the tree, that may represent the starting point to derive the support of each radial basis function in the network. As a result, our proposal makes the complexity of the radial basis functions depending on the classification tree, being the number of input and hidden units derived by the tree. The radial basis function centres are quite naturally determined by the centres of gravity of the regions selected by the tree. This approach, unlike the unsupervised one, determines a partition of the predictor space taking into account also the class-membership

information. Furthermore, the ability of the classification tree to select the variables that play the main role in discriminating between classes reduces the radial basis function networks dependence on irrelevant predictors. On the contrary, when the basis function centres are chosen using only the input data, there is no way to distinguish relevant from irrelevant variables.

The joint use of the two methods to generate a single classifier may be preferable not only for radial basis function networks, but also for recursive partitioning techniques. Similar to multilayer neural networks, a binary classification tree can be written using an expansion in a set of basis functions. In fact, each terminal node in a tree corresponds to a region of the predictor space and the collection of such nonoverlapping regions can be viewed as a set of basis functions, each represented by the product of univariate step functions. Recursive partitioning methods have the ability to exploit the local dimensionality of the relationship between the response variable and the predictors. However, one of the main drawbacks of this class of methods is that the classifiers generated by decision trees are discontinuous at the subregion boundaries. Unlike multilayer perceptrons, radial basis function networks have a "local" nature, because their radially symmetric functions have support over a local region of the predictor space. In other words, only few hidden units have significant activations for a given predictor vector. It is important to note, however, that this does not necessarily mean that the basis functions have non-overlapping support, such as trees. Thus, radial basis function networks derived from the results obtained by classification trees may represent a way to overcome the discontinuity problem.

4 Simulation study and concluding remarks

The proposed classifier has been applied to a simulated data set and compared to a classification tree and to two different radial basis function networks. The simulation study is based on a two-class data set with 6 dimensional measurement vectors. Class 1 is drawn with equal probability from a unit multivariate normal with mean (a, \dots, a) and from a unit multivariate normal with mean $(-a, \dots, -a)$. Class 2 is drawn from a unit multivariate normal with mean $(a, -a, a, -a, a, -a)$. Several situations were considered according to the inclusion of a different number of normally distributed noisy variables.

Fifty training samples of 500 observations each were generated and 5000 observations have been considered as a test set. The misclassification error rate of each classifiers has been estimated by averaging the fifty error rates on the test set.

Radial basis function networks, with a different number of hidden units, have been constructed following two different procedures, implemented by GAUSS algorithms. Firstly, we considered the networks obtained by the random selection of the basis function centres from the training sets and with constant widths proportional to the maximal Euclidean distance between

each pair of centres. Secondly, following Moody and Darken (1989), we implement the k -means clustering algorithm to determine the centres of the hidden layers and a P -nearest neighbours heuristic for the widths ($P=10$).

The classification trees have been generated by CART 3.6, and selected through a 10-fold cross-validation. Finally, to implement the proposed methodology we calculate the basis function centres from the elements of the partition obtained with CART, and a P -nearest neighbours heuristic has been considered to determine the width parameters. Tables 1 and 2 report the

Table 1. Test set error rate (%) with randomly selected centres.

Noise variables	Hidden units				
	6	10	12	18	24
0	20.12	17.52	17.95	17.94	18.01
1	20.56	19.81	18.97	18.92	18.01
2	23.78	21.65	19.05	19.25	20.52
3	23.53	20.42	20.17	21.05	21.01
4	26.09	22.35	22.86	22.21	23.39
5	25.52	23.22	23.62	23.94	24.07
6	26.45	24.91	23.95	26.16	26.76
12	30.56	29.97	30.62	30.19	32.32

Table 2. Test set error rate (%) with centres selected via k -means clustering technique.

Noise variables	Hidden units				
	6	10	12	18	24
0	15.58	16.87	16.76	15.85	15.34
1	15.91	17.11	17.05	16.19	15.89
2	15.96	17.22	17.13	16.42	16.38
3	16.37	17.86	17.74	17.06	17.08
4	16.67	18.15	17.99	17.33	17.50
5	17.02	18.32	18.12	17.70	17.92
6	17.70	18.85	18.47	18.26	18.44
12	21.16	23.27	22.40	21.99	22.07

average test set performances over 50 training samples of radial basis function networks with a randomly selection of centres and k -means clustering approach respectively. The tables present the results obtained for different neural network architectures and for different number of noise variables. The results show that the clustering solution outperforms the naive method of

centre selection. The performance of both classifiers gets worse as the number of noise variables increases. This effect is less evident for classifiers based on the clustering technique, whose performance considerably decreases when the number of irrelevant variables doubles the one of relevant predictors. The second column of Table 3 shows that the classifier we propose maintains a stable behaviour as the number of irrelevant predictors does increase. Hence, this classifier and the one based on clustering technique have quite similar performances only in absence of noise variables. The obtained results clearly highlight the good performance of the proposed classifiers also in presence of several noise variables. The first column of Table 3 shows that the tree classifier is not influenced by the noise variables and reveals the crucial contribution of binary partitioning techniques in the selection of the variables with the highest discrimination power. At the same time the method proposed outperforms the tree classifier, whatever the number of irrelevant variables. The results are particularly interesting because the preliminary use of decision trees allows to define a more parsimonious network. Following our proposal the average number of hidden units in the radial basis function network architecture is 11.4 and the modal value is 11.

The simulation results seem therefore to confirm that the combination of radial basis function networks and decision trees may help to integrate the knowledge acquired by the single predictors and to produce a better classifier.

Table 3. Test set error rate (%) for classification trees (CART) and radial basis networks with centres selected via classification trees (RBF-tree)

		Noise variables							
		0	1	2	3	4	5	6	12
CART	18.62	18.69	18.72	18.72	18.73	18.77	18.80	18.80	
	15.40	15.53	15.66	15.66	15.76	15.97	16.43	16.43	
RBF-tree									

References

- BISHOP, C.M. (1995): *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.
- BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE C.J. (1984): *Classification and Regression Trees*, Wadsworth and Brooks/Cole, Monterey.
- MOODY, J. and DARKEN, C.J. (1989): Fast Learning in Networks of Locally-Tuned Processing units. *Neural Computation 1* (2), 281-294.
- RIPLEY, B.D. (1996): *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.
- TARASSENKO, L. (1994): discussion of Ripley B. D. Neural networks and related methods for classification, *Journal of the Royal statistical Society, B* 56(3), 409-456.

Clustered Multiple Regression

Luis Torgo¹ and J. Pinto da Costa²

¹ LIACC-FEP, University of Porto,
R. Campo Alegre, 823, 4150 Porto, Portugal
(e-mail: ltorgo@ncc.up.pt)

² LIACC-DMA-Faculty of Sciences, University of Porto,
R. Campo Alegre, 823, 4150 Porto, Portugal
(e-mail: jpcosta@ncc.up.pt)

Abstract. This paper describes a new method for dealing with multiple regression problems. This method integrates a clustering technique with regression trees, leading to what we have named as *clustered regression trees*. We use the clustering method to form sub-samples of the given data that are similar in terms of the predictor variables. By proceeding this way we aim at facilitating the subsequent regression modeling process based on the assumption of a certain smoothness of the regression surface. For each of the found clusters we obtain a different regression tree. These clustered regression trees can be used to predict the response value for a query case by an averaging process based on the cluster membership probabilities of the case. We have carried out a series of experimental comparisons of our proposal that have shown a significant predictive accuracy advantage over the use of a single regression tree.

1 Introduction

This paper describes a new approach to multiple regression whose main distinguishing feature is the integration of a clustering procedure with a standard regression method. This new approach, that we refer to as *clustered regression trees* (CLRT), performs an initial clustering of the given sample of data and then applies the regression method to each found cluster. Thus clustering plays the role of a kind of internal resampling strategy, that groups *similar* observations. The result of CLRT is a set of models: one for each of the clusters found before. Given a test case for which we want to predict the response variable value, CLRT starts by obtaining its cluster membership in the form of a probability. The regression models associated to the clusters to which the test case may belong are used to obtain a prediction. The final CLRT prediction is a weighted average of these individual predictions.

In the implementation of the method briefly described above we have used AUTOCLASS C¹ (Cheeseman et al., 1988; Cheeseman and Stutz, 1995) as the clustering method. Regarding the multiple regression method we have used RT (Torgo, 1999)² that generates regression trees from samples of data.

¹ <http://ic-www.arc.nasa.gov/ic/projects/bayes-group/autoclass/>

² <http://www.ncc.up.pt/~ltorgo/RT/>

2 Clustered regression trees

The main motivation behind our proposed method is the assumption that the unknown regression function being modeled should have a certain degree of smoothness. This means that cases that are “nearer” in terms of the multi-dimensional space defined by the predictor variables, should have similar values of the target (response) variable. In a certain way this is the same assumption underlying local regression methods (*e.g.* Cleveland and Loader, 1995). Furthermore, we assume that it should be easier for a regression technique to model a set of similar cases than a larger set containing quite different observations. Based on these assumptions, we propose to use a clustering method to try to divide a given sample of observations in a set of sub-samples based on the similarities of the cases that are discovered by the clustering algorithm. Each of these sub-samples is then used to obtain different regression trees that all together form what we referred to as clustered regression trees.

2.1 Clustering-based resampling

We use AUTOCLASS C (Cheeseman et al., 1988; Cheeseman and Stutz, 1995) to obtain a clustering of the given training sample. For this clustering task we use only the information of the predictor variables. Why do we not use the response variable values? The answer is again related to the motivations behind local regression modeling (*e.g.* Cleveland and Loader, 1995). In effect, these approaches obtain one model for each test (query) case using only the training observations within a certain neighborhood (bandwidth) of the query case. This neighborhood is obtained using a metric over the input space defined by the predictors. This means that only the most similar (in terms of the predictors) training observations are used to obtain each local model. The assumption behind these approaches is that if the unknown regression surface is reasonably smooth only nearby cases are necessary to obtain the prediction for a test case. Our approach is based on the same motivation. However, instead of “building” a neighborhood for each test case, we use a clustering method to obtain a fixed set of relevant neighborhoods. As in local modeling, only the predictors are used to obtain these sub-sets of the training observations. Compared to local modeling our method is significantly more efficient in computation terms because the neighborhoods are fixed from the start and are not obtained for each test case. Moreover, as we have a fixed set of neighborhoods (the clusters) we also have a fixed set of models built for each cluster, which means that our approach is more interpretable to the user than local modeling.

As a result of the clustering process AUTOCLASS outputs the “best” number of classes (clusters) and also attaches to each given training case a cluster membership probability. AUTOCLASS follows a Bayesian approach to the

automatic clustering of data. It takes a database of cases described by a combination of real and discrete valued attributes, and automatically finds the natural classes in that data. The fundamental model is the classical finite mixture distribution. In the first part of this model, each case, X_i , has a certain probability of belonging to class C_j . In the second part, each class, C_j , is modeled by a class p.d.f. belonging to a certain family. One important characteristic of this system is that it uses weighted assignment of the cases to the classes. The classes are therefore described probabilistically, so that an object can have partial membership in the different classes, and the class definitions can overlap. AUTOCLASS does not need to be told how many classes are present or what they look like – it extracts this information from the data itself.

The output of AUTOCLASS is a set of j clusters of the given sample. Based on this clustering we generate sub-samples of the cases. For each case in the initially given sample we check its cluster membership and “attach” it to the respective sub-sample. Cases that have a certain probability of belonging to more than one cluster are inserted in the respective clusters. This means that there may exist some overlap between the j sub-samples of cases.

2.2 Growing regression trees for the sub-samples

We use RT to grow a regression tree for each of the j sub-samples. RT (Torgo, 1999) is able to obtain many variants of tree-based regression models. In our experiments we have used a “standard” least squares regression tree similar to the ones generated by the CART system (Breiman et al., 1984). A tree-based regression model consists of a hierarchy of nodes, starting with a top node known as the root node. Each node of the tree contains logical tests on the input (predictor) variables, with the exception of the bottom nodes of the hierarchy. These latter are usually known as the leaves of the tree. The leaves contain the predictions of the tree-based model. Most existing methods of growing a regression tree try to minimize the mean squared error of the resulting model. Using this criterion, each test on a tree node is chosen as to minimize the weighted variance of the resulting partition entailed by the test. Trees are grown using a recursive partitioning algorithm that on each iteration tries to find the split test that minimizes this weighted variance criterion. Trees built using this algorithm are usually post-pruned so as to minimize overfitting effects originated by too much partitioning. We use a Chi-square pruning method (Torgo, 1999) that selects a tree from a candidate set of sub-trees of the initially grown tree, using reliable estimates of the mean squared error based on the Chi square distribution.

2.3 Predictions using clustered regression trees

AUTOCLASS C is able to attach to any observation the probability of belonging to each of the j clusters. As described in the previous section, we grow

one regression tree for each cluster. Each of these models can be used to obtain a prediction for a given query case. The final prediction for any query case is obtained through a weighted average of the predictions of the j regression trees. We use as weights of each prediction the probabilities obtained by AUTOCLASS C regarding the cluster membership of the query case, i.e.,

$$Y'(\mathbf{q}) = \sum_{k=1}^j (P_k(\mathbf{q}) * RT'_k(\mathbf{q})) \quad (1)$$

where,

\mathbf{q} is a query case;

j is the number of clusters found by AUTOCLASS C;

$P_k(\mathbf{q})$ is the probability of query case \mathbf{q} belonging to cluster k ;

$RT'_k(\mathbf{q})$ is the prediction of the regression tree obtained with cluster k .

3 Experimental evaluation

In this section we describe a series of experiments whose main goal is to evaluate our proposed clustered regression trees (CLRT). These experiments were designed to compare CLRT with the alternative of using a single regression tree (RT) on all available sample.

Table 1 provides some details of the data sets we have used in our experimental comparisons³.

<i>Data Set</i>	<i>Main Characteristics</i>
Abalone	4177 cases; 7 continuous variables; 1 nominal variable Predicting the age of abalone.
Elevators	8752 cases; 40 cont. var. Aircraft control problem (prediction of elevators level).
F	3000 cases; 5 cont. var. Artificial domain with marked clusters of data points.
Kinematics	8192 cases; 8 cont. var. Robot arm control problem.
Telecomm	15000 cases; 48 cont. var. Telecommunications problem.
Computer	8192 cases; 22 cont. var. Prediction of CPU activity level in a computer network.
Computer(small)	8192 cases; 12 cont. var. Simplified version of the previous data set.

Table 1. The Used Data Sets.

For each of these data sets we have carried out 5 repetitions of a standard 10-fold Cross Validation process, to estimate the mean squared error (MSE) of

³ Further details in <http://www.ncc.up.pt/~ltorgo/Regression/DataSets.html>

both alternatives (CLRT and RT). We also present the statistical significance level of the observed differences, asserted by paired t -tests.

The results of the comparison of CLRT with RT are presented in Table 2. For each alternative we present the average MSE together with the respective standard deviation. In the fourth column we present the statistical significance results regarding the observed differences in the average MSE of both methods. A “+” sign means that the score of CLRT is significantly better than the result of RT with 99% confidence. A “-” sign means the same but favoring RT.

<i>Data Set</i>	<i>MSE of CLRT</i>	<i>MSE of RT</i>	<i>Significance</i>	<i>N. of Clusters (CLRT)</i>
Abalone	2.468 ± 0.712	5.406 ± 0.653	+	7.62 ± 0.98
Elevators	9.733 ± 4.180	17.380 ± 2.747	+	11.72 ± 6.1
F	4.312 ± 2.291	7.561 ± 1.802	+	10.98 ± 2.1
Kinematics	0.0321 ± 0.0112	0.0389 ± 0.0023	+	12.61 ± 2.6
Telecomm	76.706 ± 16.470	61.901 ± 7.785	-	11.48 ± 3.8
Computer	8.214 ± 0.914	12.348 ± 1.892	+	11.08 ± 3.9
Computer(small)	7.648 ± 1.496	14.362 ± 2.349	+	11.36 ± 2.2

Table 2. The Results of Comparing CLRT with RT.

These results show an overwhelming advantage of our proposed method over the non-clustered regression tree variant. There is a single exception in the *Telecomm* data set for which we still do not have a convincing explanation.

There are several possible causes for the results of CLRT. The first is our hypothesis concerning the fact that applying a regression method on similar training cases is advantageous as it simplifies the modeling task. A second possible explanation for the observed error decrease is the use of multiple models. In effect, several authors have provided strong evidence towards the advantages of building multiple models based on different samples of the given data. Examples of such multiple model approaches are *bagging* (Breiman, 1996) and *boosting* (Schapire, 1990; Freund and Schapire, 1996). We have carried out an initial set of experiments comparing our CLRT method with a *bagged* RT, and these have shown that CLRT manages to have some advantage on several data sets (although the opposite also occurs). These results indicate that we can be confident on the hypothesis that the error advantages that we have observed when compared to RT are also due to the clustering-based resampling strategy that we have proposed.

4 Conclusions

We have presented a new multiple regression method that combines a Bayesian clustering technique with regression trees. In this methodology clustering is

used to find sub-samples of similar cases in terms of the predictor variables with the aim of facilitating the task of the regression technique. The result of this process can be regarded as a multiple model approach in the sense that several regression trees are built, one for each sub-sample. Predictions using our proposed clustered regression trees are obtained by an averaging process based on cluster membership probabilities.

The results of the experimental evaluation we have carried out show that clustered regression trees are able to significantly outperform “standard” regression trees. These predictive accuracy gains come at the cost of loosing some comprehensibility of the models and of additional computation costs as a result of the clustering process.

In this work we have “integrated” clustering-based resampling with regression trees. Still, the same idea can be carried out with other regression methods, or even with discriminant analysis techniques. In the near future we intend to explore these other directions of research.

References

- BREIMAN, L. (1996): Bagging Predictors. *Machine Learning*, **24**(3), 123–140. Kluwer Academic Publishers.
- BREIMAN,L., FRIEDMAN,J., OLSHEN,R. and STONE,C. (1984): *Classification and Regression Trees*. Wadsworth Int. Group.
- CHEESEMAN, P., KELLY, J., SELF, M. and STUTZ,J. (1988): AutoClass: A Bayesian Classification System. In: Proceedings of the Fifth International Conference on Machine Learning, Ann Arbor, MI. June 12-14 1988. Morgan Kaufmann, San Francisco, 54–64,
- CHEESEMAN, P. and STUTZ, J. (1995): Bayesian Classification (AutoClass): Theory and Results. In: Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy (Eds.): *Advances in Knowledge Discovery and Data Mining*. The AAAI Press.
- CLEVELAND, W. and LOADER, C. (1995) : Smoothing by Local Regression: Principles and Methods (with discussion). *Computational Statistics*.
- FREUND,Y. and SCHAPIRE,R. (1996) : Experiments with a new boosting algorithm. In: L. Saitta (Ed.): Proceedings of the 13th International Conference on Machine Learning. Morgan Kaufmann.
- FRIEDMAN, J. (1991): Multivariate Adaptative Regression Splines. *Annals of Statistics*, **19**:1, 1–141.
- SCHAPIRE,R. (1990) : The strength of weak learnability. *Machine Learning*, **5**, 197–227
- TORGO, L. (1999) : Inductive Learning of Tree-based Regression Models. Ph.D. Thesis. Department of Computer Science, Faculty of Sciences, University of Porto. (<http://www.ncc.up.pt/~ltorgo/PhD/>).

Constructing Artificial Neural Networks for Censored Survival Data from Statistical Models

Antonio Ciampi¹ and Yves Lechevallier²

¹ Department of Epidemiology & Biostatistics,
McGill University, Montreal, P.Q., Canada
and IARC, 150 Cours Albert-Thomas, Lyon
(e-mail: aciampi@epid.lan.mcgill.ca)

² INRIA-Rocquencourt,
78153 Le Chesnay CEDEX, France
(e-mail: Yves.Lechevallier@inria.fr)

Abstract. A general approach to the design and training of ANNs for censored survival data is presented, with statistical models used as building blocks. This provides efficient initialization and an aid to interpretation.

1 Introduction

Since Artificial Neural Networks (ANNs) are a powerful tool for constructing predictors, it is natural to use them for survival data. For any approach to prediction model building in survival analysis, the challenge consists in properly handling (*right*) censoring, *i.e.* the fact that for some individuals, indeed all those who survive or are lost to follow up, only a (*lower*) limit to the survival time (time to failure) is known (Cox and Oakes (1984)). Another challenge, particularly compelling, in view of the goals of survival analysis, is to find reasonably simple interpretations for the prediction. Now, ANNs are known to provide good predictions but these predictions are not easily interpretable. As a consequence, the application of neural nets to survival data has been limited. Typically, censoring is handled by discretizing the time variable, *e.g.* Biganzoli et al.(1998), Brown et al. (1997). By contrast, Faraggi and Simon (1995) use the Cox model to avoid time discretization: however their approach does not easily lend itself to 'on line' training, a potential limitation when working with very large data bases.

We have developed an alternative approach to the design and training of ANN for survival analysis. We do not discretize the time variable and do use 'on line' back-propagation for training the neural net. Furthermore, we base the ANN architecture on statistical models, thus extending the work presented in Ciampi and Lechevallier (1997), where ANN were used as predictors of a binary response. In this paper we present the general approach specifically focusing on architectures based on statistical models. The result is a viable approach to ANN for survival analysis which allows a partial interpretation via the imbedded statistical models.

2 Basic notions of survival analysis

Proportional Hazards (PH) is the most common simplifying assumption needed to develop a predictive model for survival. It concerns the form of the *hazard function* $h(t; z)$, which is defined as the negative logarithmic derivative of the *survival function* $S(t; z)$ with respect to time. The PH assumption is that the hazard can be factorized as a product of two functions, one depending on t , the other on \mathbf{z} only. This is usually written :

$$h(t; \mathbf{z}) = e^{\varphi(\mathbf{z})} \cdot h_0(t) \quad (1)$$

where $\varphi(\mathbf{z})$, the *log relative hazard*, is a real valued function, arbitrary except for the usual convention $\varphi(0) = 0$, and $h_0(t)$ is the *baseline hazard*, *i.e.* the hazard function for subjects with null covariate vector. In this paper, we will limit ourselves to models based on the PH assumption. However, most of the ideas discussed here can be extended to regression models based on other assumptions, such as Accelerated Failure Times, and Proportional Odds.

Let us denote by $\{(\mathbf{y}^{(n)}; \mathbf{z}^{(n)}) = (t^{(n)}, \delta^{(n)}; \mathbf{z}^{(n)}); n = 1, \dots, N\}$, a sample of N observations from the target population, where δ is the *censoring indicator*, a binary variable which takes values 1 or 0 according as t is an observed or a censored survival time. Under the PH assumption and the common further assumption of random censoring, see Cox and Oakes (1984), a general form for the log-likelihood of observing $(t^{(n)}, \delta^{(n)}; \mathbf{z}^{(n)})$ is :

$$l^{(n)} = \delta^{(n)} (\varphi(\mathbf{z}^{(n)}) + \log(h_0(t^{(n)}))) - e^{\varphi(\mathbf{z}^{(n)})} H_0(t^{(n)}) \quad (2)$$

with : $H_0(t) = \int_0^t h_0(u) du$. To estimate h_0 and φ from the data, additional assumptions about these functions are usually needed. Traditionally, linearity of φ is assumed : $\varphi(\mathbf{z}) = \beta \cdot \mathbf{z}$ and h_0 is either specified parametrically, or is left unspecified as in the approach known as Cox regression. In the latter case, the log-likelihood is replaced by the *partial likelihood* : this is a function of the regression vector and the data which does not contain the baseline hazard.

As larger and larger data bases become available, more information may be extracted directly from the data and the linearity assumption appears unnecessarily restrictive. Of the many non-linear methods that have been proposed, two stand out for their flexibility and interpretability: (generalized) additive models and (generalized) tree-growing, available also for censored survival data (Hastie and Tibshirani (1990), Ciampi (1991)). Though much more flexible than linear regression, they also impose restrictions on the shape of the log relative hazard. The next logical step is to use the output function of an ANN to approximate $\varphi(\mathbf{z})$, since even a relatively simple architecture like the 1-hidden layer, is (theoretically) sufficient to reproduce any real valued multi-variable function (Bishop (1995)).

3 A Multilayer Perceptron for survival analysis

The most common ANN architecture is the multilayer perceptron or feed-forward ANN. This is an input-output system consisting of a series of layers of artificial neurons (AN's), with N_h ANs in the h -th hidden layer, such that connections are allowed only between AN's of one layer and those of the next layer. Each AN is characterized by an activation functions, which transforms the weighted sum of its input into its output, which in turn feeds the next layer. Figure 1 shows a simple multilayer perceptron with only one hidden layer and a single output unit.

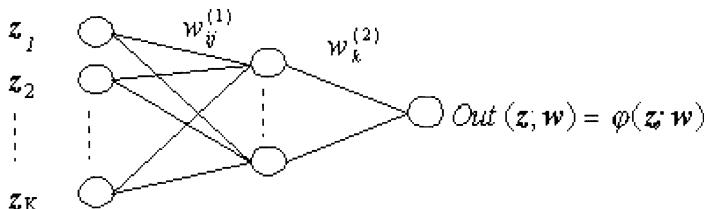


Fig. 1. A multilayer perceptron (1-hidden layer).

The connection weights, the w 's of Figure 1, denoted globally by \mathbf{w} , are determined by training the net with actual data. At the m -th passage, the net associates to $\mathbf{z}^{(m)}$ the output $Out(\mathbf{z}^{(m)}; \mathbf{w}^{(m-1)})$, compares it with the desired output $\mathbf{y}^{(m)}$ by means of a cost function $c(Out(\mathbf{z}^{(m)}; \mathbf{w}^{(m-1)}); \mathbf{y}^{(m)})$, and modifies the connection weights \mathbf{w} according to a learning law. The back-propagation algorithm (Bishop (1995)) is based on the learning law :

$$\mathbf{w}^{(m)} = \mathbf{w}^{(m-1)} - \alpha^{(m)} \text{grad}_{\mathbf{w}^{(m-1)}}(c) \quad (3)$$

commonly used in stochastic optimization. Thus a specification of a cost function and a calculation of its gradient are the two crucial steps in developing an ANN training algorithm.

From this general point of view, it is not difficult to see how to treat censored survival data. For any input \mathbf{z} , the ANN computes the network output function through $Out(\mathbf{z}; \mathbf{w}) = \varphi(\mathbf{z}; \mathbf{w})$, an estimate of the quantity $\varphi(\mathbf{z})$ defined by the PH assumption of equation (1). It remains to define the cost function : several forms of negative log-likelihood would be natural candidates. Although many possibilities are open, in this paper we adopt, for simplicity, a full-likelihood, parametric approach to modeling the baseline hazard, with the further restriction to a hazard constant in time, which defines the exponential model. Setting $h_0(t) = \lambda_0 = \text{constant}$, and redefining φ as $\varphi + \log(\lambda_0)$, in equation (2), the cost function becomes :

$$c(\varphi(\mathbf{z}^{(n)}; \mathbf{w})) = -\delta^{(n)}\varphi(\mathbf{z}^{(n)}; \mathbf{w}) + e^{\varphi(\mathbf{z}^{(n)}; \mathbf{w})}t^{(n)} \quad (4)$$

where now the constraint $\varphi(0) = 0$ is replaced by $\varphi(0) = \log(\lambda_0)$, making φ arbitrary. Derivation of (3) with respect to φ yields :

$$\frac{\partial c}{\partial \varphi} = -\delta^{(n)} + e^\varphi t^{(n)} \quad (5)$$

which can be used to write the back-propagation algorithm explicitly.

4 ANN's based on statistical models

In order to facilitate the interpretation of the input-output relationship, Ciampi and Lechevallier (1997) proposed an approach which associates, to a given statistical model, an ANN of rather special structure. For instance, the ANN of Figure 2 represents a generalized additive model. The activation functions of the hidden layer ANs are chosen as elements of a flexible basis of approximating functions, such as sigmoids, B-splines, radial functions etc. As Figure 2 shows, every component of the covariate vector feeds into a sub-net the output of which is a linear combination of basis function transformations of the original predictor. The single output AN has an identity activation function, and receives at its input a sum of arbitrary functions of the individual predictors, thus realizing a generalized additive model.

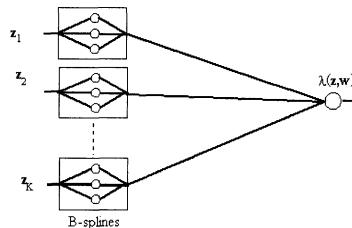


Fig. 2. An ANN realizing an additive model.

Notice that replacing the spline blocks with single ANs having linear activation functions, yields a linear model, which is a particularly important case of additive model. As another example, Figure 3 indicates how to associate an ANN to a tree. Figure 3a shows a RECPAM tree (Ciampi, 1992) and Figure 3b shows an ANN obtained by approximating this tree, with soft nodes replacing hard nodes.

The ANN has 3 hidden layers, all with $\text{sigmoid}[-1, 1]$ activation functions: the first layer defines splits based on individual predictors (internal nodes). The second, contains as many AN's as there are leaves in the tree; using the conjunction 'and', to define paths from the root node to the leaves (squares in Figure 3a). The third, contains as many AN's as there are RECPAM classes (hexagons).

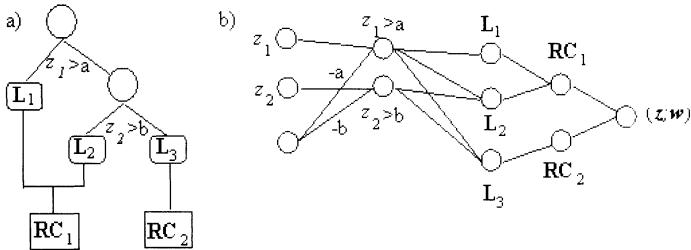


Fig. 3. Correspondence between a tree and the associated ANN (3 hidden layers).

In practice, one can start from a classical statistical analysis of a data set, based on a well known class of models, such as additive models or trees. Then the weights obtained from the classical analysis can be used as initialization of the corresponding ANN. ANN training will cause some departure from the initial conditions, but the final ANN will remain close to the original model, especially if no additional connections are allowed. But one can achieve an even higher flexibility by combining model-based ANNs into a network of networks, as shown in Figure 4 which combines two ANNs, based on an additive model and on a tree model respectively.

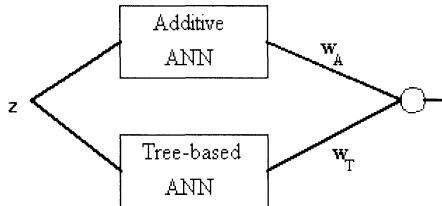


Fig. 4. Network of Neural Networks.

5 An example

The data analyzed here are from cardiology. Thirteen covariates describing cardiac rhythms were used to construct ANN based predictors of survival for a series of 1550 patients with heart disease. We divided the sample into learning (1000 subjects) and test sample (550 subjects). The learning sample was used in its entirety to obtain the regression and the tree models. On the other hand, for the training of the neural nets, we split the learning sample into a training sample and a validation sample, the latter being used to implement *early stopping*, see Bishop (1995).

The results on the test set automatically correct for the predictors' varying degree of flexibility, and clearly indicate that the ANN-based predictors perform substantially better than both regression-based and tree-based ones. The following table summarizes the results of our training experiments.

	Constant	Reg	Tree	ANN-Rand	ANN-Reg	ANN-Tree
Learning	607.45	573.73	563.14	559.36	556.53	550.53
Test	312.18	327.98	309.17	294.28	298.33	295.17

Table 1. Cost function (negative log-likelihood). **Constant** : Null model (without covariates). **Reg** : Exponential regression model, estimated in Splus. **Tree** : A tree obtained with RECPAM; the algorithm yields nine leaves and no amalgamation (RC classes and leaves are identical). **ANN-Rand** : A 1-hidden layer ANN with random initialization; using cross-validation we determined both the early stopping and the optimal number of hidden units, 8. **ANN-Reg** : A 1-hidden layer ANN with initialization corresponding to the regression model and the optimal number of hidden units, 5. **ANN-Tree** : A 2-hidden layer ANN based on the tree of Fig. 3.

6 Conclusion

The work outlined in this paper extends the applicability of the ANN paradigm to the prediction of censored survival data. Unlike most previous work, time discretization is avoided; furthermore, the use of statistical models as building blocks, provides a constructive approach to ANN implementation which facilitates interpretation. The example and our experience so far, show that even with a relatively small sample size (1000 subjects for 10-20 variables) ANN can achieve a substantial predictive improvement on classical statistical models. An in depth comparative study which would provide precise guidelines is highly desirable and would establish ANN in the practice of data analysis for censored survival data. It is, however, well beyond the scope of this work.

References

- BIGANZOLI, E., BORACCHI, P., MARIANI, L., MARUBINI, E. (1998): Feed Forward Neural Networks for he analysis of censored survival data : a partial logistic regression approach. *Statistics in Medicine*, 17, 1169-1186
- BISHOP, C.M. (1995): *Neural Networks for Pattern Recognition*. Clarendon Press.
- BROWN, S.F., BRANFORD A.J., MORAN W., (1997): On the use of Artificial Neural networks for the Analysis of Survival Data. *IEEE Transactions on Neural Networks*, 8, 1071-1077.
- CIAMPI, A. (1991): Generalized Regression trees. *Computational Statistics and Data Analysis*, 57-78.
- CIAMPI, A. and LECHEVALLIER, Y. (1997): Statistical Models as Building Blocks of Neural Networks. *Communications in Statistics*, 26(4), 991-1009.
- COX, D.R., OAKES, D. (1984): *Analysis of Survival Data*. Chapman & Hall.
- FARAGGI, D., SIMON, R. (1995): A Neural Network Model for Survival Data. *Statistics in Medicine*, 14, 73-82.
- HASTIE, T., TIBSHIRANI, R. (1990): *Generalized additive models*. Chapman & Hall.

Visualisation and Classification with Artificial Life

Alfred Ultsch

Department of Computer Science, Phillips-University of Marburg,
Hans-Meerwein-Str, 35032 Marburg, Germany
(E-mail: ultsch@informatik.uni-marburg.de)

Abstract. Systems that possess the ability of emergence through self-organization are a particular promising approach to Data Mining. In this paper, we describe a novel approach to emerging self organizing systems: artificial life forms, called DataBots, simulated in a computer show collective behavioural patterns that correspond to structural features in a high dimensional input space. Movement strategies for DataBots have been found and tested on a real world data set. Important structural properties could be found and visualized by the collective organisation of the artificial life forms.

1 Introduction

Systems that possess the ability of emergence through self-organization are a particularly promising approach to Data Mining. Self-organization means the ability of a biological or technical system to adapt its internal structure to structures sensed in the input of the system without external intervention. A biological example for self-organization is the organisation of swarms, e.g., bee swarms. Emergence means the ability of a system to produce a phenomenon on a new, higher level. This change of level is termed in physics "mode-" or "phasechange". It is produced by the cooperation of many elementary processes. Important technical systems that are able to show emergence are in particular laser and maser. In those technical systems billions of atoms (elementary processes) produce a coherent radiation beam (Haken (1974)).

Self-Organizing Neural Networks with emergent properties have been extensively studied by us in the past (Kohonen (1982), Ultsch et al. (1990), Ultsch (1993), Ultsch (1995), Ultsch (1998a)). In this paper we describe a novel approach to emerging self organizing systems: artificial life forms. The central idea is, that a large number of artificial life forms simulated in a computer show collective behavioural patterns that correspond to structural features in a high dimensional input space.

2 UD - Universe and DataBots

A UD-Universe (Umgebungs-Dynamik-Universum) is a world in which artificial life forms, so called Data Robots (DataBots), dwell. A UD-Universe

consists of a space, called UD-Matrix (Umgebungs-Dynamik-Matrix), which provides locations, called UNodes, where a DataBot may be at a certain moment in time. By his presence on a UD-Matrix a DataBot changes the

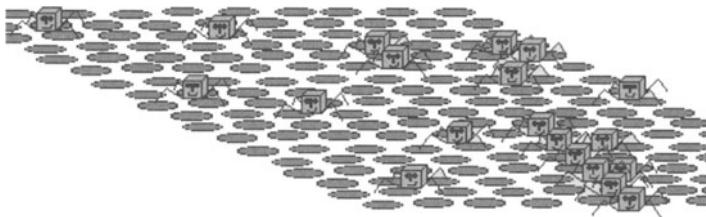


Fig. 1. Picture of a UD-Universe with DataBots.

UD-Universe. Especially a DataBot articulates its opinion (see below). A UD-Matrix is able to transmit news (opinions, scents, traces) and may be able to alter the transmissions. It may, for example, be possible to mix or weaken them. The UD-Matrix forms a constantly changing landscape for the representation of the opinions "at a glance" using U-Matrix-Methods (Ultsch (1993)). A DataBot is an artificial life form living in a UD-universe. A DataBot is able to maintain its own independent existence. By doing so it can take in food; consume food; store quantities of food (foodstuffs); stay at a certain location (UNode) in a UD-Matrix; propagate its opinion at its UNode with maximum weight and move on the UD-Matrix.

3 Movement

Grid UD-Matrices consists of neighbours in four directions. We call them North, East, South and West. The UNode itself is called O. The movement apparatus of a DataBot consist of five bins corresponding to O, N, E, S and W. These direction bins may contain positive numbers. When performing a move these numbers are rescaled to percentages, which may be regarded as probabilities for a certain direction of movement. That is, a probabilistic choice of directions is done on the basis of these percentages. If a movement is chosen by this process, all numbers in the direction bins are reset to zero. The direction of the move taken is stored in a movement memory that can be read by movement programs. Movement strategies manipulate the content in the direction bins. These programs may act simultaneously and concurring to each other. In analogy to nature we envision such movement strategies to evolve. The newer, fancier programs do not replace older ones, but work on top of them. In particular a strategy is sought for which the final location of a DataBot corresponds to the data distribution in the highdimensional vector-space of the opinions of the DataBots.

4 Movement programs for Data Mining

A movement program is sought for which the DataBots reveal structure in a highdimensional input-dataset. The idea is to provide each DataBot with an *n-dimensional* input-vector which is the opinion of a DataBot represented in the UD-Universe. One can imagine this as a scent or smell that a DataBot emits. By sensing or smelling other DataBots, more precisely by sensing the smells that the UD-Matrix is transmitting, a DataBot searches for locations where it likes the aroma. A DataBot searches, so to speak, for UNodes where its friends are. At the same time the DataBot tries to avoid bad smells, that is, it wants to get away from enemies. One goal of our research was to find movement programs for DataBots such that the location of a DataBot, that is, the UNode where the DataBot wishes to stay on the UD-Matrix, reveals the structure of the highdimensional input-dataset. In particular, if there are structures like clusters in the input-dataset the DataBots should cluster too. DataBots that have data (opinions) from the same highdimensional cluster should cluster together on a UD-Matrix. They should separate themselves from other DataBots that do not belong to the same cluster. We have tried several movement programs and found in particular one of them very useful for data clustering. This movement program, called "friends_and_foes" works as follows: a DataBot gets transmitted from the UD-Matrix all smells in the neighbourhood of a certain radius. The DataBot ranks the similarity respectively dissimilarity of the highdimensional smells. The 10 % best fitting smells are considered to be from friends and the 10 % worst smelling are considered to be foes. The movement program consists of a vector addition of the direction towards the friends plus a vector addition away from the foes. The resulting direction is converted to numbers for the directional bins of the DataBot. This movement program has been successfully tested on artificial data sets containing clusters. An example of a clustering problem from real data is described in Section 6.

5 Softwaresystem to simulate UD-Universes

We have implemented a simulation program for UD-Matrices called DataBots. The software is written in C++ using the QT graphical library (Malorny et al. (1998)). The simulation software can display the UNodes containing the DataBots and movements. Besides that the simulation program creates a visualization of the highdimensional structure of the data using U-Matrix technologies (Ultsch (1993)). Movement strategies, which are in the focus of our present research, may be programmed and modified while a simulation of a UD-Universe is running. The movement strategies can be expressed as ASCII text using elementary operations from a DataBots functional anatomy.

6 DataBots for Data Mining

The main difference of the artificial life approach to other clustering techniques is that local movement rules for each DataBot develop a nonlinear mapping of the highdimensional data onto a two dimensional grid that take not only the closest but also the whole topology into account. The difference in performance can be seen using a data set that consists of x/y coordinates of points evenly distributed inside two tangent circles. Figure 2 shows the performance of the DataBot clustering vs. Single Linkage and Ward clustering. As can be seen errors are minimal using DataBots. In order to test the

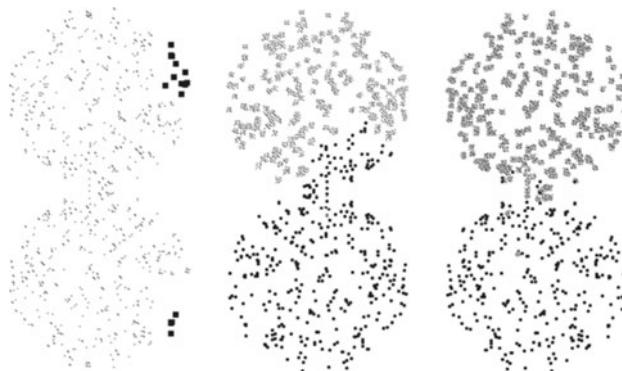


Fig. 2. Single Linkage (left), Ward (middle) and DataBot (right) Clustering.

clustering and self-organizing properties of the friends_and_-foes movement strategy for DataBots on a practical example, we used a dataset made available to us by Prof. Gasteiger described in Zupan et al. (1993). This dataset has been extensively studied using statistical and pattern recognition methods, see Zupan et al. (1993) pp. 168. The dataset consists of analytical data from 572 Italian olive oils produced in nine different regions of Italy. Figure 3 contains a map of the different regions from which the olive oils are taken. For each oil the percentual contents of 8 different fatty acids are measured. That is, the aroma of each DataBot is an 8-dimensional real-valued vector. Each DataBot was loaded with an 8 dimensional vector describing one olive oil. The number of DataBots used corresponds to the number of datavectors in the inputset. In this example we had 572 DataBots. The following picture shows the organisation of the DataBots on a 64 by 64 grid. To interpret the picture it must be understood that the picture is circular in each direction. The resulting clustering is more or less topology preserving. It can be concluded that the consistency of the olive oils vary according to the producing regions.

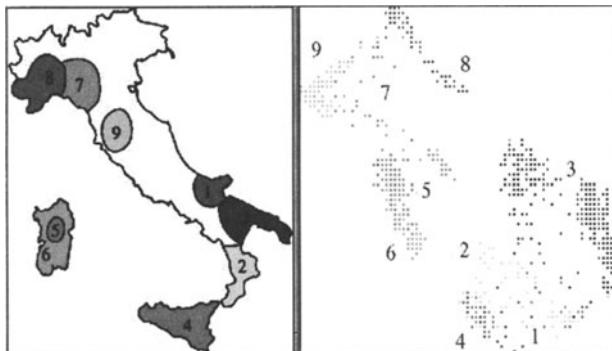


Fig. 3. Italian regions of origin of the olive oils compared to the distribution of DataBots on Olive Oil Dataset after 400 steps.

7 Conclusion

In this work we describe a novel approach to Data Mining using artificial life forms. To our knowledge this is one of the first attempts to use artificial life forms for Data Mining and Knowledge Discovery, while definitely in its first steps the approach shows surprisingly good performance. By the usage of self-organization the system shows emergent properties, see Ultsch (1999a). It could be shown that a very simple anatomy and very simple strategies lead to surprising results concerning the detection of clusters and preserving the overall structure of highdimensional datasets.

DataBots construct a mapping from highdimensional data space onto the two dimensional grid of a UD-Universe. Thus they are similar to principal components analysis (Hotelling 1933), multidimensional scaling (Shepard 1962) and Sammon's mapping (Sammon 1969). While principal components analysis and multidimensional scaling construct linear mappings, Sammon's mapping and DataBots construct nonlinear mappings. The latter is in particular advantageous, if the data space is linearly non separable like in the case of the chainlink data set (Ultsch 1996). Sammon's mapping emphasizes however the preservation of local distances while DataBots aim to preserve the overall topology. In this sense they are similar to emergent self organizing feature maps with U-matrix methods (Ultsch 1999). Our approach may lead to a very natural realization on parallel hardware using simple processors that cost less than 2 \$ in order to formulate U-Nodes and DataBots on a UD-Matrix. While the first version of DataBots was designed in April 1998 it took the work of several students mentioned below in order to provide us with a simulation tool that uses artificial life forms for the evaluation of high dimensional data structure (Ultsch 1999b).

Acknowledgements The author wishes to thank J. Gasteiger from University of Erlangen - Nuremberg and Prof. Fiorna from University of Genoa,

Italy for the olive-oil dataset. A first version of the UD-Universe has been implemented by Ingo Felger. A second version by Dirk Malorny, Ingo Müller and Falko Münchberg. The third version is currently being developed by Dirk Malorny.

References

- HAKEN, H. (1974): Synergetics, an Introduction, Springer, Berlin 1974
- HOTELLING, H. (1933): Analysis of complex statistical variables into principal components, *Journal of Educational Psychology*, 24, 417-441, 498-520, 1933.
- KOHONEN, T.(1982): Self-Organized Formation of Topologically Correct FeatureMaps, *Biological Cybernetics Vol. 43*, pp 59 - 69, 1982
- MALORNY, D., MÜLLER,I. and MÜNCHBERG, F.(1998): Realization of UD-Universes, Technical Note, *Department of Computer Science, University of Marburg, Hans-Meerwein- Str., 35032 Marburg, 25. Apr. 1998*
- SAMMON, J.R. (1969): A nonlinear mapping for data structure analysis., *IEEE Transactions on Computers*, 18:401-409, 1969.
- SHEPARD, R.N. (1962): The analysis of proximities: multidimensional scaling with an unknown distance function, *Psychometrika*, 27:125-140;219-246, 1962.
- ULTSCH, A.(1993): Self-organizing Neural Networks for Visualization and Classification, in *O. Opitz, B. Lausen and R. Klar, (Eds.) Information and Classification, Berlin: Springer-Verlag, 307-313, 1993*
- ULTSCH, A.(1995): Self-Organizing Neural Networks Perform Different from Statistical k-means clustering, *Gesellschaft f. Klassifikation, Basel 8th - 10th March, 1995*
- ULTSCH, A. (1996): Self Organizing Neural Networks perform different from statistical k-means clustering, In:*M. van der Meer, R. Schmidt, G. Wolf, (Eds.): BMBF Statusseminar, pp. 433 - 443, München. April 1996*, .
- ULTSCH, A.(1998a): The Integration of Connectionist Models with Knowledge-based Systems: *Hybrid Systems, Proceedings of the 11th IEEE SMC 98 International Conference on Systems, Men and Cybernetics, 11 - 14 October 1998, San Diego*
- ULTSCH, A.(1998b): Umgebungs dynamik Universen, Technical Note, *Department of Computer Science, University of Marburg, Hans-Meerwein- Str., 35032 Marburg, 25. Apr. 1998*
- ULTSCH, A.(1999a): Data Mining and Knowledge Discovery with Self-Organizing Feature Maps for Multivariate Time Series,in *Oja,E., Kaski,S.: Kohonen Maps, p 33- 46, Elsevier, 1999*.
- ULTSCH, A.(1999b): Clustering with Data Bots, *Research Report No. 19, Department of Computer Science, University of Marburg, 1999*
- ULTSCH, A. and SIEMON, H.P.(1990): Kohonen's Self Organizing Feature Maps for Exploratory Data Analysis, *Proc. Intern. Neural Networks, Kluwer Academic Press, Paris, 1990, pp. 305 - 308*
- ZUPAN, J. and GASTEIGER, J.(1993): Neural Networks for Chemists, *VCH, Weinheim New York, 1993*

Exploring the Periphery of Data Scatters: Are There Outliers?

Giovanni C. Porzio¹ and Giancarlo Ragozini²

¹ Dipartimento di Scienze Statistiche,
Università degli Studi di Napoli Federico II,
Via G. Sanfelice 46, 80134 Napoli, Italy (e-mail: gporzio@unina.it)

² Dipartimento di Matematica e Statistica,
Università degli Studi di Napoli Federico II
Via Cintia, 80126 Napoli, Italy (e-mail: giragoz@unina.it)

Abstract. Outliers are observations that are particularly discordant with respect to others, lying hence on the periphery of the data region. In the literature, many tools have been proposed with the aim of detecting multiple outliers. Most of the recent and attractive methods are based on some measure of the distance of each data point from a center. However, they are really effective only if the shape of the data scatter is symmetrical with respect to such a center. Otherwise, asymmetry will make these measures misleading. For this reason, we propose a method that allows direct exploration of the periphery of the data scatter, without considering any center. The methodology we propose is based on a two-step procedure that exploits the sample convex hull and radial projections. It explores gaps in the data scatter and proximities to its boundary, highlighting how the data structure is sparse at its periphery. A complementary graphical display is finally offered as a useful tool to visualize boundary features.

1 Introduction

The detection of multiple outliers in multivariate data is one of the most difficult statistical tasks, and the literature on this topic is extensive. Graphical methods can be particularly effective when few variables are involved, while in higher dimensions many of them can easily fail.

Alternatively, numerical methods can be considered. Some of them are based on a univariate index that is a measure of the distance of each observation from the center of the data cloud. The Mahalanobis distance (Mahalanobis, 1936) is the most widely used in practice. Other methods are based on the so-called omission approach. They evaluate the outlyingness of a point considering changes of some statistic due to the point deletion (one of the first proposals in such a direction was the λ statistic (Wilks, 1963)).

However, both these approaches are subject to the well-known masking effect. That is, these methods fail when two or more outliers lie close in a region of the space, masking themselves from the detection tools.

To overcome this main drawback, many solutions have been proposed. In an omission framework, the identification procedure could be iteratively

applied, deleting from the data set the most outlying point at each step (see Barnett and Lewis (1994) and references therein). Yet, it is well known that this sequential approach could fail too. Alternatively, a block omission approach can be considered. It consists of extending single deletion indices so that more than one observation at a time is omitted (e.g. Wilks, 1963). Block omission is very effective but it is not feasible when the number of the potential outliers increases, because of its combinatorial computational cost.

Distance measures have been devised to avoid masking as well. The most appealing idea considers distances from the center of a *clean set* as a starting point for an outlier search (Rousseeuw and van Zomeren, 1990; Hadi, 1992; Atkison, 1994). This last approach is essentially based on a robust estimate of the location parameter and of the covariance matrix in the Mahalanobis distance. Observations are ordered with respect to their distance from the center, and the most extreme points are taken as candidate outliers.

However, it may happen that data scatters present irregular structures so that it is not appropriate to consider a center. In this paper, we address this point and propose a new outlier detection method that enables us to deal with irregular and asymmetric data clouds. The method is based on an exploration of the data region boundary through a numerical and visual analysis of the sample convex hull vertices and their neighbourhoods. It is computationally feasible even in high dimensions and it also turns out to be an appropriate tool to reduce computation in a block omission approach.

2 Outliers in asymmetrical data scatter

The literature contains many definitions of outliers. For multivariate data, if a model is not assumed, outliers are those sample points that are particularly discordant with respect to the others. In the case of unimodal observed distributions, outliers will lie on the boundary of the data region and far from the bulk of the data (Barnett and Lewis, 1994; p.21).

When distance measures are used to detect outliers, the distance of each data point from the center of the data scatter is evaluated, and the furthest points are taken as candidate outliers. Let $\mathcal{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ be a sample observed on a p -dimensional random vector $\mathbf{X} \in \mathbb{R}^p$. The distance of a point \mathbf{x}_i from the center \mathbf{c} is evaluated through

$$D_i = (\mathbf{x}_i - \mathbf{c})' \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{c}),$$

where \mathbf{V}^{-1} is the inverse of the covariance matrix. The Mahalanobis distance has $\mathbf{c} = \bar{\mathbf{x}}$ and \mathbf{V} as the sample covariance matrix. Other proposed distances rely on robust estimates of location and scale parameters. In the following we will refer to any possible version of D_i as Mahalanobis-like distances.

The use of Mahalanobis-like distances to detect outliers corresponds to a reduced subordering (Barnett, 1976), i.e. a mapping from \mathbb{R}^p to \mathbb{R} through D_i . Therefore, candidate outliers are the extreme observations in the univariate

ordering in \Re . The point is that the same rank is assigned to observations symmetric with respect to the center \mathbf{c} (in \Re^p in the metric \mathbf{V}). Indeed, points with the same distance from the center lie on the same ellipses, as eqn. (1) describes *elliptical isodistance contours* with respect to \mathbf{c} .

Unfortunately, data scatters can have irregular or asymmetrical structures, and then Mahalanobis-like distances can fail in ordering data, and hence in outlier detection. With asymmetries, even if we can always compute a center that minimizes some function, it is not a symmetry center. In such cases, it may happen that an outlier has the same distance from the center (sharing the same ellipse) as other non-discordant observations.

As an example, consider the artificial data set in Fig. 1 that counts 350 observations, with five outliers sticking out on the left of the data cloud (crossed points). We will use this data set as an illustrative example throughout the paper. The ellipse corresponding to the largest value of the Mahalanobis distance has been superimposed on the scatter plot. We have also superimposed the convex hull of the data swarm with its 15 vertices, that we will discuss shortly. For the moment, consider first that the sample scatter has an asymmetrical shape (the mean values of the data are (0.0)), and, more important, note that the outliers are not so far from the center compared to other observations. However, they are isolated and discordant with respect to the main structure even if they lie inside the ellipse.

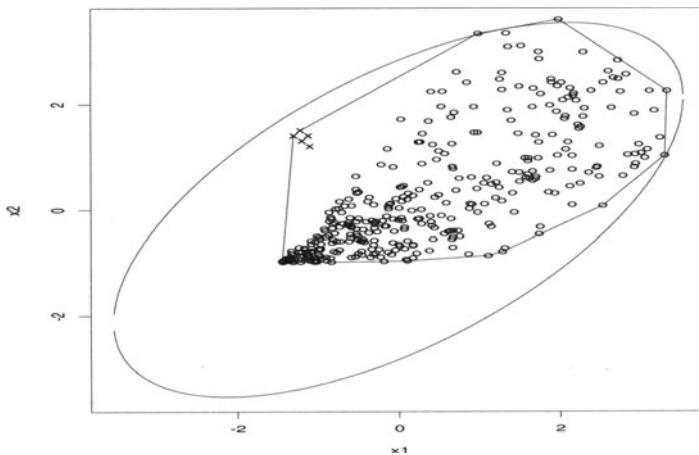


Fig. 1. Artificial asymmetric data set with five outliers.

Considering outliers just those data that lie far from the center is then not always appropriate. Rather, after Rolf (1975), it is more convenient to define anomalous those data that are somehow *isolated* and, hence, that generate a

gap in the scatter along *some directions*. This point of view is more general, and seems to work well with both symmetrical and asymmetrical shaped data. Starting from this idea, we propose a method that does not analyze data from a center towards the outside, but explores their periphery directly.

3 Selecting candidate outliers

To summarize, according to the above discussion, outliers are those points that both lie on the periphery of the data scatter *and* generate a gap along some directions. Consequently, possible outliers will lie on the boundary of the data region. To identify such a boundary, without any additional information, we propose to use the sample convex hull $\mathcal{CH}(\mathcal{X})$. It is defined as

$$\mathcal{CH}(\mathcal{X}) = \left\{ \mathbf{x} \mid \mathbf{x} = \alpha_1 \mathbf{x}_1 + \dots + \alpha_n \mathbf{x}_n, 0 \leq \alpha_i \leq 1, \sum_{i=1}^n \alpha_i = 1 \right\},$$

that is the smallest convex set containing the data points.

We consider as candidate outliers all of the convex hull vertices, as they correspond to the extremes of a multivariate sample (Barnett, 1976).

If more than one outlier occurs, and hence the chance of the masking effect, it may happen that some of them will not lie exactly on the boundary of the data region, but on its neighbourhoods (back to Fig. 1, two of the five outliers are convex hull vertices, while the others lie close to them, isolated from the remaining data). To allow for such an occurrence, we propose to identify as candidate outliers not only the convex hull vertices, but also all the neighbour observations that generate a gap with respect to the remainder.

In the univariate case a gap is simply spotted by a large empty interval in the observed values, while in the multivariate case it is a large empty region without a specific shape. Such an empty region will correspond to some univariate gaps if we project the data along some particular directions. As we cannot examine all the infinite possible directions, we use a radial projection that allows us to consider at the same time all the directions.

The procedure we propose will work as follows. In an initial step the convex hull vertices are included in the set of candidate outliers. Then we select as additional candidate the observations lying in a neighbourhood of the boundary region, if they are closer to the vertices than to the rest of the data. The distance of each data point from each vertex is measured along a radial projection. The presence of gaps along these directions will highlight empty spaces in the data structure, splitting the more peripheral observations from the data set core.

Specifically, the following steps can be considered.

Step 1: Construct the convex hull of \mathcal{X} , $\mathcal{CH}(\mathcal{X})$, and let $\mathcal{V} = \{\mathbf{x}_{i^*}\}$, with $i^* \in I^*$, be the set of v vertices of $\mathcal{CH}(\mathcal{X})$.

Step 2: Calculate the $v \times n$ distance matrix $\mathcal{D}(\mathcal{V}, \mathcal{X}) = \{d(i^*, j)\}$, $i^* \in I^*$, $j = 1, \dots, n$ with $d(i^*, j) = \|\mathbf{x}_{i^*} - \mathbf{x}_j\|_2$ the distance between the i^* -th vertex and the j -th data point. Let $d(i^*, \cdot)$ be the row vector of $\mathcal{D}(\mathcal{V}, \mathcal{X})$, which collects the distances between \mathbf{x}_{i^*} and $\mathbf{x}_j \in \mathcal{X}$.

Step 3: For each vertex i^* , sort in ascending order the $d(i^*, \cdot)$ vector; call $d_{(k)}(i^*, \cdot)$ the k -th element in the sorted sequence. Then calculate the first difference of the sorted distance vectors $[d_{(k+1)}(i^*, j) - d_{(k)}(i^*, j')]$, for $k = 1, \dots, n-1$. For each vector compute the maximum of such first differences and let k_{i^*} be the corresponding k :

$$k_{i^*} = \arg \max_k [d_{(k+1)}(i^*, j) - d_{(k)}(i^*, j')] .$$

For each vertex \mathbf{x}_{i^*} , the maximum identifies a *gap* and then a group of possible isolated observations. This group will consist of all data before the gap in the sorted distances sequence. Then any gap and any group have to be further analysed to decide towards outlyingness.

Note that, unlike Mahalanobis-like distance methods, we do not consider distances from a center, and then a unique univariate ordering, but rather the proximity relationships among points.

4 A graphical tool to explore outlyingness

Before applying any confirmatory analysis we suggest more in-depth analysis of the data structure and the set of candidate outliers through an *ad hoc* graphical tool. We propose to represent the distance sequences $d(i^*, \cdot)$ as side-by-side dotplots, one for each vertex.

This graphical display provides information about all existing gaps and other sparse structures at the data scatter periphery. To enhance the analysis, we suggest comparing the dotplots, and then the gaps, taking also into account the relative variabilities.

Fig. 2 has the proposed representation for the artificial data set displayed in Fig. 1. We will refer to the convex hull vertices, from the left to the right clockwise (Fig. 1), as $1^{st}, 2^{nd}, \dots, 15^{th}$ vertex. The 1^{st} and $13^{th} - 15^{th}$ dotplots show a distribution without gaps, (corresponding to the convex hull vertices close to the most dense part of the data cloud). By contrast, the 2^{nd} and 3^{rd} dotplots, corresponding to the two outliers taken as convex hull vertices, indicate two considerable gaps and the corresponding outlying group. We note also a large gap in the 5^{th} dotplot, although it is not so large if compared to its distribution variability. The other dotplots highlight sparse tails requiring further analysis with appropriate confirmatory tools.

Finally, it is worth noting that our method can be used as a preliminary step within a block omission approach. Indeed, classical block omission procedures waste computations, combining in the same block observations lying far from each other, although masking occurs only if outlying data lie close

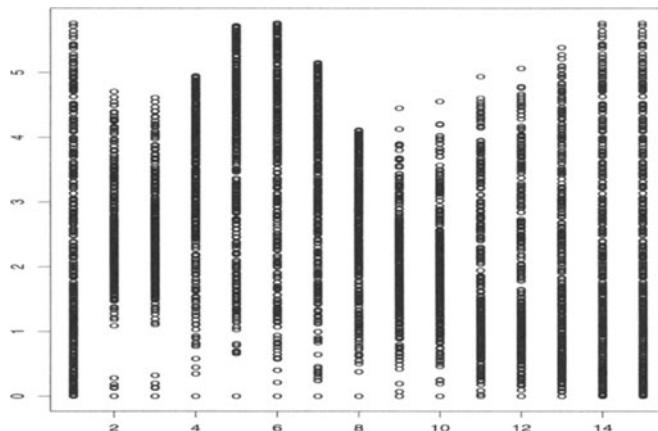


Fig. 2. Artificial data set. Dotplots of the distances of each point from the convex hull vertices.

in the space. Through our method, selecting groups of adjacent candidate outliers, block omission can be performed just within these small groups, reducing the combinatorial computational cost.

Acknowledgments: This work has been supported by MURST funds.

References

- ATKINSON, A.C. (1994): Fast Very Robust Methods for the Detection of Multiple Outliers. *Journal of the American Statistical Society*, 89, 1329–1339.
- BARNETT, V. (1976): The ordering of multivariate data (with discussion). *Journal of Royal Statistical Society A*, 139, 318–54.
- BARNETT, V. and LEWIS.T.(1994): *Outliers in Statistical Data* (3rd ed.). Wiley, New York.
- HADI, A.S. (1992): Identifying Multiple Outliers in Multivariate Data. *Journal of Royal Statistical Society, Ser.B*, 54, 761–771.
- MAHALANOBIS, P.C. (1936): On the Generalized Distance in Statistics. *Proc. Nat Inst. Sci. India A2*, 49–55.
- ROHLF, F.J. (1975): Generalization of the gap test for the detection of multivariate outliers, *Biometrics*. 31, 93–101.
- ROUSSEEUW, P.J. and van ZOMEREN, B.C. (1990): Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Society*, 85, 633–639.
- WILKS, S.S.(1963): Multivariate Statistical Outliers. *Sankhya, Ser. A*, 25, 407–426.

Discriminant Analysis Tools for Non Convex Pattern Recognition

Marcel Rémon

Département de Mathématique
Facultés Universitaires Notre-Dame de la Paix
Rempart de la Vierge, 8, B-5000 Namur (Belgium)
(e-mail: marcel.remon@fundp.ac.be)

Abstract. Estimation of non convex domains when inside and outside observations are available is often needed in current research applications. The key idea of this paper is to propose a solution based on convex and discriminant analysis tools, even when non convex domains are considered. Simulations are done and comparisons are made with a natural candidate, based on the Voronoï tessellation, for estimation of non convex bodies. However, this solution has irregularity problems.

The question of how to get smooth estimate of the unknown non convex domain is the core of this research. Our solution gives a smooth estimate of the domain and a gain of around 40 percent with respect to the symmetric difference criterion.

1 A discriminant analysis algorithm for convex bodies

Suppose that X is a Poisson point process within a fixed finite window $F \subset \mathbb{R}^d$. In F , we have a compact convex domain D . We suppose that the Poisson process X is homogeneous on F , with density λ . We observe a number $t \geq 1$ of realizations of X in F , from which n turn out to be inside the domain D and m outside D ($t = n + m$). We want to estimate the unknown convex domain D .

The solution we propose in Rémon (1996) is the modification of a classical discriminant analysis criterion developed by Baufays and Rasson (1985) to distinguish between two disjoint convex domains. The situation here is quite similar as we have two disjoint domains, D and its complementary $\sim D = F \setminus D$. The main difference lies in the non convexity of $\sim D$.

Here we note $(x, y) = \{x_1, y_1, \dots, x_{n+m}, y_{n+m}\}$ the realizations of the homogeneous Poisson process X and the labeling variable $Y : y_i = 1$ if $x_i \in D$ and $y_i = 2$ otherwise. Suppose $y_i = 1$ for $i = 1, \dots, n$ and $y_i = 2$ for $i = n+1, \dots, n+m$, without loss of generality.

Conditionally on n and m fixed, the likelihood function for (x, y) is

$$\begin{aligned} L(x, y) &= \left(\frac{1}{m(D)^n} \prod_{i=1}^n \mathbf{1}_{[x_i \in D]} \right) \left(\frac{1}{m(\sim D)^m} \prod_{i=n+1}^{n+m} \mathbf{1}_{[x_i \in \sim D]} \right) \\ &= \left(\frac{1}{m(D)} \right)^n \left(\frac{1}{m(\sim D)} \right)^m \mathbf{1}_{[H(x_1, \dots, x_n) \subseteq D]} \mathbf{1}_{[J(x_{n+1}, \dots, x_{n+m} | x_1, \dots, x_n) \subseteq \sim D]} \end{aligned}$$

where $m(C)$ is the Lebesgue measure of C , $\mathbf{1}_{[A]} = 1$ if A is true and 0 otherwise, and where $J(x_{n+1}, \dots, x_{n+m} | x_1, \dots, x_n)$ is the “shadow” statistic defined in Hatchel et al. (1981) as:

$$J(x_{n+1}, \dots, x_{n+m} | x_1, \dots, x_n) = \bigcup_{i:y_i=2} \{x_i + \lambda(x_i - b) \in \mathbb{R}^d | \lambda \geq 0, b \in H(x_1, \dots, x_n)\}.$$

$(H(\cdot), J(\cdot|\cdot))$ is a minimal sufficient statistic for the estimation of D .

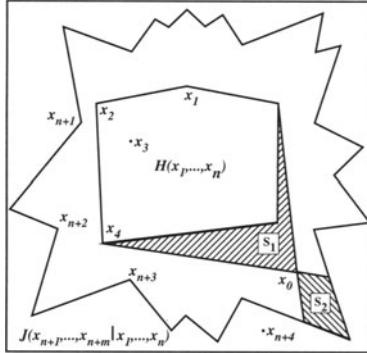


Fig. 1. Discriminant rule for convex domains

It is known that $J(x_{n+1}, \dots, x_{n+m} | x_1, \dots, x_n)$ has similar properties as the convex hull statistic $H(x_1, \dots, x_n)$. It is a consistent estimate of $\sim D$. It is robust with respect to small changes in the location of the data points and it satisfies the Equivariance requirement (see Rémon (1994) for more details).

One then gets that the following boundary of the allocating regions for D and $\sim D$ is the set of points x_0 such that:

$$\begin{aligned} & \frac{1}{m(H(x_1, \dots, x_n, x_0)) + m(J(x_{n+1}, \dots, x_{n+m} | x_1, \dots, x_n))} \\ &= \frac{1}{m(H(x_1, \dots, x_n)) + m(J(x_{n+1}, \dots, x_{n+m}, x_0 | x_1, \dots, x_n))} \end{aligned}$$

i.e.

$$S_1(x_0) = S_2(x_0)$$

where

$$S_1(x_0) \equiv m(H(x_1, \dots, x_n, x_0)) - m(H(x_1, \dots, x_n))$$

and

$$S_2(x_0) \equiv m(J(x_{n+1}, \dots, x_{n+m}, x_0 | x_1, \dots, x_n)) - m(J(x_{n+1}, \dots, x_{n+m} | x_1, \dots, x_n)).$$

This boundary gives us a practical and easily computable estimate \hat{D} for the unknown domain D (see Figure 1).

2 The Voronoï solution to the inside/outside problem for non convex bodies in \mathbb{R}^2

Let the notation be the same as before except that the unknown domain D is no longer required to be convex. The Voronoï estimation is a very natural one. It allocates any new point x_o to D if

$$\min_{1 \leq i \leq n} d(x_o, x_i) \leq \min_{n+1 \leq i \leq n+m} d(x_o, x_i)$$

where $d(.,.)$ is the Euclidean distance in \mathbb{R}^2 . The set of all these points will be the estimate \hat{D} . If one computes the Voronoï cell V_i associated to each point x_i of the training set, i.e.

$$V_i = \{x_o \in \mathbb{R}^2 : d(x_o, x_i) \leq \min_{j \neq i} d(x_o, x_j)\},$$

then $\hat{D} = \bigcup_{1 \leq i \leq n} V_i$, as showed in Figure 2.

The problem with such a solution is the non smoothness of the result. This irregularity of \hat{D} is not reduced by increasing the size of the observed training set. One could think of a smoothing algorithm applied in a second stage to the border of \hat{D} , but this leads to a “smooth estimate” incompatible with the observed data (especially when the data are numerous).

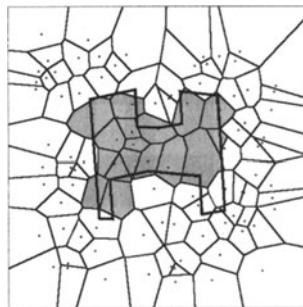


Fig. 2. Voronoï estimation for non convex domains

3 A discriminant analysis algorithm for non convex domains

Let the notation be the same as for the Voronoï estimation. The algorithm is the following.

It first labels the observed points as $x_1^1, \dots, x_{n_1}^1, x_{n_1+1}^1, \dots, x_{n_1+m_1}^1$. By convention, the first n_j points are considered as inside points while the m_j following ones are outside points. Here the algorithm takes j as 1.

In the second step, it takes the convex hull statistic of the inside points $H(x_i^j : 1 \leq i \leq n_j)$, and disregards the outside points in it. Let us denote these disregarded points $x_1^{j+1}, x_2^{j+1}, \dots, x_{n_{j+1}}^{j+1}$. With the remaining points, it performs the discriminant algorithm for convex domains, as explained in Section 1. This gives an estimating domain \hat{D}^j .

The next step consists in relabeling the inside points :

$$x_1^j, \dots, x_{n_j}^j \rightarrow x_{n_{j+1}+1}^{j+1}, \dots, x_{n_{j+1}+m_{j+1}}^{j+1} \quad \text{with} \quad m_{j+1} = n_j.$$

Then, considering $\{x_1^{j+1}, \dots, x_{n_{j+1}}^{j+1}\}$ as our new set of inside points and $\{x_{n_{j+1}+1}^{j+1}, \dots, x_{n_{j+1}+m_{j+1}}^{j+1}\}$ as our set of outside points, the algorithm goes back to the second step with $j := j + 1$. See Figure 3.

The algorithm stops as soon as no outside points can be found within the convex hull $H(x_i^j : 1 \leq i \leq n_j)$. The final solution is then :

$$\hat{D} = \hat{D}^1 \setminus (\hat{D}^2 \setminus (\hat{D}^3 \setminus (\hat{D}^4 \dots))).$$

If the algorithm requires more than three recursive calls, it divides the original figure in two, works on each half figure separately, and combines the results together. The division follows the first principal component axis of the inside points.



Fig. 3. Letter E recognition. From left to right, $H(x_1^1, \dots, x_{n_1}^1)$, $H(x_1^2, \dots, x_{n_2}^2)$ and $H(x_1^3, \dots, x_{n_3}^3)$.

4 Examples and comparison with Voronoï solution

Figure 4 and Figure 5 are two examples of letter recognitions with inside and outside points from an homogeneous Poisson point process.

We have also tested our algorithm on a chromosome picture with very low resolution : the original image is 25 by 46 pixels, with 22 different grey values. We have taken three thresholds to estimate the contours of the chromosome. The point process was no longer an homogeneous Poisson one. The point were generated from a grid of 13×26 nodes, with small random deviation to

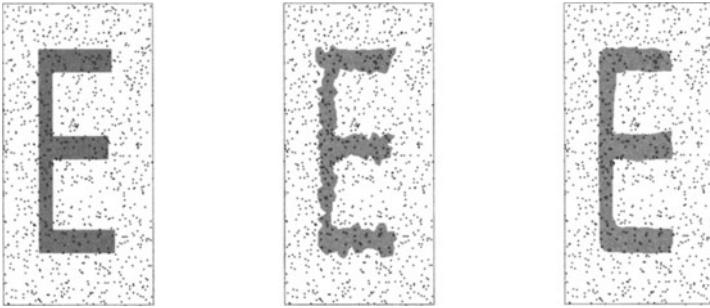


Fig. 4. Example with $n = 150$, $m = 850$ and $m(D) = 0.077$. Voronoï solution (in the center) yields $m(\hat{D}) = 0.081$ and $m(D\Delta\hat{D}) = 0.017$, while the discriminant algorithm (on the right) yields $m(\hat{D}) = 0.081$ and $m(D\Delta\hat{D}) = 0.010$. The Voronoï and discriminant algorithms use respectively 1.754 and 0.574 CPU sec.

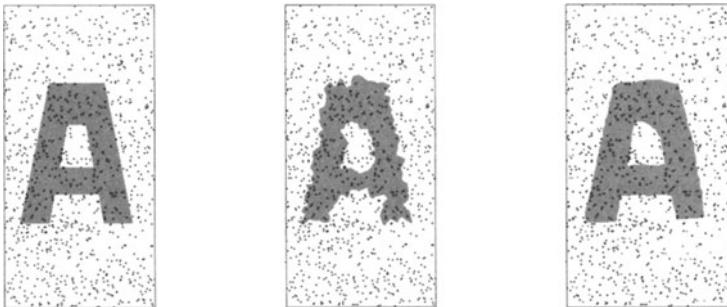


Fig. 5. Example with $n = 201$, $m = 799$ and $m(D) = 0.102$. Voronoï solution (in the center) yields $m(\hat{D}) = 0.101$ and $m(D\Delta\hat{D}) = 0.015$, while the discriminant algorithm (on the right) yields $m(\hat{D}) = 0.104$ and $m(D\Delta\hat{D}) = 0.008$. The Voronoï and discriminant algorithms use respectively 1.774 and 0.577 CPU sec.

avoid strict linearity between points (see Figure 6). In future researches, we hope to apply this algorithm to spatial non parametric density estimation.

We have done 100 simulations of an homogeneous Poisson point process over the non convex domains D shown in the previous figures. We have compared the Voronoï algorithm and the discriminant analysis algorithm in terms of the relative symmetric difference $m(D\Delta\hat{D})/m(D) = m(D \cup \hat{D} \setminus D \cap \hat{D})/m(D)$, the relative recovering measure $m(\hat{D} \cap D)/m(D)$ and the relative estimated measure $m(\hat{D})/m(D)$. The comparisons are made on exactly the same set of data (see Table 1).

Except for the last criterion, the estimated measure $m(\hat{D})$, the discriminant analysis algorithm seems to give better estimation for D . The results show a gain of more than 40 percent with respect to the symmetric difference criterion for our solution.

In terms of computing time, both algorithms have linear dependency with respect to the number of points : the Voronoï solution takes around 8.61 CPU

sec. for 500 observed points and 13.7 CPU sec. for 800 observed points, while the discriminant algorithm takes around 2.69 CPU sec. for 500 observed points and 6.94 CPU sec. for 800 observed points. These measurements were done on a Digital Workstation Dec Alpha 600 MHz.

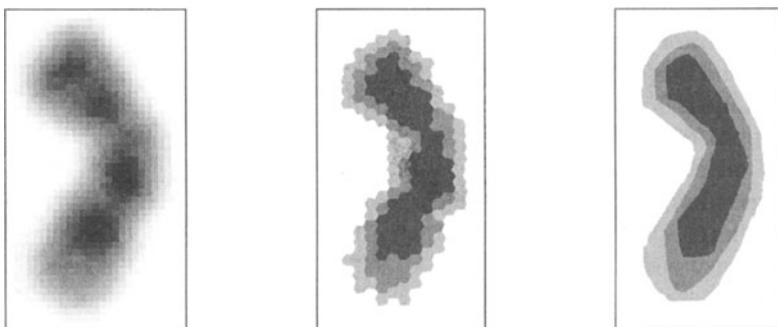


Fig. 6. Estimation of the density of a chromosome : on the left, the original image (25×46 pixels), in the center, the Voronoï estimation based on a grid (13×26) point process and on the left, the discriminant estimation based on exactly the same grid point process.

Domain D [$m(D)$]	$m(D\Delta\hat{D})/m(D)$		$m(D \cap \hat{D})/m(D)$		$m(\hat{D})/m(D)$	
	[$t = 2s/\sqrt{100}$]	Voronoi	[$t = 2s/\sqrt{100}$]	Voronoi	[$t = 2s/\sqrt{100}$]	Discr. Alg.
Method	Voronoi	Discr. Alg.	Voronoi	Discr. Alg.	Voronoi	Discr. Alg.
Letter E [0.0771]	0.2653 [0.0046]	0.1565 [0.0175]	0.8673 [0.0074]	0.9031 [0.0186]	0.9999 [0.0140]	0.9628 [0.0201]
Letter A [0.1016]	0.2015 [0.0041]	0.1217 [0.0046]	0.9007 [0.0041]	0.9367 [0.0041]	1.0029 [0.0067]	0.9952 [0.0072]

Table 1. Comparison of methods for 100 simulations : the number of generated points were $n + m = 800$ for the letter E and $n + m = 500$ for the letter A.

References

- BAUFAYS, P. and RASSON, J.-P. (1985) : A new geometric discriminant rule. *Computational Statistics Quarterly*, 2, 15–30.
- GRENANDER, U. (1976) : *Pattern Synthesis : Lectures in Pattern Theory, vol.1*, Springer Verlag, New York.
- HACHTEL, G.D., MEILIJSION, I. and NADAS, A. (1981) : The estimation of a convex subset of \mathbb{R}^k and its probability content, *IBM research report*, Yorktown Heights N.Y.
- REMON, M. (1994) : The estimation of a convex domain when inside and outside observations are available, *Supplemento ai Rendiconti del Circolo Matematico di Palermo*, 35, 227–235.
- REMON, M. (1996) : A discriminant analysis algorithm for the inside/outside problem, *Computational Statistics and Data Analysis*, 23, 125–133.

A Markovian Approach to Unsupervised Multidimensional Pattern Classification

A. Sbihi¹, A. Moussa¹, B. Benmiloud¹, and J.-G. Postaire²

¹ Université Ibn Tofail, FSK, Laboratoire Image et Reconnaissance des Formes
BP133, 14000 Kénitra, Morocco.

² Université des sciences et Technologies de Lille, Laboratoire I^3D
59655 Villeneuve d'Ascq, France.

Abstract. This paper proposes a new method for core cluster detection prior to unsupervised automatic classification. Based upon a Markov random field model, this approach transforms the set of multidimensional observations into a normalised discret binary set, which represents the observable field. The field of classes is then represented by connex components corresponding to the cores, or prototypes, inside the samples. Classification results of artificially generated data are compared with results obtained by a classical clustering method.

1 Introduction

Many statistical approaches to unsupervised automatic classification consist in detecting the modes of the underlying probability density function, p.d.f, estimated from the available patterns. Some of the existing approaches are sensitive to irregularities in distribution or/and are badly adapted to linearly non-separable distributions (Vasseur and Postaire (1980), Devijver and Kittler (1982)); while others need a great number of observations (Sbihi and Postaire (1995)). In this paper, we propose a new method which links the problem of mode detection to the Markovian model theory. After a short review of this theory (section 2), the normalised data space is transformed into a discret binary space representing the measurement field Y (section3). Section 4 exposes the proposed detection technique, from Y. of the hidden field X which represents the cores of clusters inside the samples. These cores, assimilated to regions of high concentration of observations in the data space, are the prototypes which are then used for data classification (section 5). The application of this new approach to artificially generated data. demonstrates its interest for unsupervised pattern classification (section6).

2 Theoretical basis

Let X be is a Markov random field, which satisfies the conditions:

1. $\forall x \in \Omega \quad P(X = x) > 0$
2. $P(X_s/X_r, s \neq r, r \in S) = P(X_s/X_r, r \in V_s)$

where V_s is the neighborhood of the observation X_s in the space S and $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ the set of classes.

The second condition indicates that an observation's label depends only on its neighborhood label (Gemain and Gemain (1984)). The field X is assumed to be Markovian for all the possible realisations. Thus its distribution is assimilated to a Gibbs distribution.

Let us consider a site s which is an element of S . A neighbourhood system on s is defined as any collection of subsets $G = \{G_s, s \in S\}$ such that:

1. $s \notin G_s$
2. for all s_1 and s_2 , $s_1 \in G_{s_2}$ if and only if $s_2 \in G_{s_1}$;

while a clique is defined as any subset c of S in which the neighbours of each element of c are themselves elements of c . The set of all the cliques c will be denoted by C .

For a given set S and a neighbourhood system G , a Gibbs distribution has a density function such as : $P(X = x) = \frac{1}{Z} \exp[-U(x)]$, Z is a constant independant of X , called the partition function. $U(x)$ is termed the energy function and is expressed as : $U(x) = \sum_{c \in C} V_c(x)$, where the arbitrary functions $V_c(x)$ depend only on the intensities of the sites in the clique c .

The clique potentials are defined as follows:

- For single cliques $V_c(x) = \alpha$
- For non-single cliques, $V_c(x) = \begin{cases} \beta & \text{if all } x \text{ are equal} \\ -\beta & \text{otherwise} \end{cases}$

For a clique of type m , $m=1, \dots, M$ (M is the total number of non-single type clique with non-zero parameters), the conditional probabilities are given by : $P(X_s = \omega_k / V_s) = \frac{\exp -U(\omega_k, V_s)}{\sum_{j=1}^M \exp -U(\omega_j, V_s)}$

3 Binarisation of the representation space

In order to adapt this model to clustering, we need to define a sampling lattice in the N -dimensional data space where the density function is estimated.

Let $D_q = [D_{1,q}, D_{2,q}, \dots, D_{n,q}]^T$, $q = 1, 2, \dots, Q$, be the Q available observations defined as points in \mathcal{R}^N . The transformation defined as :

$\frac{L[D_{n,q} - \min_{q'} D_{n,q'}]}{[\max_{q'} D_{n,q'} - \min_{q'} D_{n,q'}]}$, normalises the range of each component between 0 and L . Let D'_q be the new observation obtained from D_q by this transformation. Dividing each axis of the new space of data representation into L adjacent intervals of unity width, we define hypercubic lattices y_i with the corresponding function b such that :

$$b(y_i) = \begin{cases} 1 & \text{if } y_i \text{ includes at least one observation.} \\ 0 & \text{if } y_i \text{ is empty.} \end{cases}$$

Let Y denote this discrete binary set.

4 Core cluster detection

4.1 The principle of the proposed approach

The spatial structure of the data distribution set depends on the concentration of the observations within each region. Inside the clusters, each non-empty hypercube tends to have a great number of non-empty hypercubes as its nearest neighbours. On the other hand, this is not the case for non-empty hypercubes standing out of the core of clusters. To detect this disparity, a local "neighbourhood" analysis based upon Markov random fields is proposed. The binary estimated set from the available observations is considered as the field Y which models the field of measures (Markovian language), from which we try to extract the hidden field X representing the cores of clusters.

Let us consider Y as the initial state of the class field X . Knowing the geometrical shape of the neighbourhood and the clique potential energy, we can compute explicitly the conditional probability. The approach is described by an iterative algorithm, similar to the Gibbs algorithm, which eliminates progressively the isolated non-empty hypercubes in order to detect the field of class.

A stop criterion of the procedure is based on the computation of the global energy function $U(k)$ during the iteration process. In fact, experimental results shows that $U(k)$ remains constant after some iterations. Consequently the exchange rate of sites between successive iterations is stabilised. Thus, we may apply a stabilisation criterion to stop the core detection procedure.

4.2 Core cluster detection algorithm (CCDA)

1. Initialisations : $(t, \Delta^t, E^t, \epsilon, k)$;
2. While $(\Delta^t > \epsilon) : t = t + 1$;
 - For each observation;
 - Compute the energy of each configuration U_k^i and the conditional probability P_k^i ;
 - Classify according to P_k^i (*if* $P_0^i > P_1^i \Rightarrow X_i = 0$ *else* $X_i = 1$);
 - Compute the global energy $E^t = \sum_{i=k}^n \sum_{k=0}^1 U_k^i$ and the difference $\Delta^t = E^t - E^{t-1}$;
 - For each observation, reinitialise Y_i to $X_i (Y_i = X_i)$;
3. End ;

The behaviour of the procedure depends on the adjustment of the resolution parameter L used in the discretisation process. If L is too small, the procedure will suffer from too little resolution and small clusters will not be detected. If L is too large, the procedure will tend to find a great number of non-significant clusters. However, it can be expected that when true clusters exist, stable connected subsets corresponding to these clusters will appear for a wide range of values of L . Based on this assumption, the adjustment of L is governed

by the concept of cluster stability. Choosing such a parameter in the middle of the largest range when the number of detected clusters remains constant has proved to be a good procedure to optimise a number of classification algorithms (Postaire and Vasseur (1983)).

5 Classification

Once the different cores are identified by CCDA algorithm, many grouping procedures can be used to assign the input data points to the clusters attached to them. One approach is to use the data points falling into the cores. The remaining data points, which do not fall into one of the cores, are finally assigned to the clusters attached to their nearest (Euclidean) neighbour among these cores (Cover and Hart (1967)).

6 Experimental results

The efficiency of the proposed algorithm for core cluster detection has been demonstrated using artificially generated data sets.

6.1 Example 1

The raw data set, shown in figure 1-a, consists of observations drawn from three equiprobable gaussian clusters with the statistical parameter given in table-1. The estimate of the binary field Y is illustrated in figure 1-b with $L = 40$, while figure 1-c shows the core clusters detected by the CCDA.

The classification results are displayed in figure 1-d. The statistical parameters of these results are also given in table-1 . The error rate achieved by the procedure is equal to 2.58%, whereas it is equal to 2.66%, with the classical Isodata algorithm.

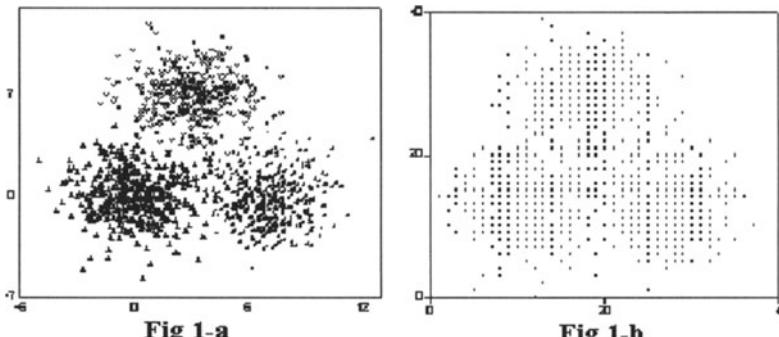
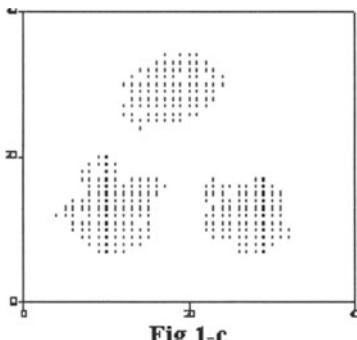
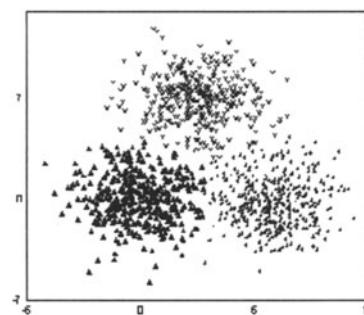


Fig 1-a

Fig 1-b

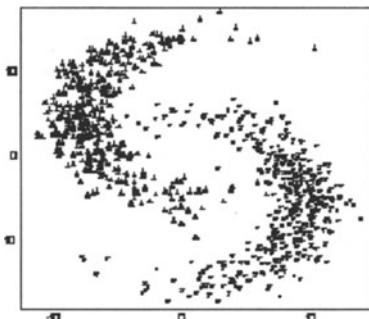
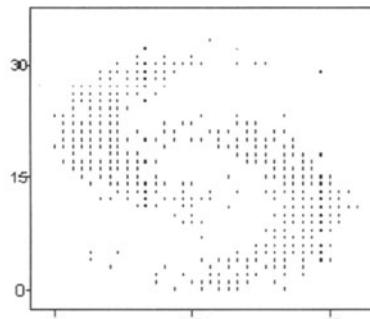
**Fig 1-c****Fig 1-d**

Generated Data				Classification Results		
Clusters	Mean Vect	Var Matrices	P.Proba	Mean Vect	Var Matrices	P.Proba
1	0.002403	2.96481	0.33	-0.100301	2.57057	0.327500
	0.300105	2.92763		0.045900	2.80975	
2	3.02523	2.96424	0.33	3.04746	2.98565	0.334167
	6.87726	2.64326		6.88555	2.568501	
3	7.07959	2.65041	0.33	7.04028	2.66426	0.338333
	-0.16626	3.02594		-0.205679	2.9684	

Table 1.

6.2 Example 2

This example is composed of two non-gaussian clusters distributed as shown in figure 2-a. The binary field Y estimated with $L = 34$ is displayed in figure 2-b . While figure 2-c shows the cluster cores detected by the CCDA.

**Fig 2-a****Fig 2-b**

The classification results, displayed in figure 2-d, show the ability of the proposed procedure to detect cores of clusters in the case of non-linearly separable data. The CCDA algorithm takes into account the geometrical

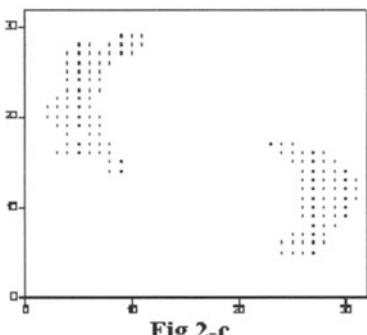


Fig 2-c

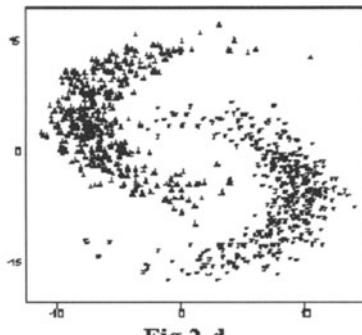


Fig 2-d

shapes of the original distribution of the data. The error rate obtained is equal to 1.36%, while the Isodata error rate is equal to 9.38%.

Note that some examples have proved the ability of this procedure to detect core clusters from distributions of higher dimensionally.

7 Conclusion

By analysing the approach presented in this paper, we come to the conclusion that a Markovian model is very efficient in non-trivial clustering situations such as linearly non-separable clusters, particularly when no a priori information is available. It therefore appears that the Markovian model has found a new field of application in the area of cluster analysis. As a natural extension of this work, the authors are now concentrating their effort on the development of other Markovian tools to pattern classification.

References

- T.M.COVER and P.E.HART (1967): Nearest Neighbor Pattern Classification. *IEEE. Trans.Inf.Theory*, Vol IT-13, N1, 21–27.
- P.A.DEVIJER and J.KITTLER (1982): Pattern Recognition. A statistical Approach. Englwood Cliff, NJ, Prentice-Hall International, 448.
- S. GEMAIN and P. GEMAIN (1984): Stochastic Relaxation, Gibbs distribution and the Baysien Restoration of Images. *IEEE Trans on pattern anal. and Machine Intell*, PAMI-6, 721–741.
- J. G. POSTAIRE and C. P. VASSEUR (1983): A Fast Algorithm for Non Parametric Probability Density Estimation. *IEEE Trans on Pattern Anal. Machine Intell*, PAMI-4 n6, 663–666.
- A. SBIHI and J. G. POSTAIRE (1995): Mode Extraction by Multivalue Morphology for Cluster Analysis. In W.Gaul and D.Pfeifer (eds) :*From DATA to Knowledge: Theoretical and Practical of aspect of Classification*. Springer, Berlin, 212–221.
- C.P.A. VASSEUR and J. G. POSTAIRE (1980): A Convexity Testing Method for Cluster Analysis. *IEEE Trans. Syst. Man. Cybern* , SMC-10, (3), 145–149.

Part III

Multivariate and Multidimensional Data Analysis

An Algorithm with Projection Pursuit for Sliced Inverse Regression Model

Masahiro Mizuta and Hiroyuki Minami

Center for Information and Multimedia Studies,
Hokkaido University, N11W5, Sapporo 060-0811, JAPAN
(e-mail: mizuta@main.eng.hokudai.ac.jp)
(e-mail: min@main.eng.hokudai.ac.jp)

Abstract. In the paper, we investigate a conditional density function of sliced response variables and propose an algorithm for the sliced inverse regression (SIR) model with projection pursuit.

The SIR model is a general model for dimension reduction of explanatory variables on regression analysis. Some algorithms for SIR model are proposed; SIR, SIR2, Bivariate SIR. We apply the algorithms to some typical data sets. They can not find suitable reductions for all of the data sets. The proposed algorithm can get reasonable results for all of them.

1 Introduction

Regression analysis is one of the fundamental methods for data analysis. A response variable y is estimated as a function of explanatory variables \mathbf{x} (a p -dimensional vector). An immediate goal of ordinary regression analysis is to find the function of \mathbf{x} . When there are many explanatory variables in the data set, it is difficult to calculate the regression coefficients stably. An approach to reduce the number of explanatory variables is an explanatory variable selection and there are many studies on the variable selection. Another approach is to project explanatory variables on a lower dimensional space that almost estimates the response variable.

Sliced Inverse Regression (SIR) proposed by Li (1991) is one of the methods to reduce explanatory variables with linear projection. SIR finds linear combinations of explanatory variables that are reductions for nonlinear regression. But the original SIR algorithm can not derive reasonable results for some artificial data that have trivial structures. Li also developed another algorithm: SIR2, which uses a conditional estimation $E[\text{cov}(\mathbf{x}|y)]$. However, SIR2 can not find trivial structures for another type of data, either.

We aim at a usage of projection pursuit for finding the linear combinations of explanatory variables. A new SIR method with projection pursuit (SIRpp) is proposed in the paper. We also present numerical examples of the proposed method.

2 Sliced Inverse Regression model

Sliced Inverse Regression is one of the approaches to reduce the number of explanatory variables in regression analysis. SIR does not get rid of some explanatory variables themselves but may reduce the dimension of a space of explanatory variables. It is based on the model (SIR model)

$$y = f(\beta_1 \mathbf{x}, \beta_2 \mathbf{x}, \dots, \beta_K \mathbf{x}, \varepsilon), \quad (1)$$

where \mathbf{x} is the vector of p explanatory variables, β_k ($k = 1, 2, \dots, K$) are unknown row vectors, ε is independent of \mathbf{x} , and f is an arbitrary unknown function on \mathbf{R}^{K+1} .

The purpose of SIR is to estimate the vectors β_k when this model holds. If we get the β_k , we can reduce the dimension of \mathbf{x} to K . Hereafter, we shall refer to any linear combination of β_k as effective dimensional reduction (e.d.r.) direction.

Li (1991) proposed an algorithm to find e.d.r. directions and named SIR. However, we call the algorithm SIR1 in order to distinguish it from the SIR model.

The main idea of SIR1 is to use $E[\mathbf{X}|y]$. $E[\mathbf{X}|y]$ is contained in the space spanned by e.d.r. directions, but there is no guarantee that $E[\mathbf{X}|y]$ span the space. For example in Li, if $(X_1, X_2) \sim N(0, I_2)$, and $Y = X_1^2$ then $E[X_1|y] = E[X_2|y] = 0$.

3 The SIR model and nonnormality of the conditional distributions

Because $E[\mathbf{X}|y]$ is not sufficient to find e.d.r. directions, we will investigate the conditional distributions $\mathbf{X}|y$ themselves. We assume that the distribution of \mathbf{x} is standard normal hereafter: $\mathbf{x} \sim N(0, I_p)$. If not, we standardize \mathbf{x} with an affine transformation. Furthermore, $\beta_i \beta_j^T = \delta_{ij}$ ($i, j = 1, 2, \dots, K$) is supposed without loss of generality. We can choose β_i ($i = K + 1, \dots, p$) such that $\{\beta_i\}$ ($i = 1, 2, \dots, p$) is an orthonormalized basis for \mathbf{R}^p .

Because the distribution of \mathbf{x} is $N(0, I_p)$, the distribution of $(\beta_1 \mathbf{x}, \dots, \beta_p \mathbf{x})$ is also $N(0, I_p)$.

It is easy to derive the density function of $(\beta_1 \mathbf{x}, \dots, \beta_p \mathbf{x}, y)$;

$$h(\beta_1 \mathbf{x}, \dots, \beta_p \mathbf{x}, y) = \phi(\beta_1 \mathbf{x}) \cdots \phi(\beta_p \mathbf{x}) \psi(\beta_1 \mathbf{x}, \dots, \beta_K \mathbf{x}, y),$$

where $\phi(x) = 1/\sqrt{2\pi} \exp(-x^2/2)$ and $\psi(\cdot)$ is a function of $\beta_1 \mathbf{x}, \dots, \beta_K \mathbf{x}, y$.

The conditional density function is

$$h(\beta_1 \mathbf{x}, \dots, \beta_p \mathbf{x} | y) = \phi(\beta_{K+1} \mathbf{x}) \cdots \phi(\beta_p \mathbf{x}) g(\beta_1 \mathbf{x}, \dots, \beta_K \mathbf{x}),$$

where $g(\cdot)$ is a function of $\beta_1 \mathbf{x}, \dots, \beta_K \mathbf{x}$ and is not generally normal density function. So, $h(\beta_1 \mathbf{x}, \dots, \beta_p \mathbf{x} | y)$ is separated into the normal distribution part $\phi(\beta_{K+1} \mathbf{x}) \cdots \phi(\beta_p \mathbf{x})$ and the nonnormal distribution part $g(\cdot)$.

Projection Pursuit is a good method to find out nonlinear structure. Specifically, Friedman (1987) regarded the nonlinear structure as nonnormality. We will apply projection pursuit to SIR in the next section in order to find out the nonnormal distribution part.

4 SIRpp algorithm

We propose an algorithm for the SIR model with projection pursuit (SIRpp). The proposed algorithm for data (y_i, \mathbf{x}_i) ($i = 1, 2, \dots, n$) as follows:

1. Standardize \mathbf{x} : $\tilde{\mathbf{x}}_i = \hat{\Sigma}_{\mathbf{xx}}^{-\frac{1}{2}}(\mathbf{x}_i - \bar{\mathbf{x}})$ ($i = 1, 2, \dots, n$), where $\hat{\Sigma}_{\mathbf{xx}}$ is the sample covariance matrix and $\bar{\mathbf{x}}$ is the sample mean of \mathbf{x} .
2. Divide range of y into H slices, I_1, I_2, \dots, I_H .
3. Conduct a projection pursuit in K dimensional space for each slice.
We get H projections: $(\boldsymbol{\alpha}_1^{(h)}, \dots, \boldsymbol{\alpha}_K^{(h)})$, ($h = 1, 2, \dots, H$).
4. Let the K largest eigenvectors of \hat{V} be $\hat{\boldsymbol{\eta}}_k$ ($k = 1, 2, \dots, K$).

Output $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\eta}}_k \Sigma_{\mathbf{xx}}^{-\frac{1}{2}}$ ($k = 1, 2, \dots, K$) for estimations of e.d.r. directions, where $\hat{V} = \sum_{h=1}^H w(h) \sum_{k=1}^K \boldsymbol{\alpha}_k^{(h)T} \boldsymbol{\alpha}_k^{(h)}$ and $w(h)$ is a weight determined by the size of the slice and the projection pursuit index.

Steps 1 and 2 are the same as those of SIR1. The data is spherized in Step 1 and is sliced in Step 2. The H projections in Step 3 are regarded as e.d.r. directions on the coordinates of $\tilde{\mathbf{x}}$. We get H projections and combine them into \hat{V} in Step 4; this is similar to a singular value decomposition.

5 Numerical examples

We use two models of multicomponents:

$$y = x_1(x_1 + x_2 + 1) + \sigma \cdot \varepsilon \quad (2)$$

$$y = \sin(x_1) + \cos(x_2) + \sigma \cdot \varepsilon \quad (3)$$

to generate data $n = 400$, where $\sigma = 0.5$. At first, we generate x_1, x_2, ε with $N(0,1)$ and calculate a response variable y with (2) or (3). Eight variables x_3, \dots, x_{10} generated by $N(0,1)$ are added to explanatory variables.

The ideal e.d.r. directions are contained in the space spanned by two vectors $(1, 0, \dots, 0)$ and $(0, 1, \dots, 0)$.

The squared multiple correlation coefficient between the projected variable \mathbf{bx} and the space B spanned by ideal e.d.r. directions:

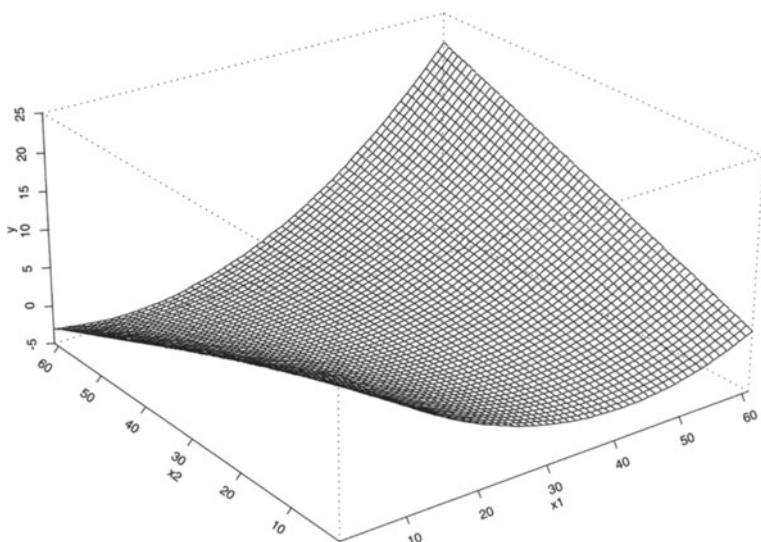
$$R^2(\mathbf{b}) = \max_{\boldsymbol{\beta} \in B} \frac{(\mathbf{b} \sum_{\mathbf{xx}} \boldsymbol{\beta}')^2}{\mathbf{b} \sum_{\mathbf{xx}} \mathbf{b}' \cdot \boldsymbol{\beta} \sum_{\mathbf{xx}} \boldsymbol{\beta}'} \quad (4)$$

is adopted as the criterion to evaluate the effectiveness of estimated e.d.r. directions. Tab.1 shows the mean and the standard deviation (in parentheses)

of $R^2(\hat{\beta}_1)$ and $R^2(\hat{\beta}_2)$ of four SIR algorithms for $H = 5, 10$, and 20 , after 100 replicates.

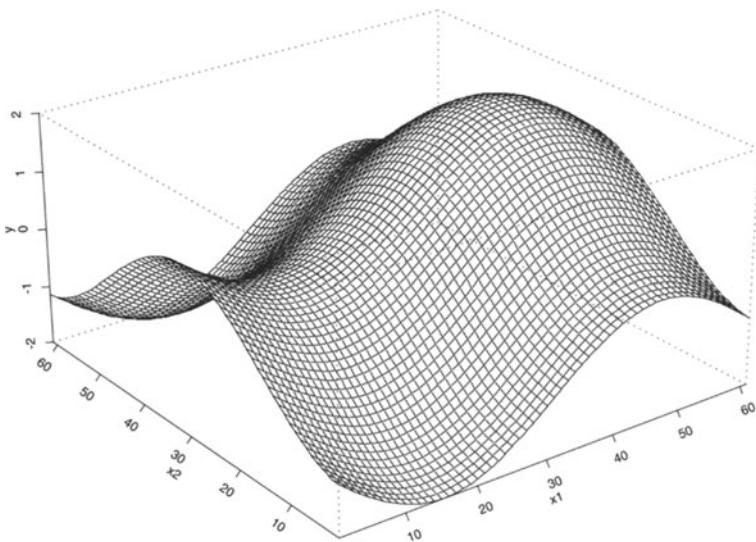
SIR2 can not reduce the explanatory variables of the example 1. The result of the second example is very interesting. SIR1 and BiSIR find the asymmetric e.d.r. direction and do not find the symmetric e.d.r. direction. Conversely, SIR2 finds only the symmetric e.d.r. direction. SIRpp succeeds in detecting both e.d.r. directions.

Tab.1.: Results of SIR1, SIR2, BiSIR, and SIRpp (Example 1)
Asymmetric function $y = x_1(x_1 + x_2 + 1) + \sigma \cdot \varepsilon$



	SIR1		SIR2		BiSIR		SIRpp	
	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$
5	.92 (.04)	.77 (.11)	.96 (.03)	.20 (.21)	.94 (.03)	.69 (.17)	.97 (.02)	.78 (.15)
10	.93 (.03)	.81 (.09)	.92 (.09)	.10 (.12)	.93 (.04)	.73 (.15)	.95 (.04)	.79 (.13)
20	.92 (.04)	.76 (.18)	.83 (.19)	.11 (.13)	.92 (.04)	.73 (.14)	.95 (.07)	.75 (.18)

Tab.2.: Results of SIR1, SIR2, BiSIR, and SIRpp (Example 2)
Function $y = \sin(x_1) + \cos(x_2) + \sigma \cdot \varepsilon$ is asymmetric with respect to the x_1 axis and symmetric with respect to x_2 axis.



H	SIR1		SIR2		BiSIR		SIRpp	
	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$	$R^2(\hat{\beta}_1)$	$R^2(\hat{\beta}_2)$
5	.97 (.02)	.12 (.14)	.92 (.04)	.01 (.10)	.95 (.02)	.34 (.28)	.92 (.05)	.88 (.11)
10	.97 (.02)	.12 (.15)	.90 (.06)	.05 (.07)	.95 (.02)	.43 (.28)	.88 (.08)	.84 (.13)
20	.97 (.02)	.12 (.14)	.85 (.09)	.05 (.06)	.95 (.02)	.47 (.26)	.84 (.10)	.73 (.22)

6 Concluding remarks

The SIRpp algorithm has excellent performance in finding e.d.r. directions. However, the algorithm requires more computing power. It is well known that the projection pursuit algorithm takes much time. We intend to improve SIRpp algorithm especially on projection pursuit method.

References

- FRIEDMAN, J. H. (1987): Exploratory Projection Pursuit. *Journal of the American Statistical Association*, 82, 249–266.

- FRIEDMAN, J. H. & TUKEY, J. W. (1974): A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Trans. on Computer*, c-23, 9, 881–890.
- KOYAMA, K., MORITA, A., MIZUTA, M., and SATO, Y. (1998): Projection Pursuit into Three Dimensional Space.(in Japanese) *The Japanese Journal of Behaviormetrics*, 25(1), 1–9.
- LEE, Y., LEE, D. and CHOI, K.(1998): Bivariate Sliced Inverse Regression and its Application. *Data Science, Classification, and Related Methods*, ISTAT, 198–201.
- LI, KER-CHAU (1991): Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86, 316–342.
- MIZUTA, M. (1998): A New Algorithm for Sliced Inverse Regression with Projection Pursuit. (in Japanese) *Proceedings of Japan Statistical Society 1998* , 158–159.
- MIZUTA, M. (1999): Sliced Inverse Regression with Projection Pursuit., In: H. Bacelar-Nicolau, F. Costa Nicolau and J. Janssen (Eds.): *Applied Stochastic Models and Data Analysis*. INSTITUTO NACIONAL DE ESTATÍSTICA), 51–56.
- MIZUTA, M. (1999): Projection Pursuit into High Dimensional Space and its Applications. *Bulletin of the International Statistical Institute*, 52nd Session, 313–314.

Testing Constraints and Misspecification in VAR-ARCH Models

Wolfgang Polasek and Shuangzhe Liu

Institute of Statistics and Econometrics, University of Basel,
Holbeinstrasse 12, CH-4051 Basel, Switzerland
(e-mail: liu@iso.iso.unibas.ch)

Abstract. Vector autoregressive models with conditional heteroskedastic errors (abbreviated as VAR-ARCH models) have become increasingly important for applications in financial econometrics. In this paper, we propose likelihood ratio and Wald tests for constraints and the White (1982) misspecification test for VAR-ARCH models which are estimated by the maximum likelihood (ML) method. The tests are discussed for a general class of multivariate conditional heteroskedastic time series models including the VAR-ARCH models. We derive the exact analytic expression for the gradient vector and the conditional information matrix from the log-likelihood function under the normality assumption.

1 Introduction

ARCH models were first introduced by Engle (1982) and currently VAR-ARCH models are applied widely in financial econometrics. Lagrange multiplier and misspecification tests in the framework of the univariate ARCH models can be found in, e.g. Mills (1994, Section 5.3) or Gouriéroux (1997, Chapter 4). Software packages like MathSoft's (1996) S+GARCH use the method of BHHH, suggested by Berndt et al. (1974), for ML estimation based on the gradient vector (of first order derivatives). Such ML estimates may be useful for a Lagrange multiplier test. In the present paper we use the ML estimates, which are obtained by the method of scoring for maximum likelihood estimation based on not only the gradient vector, but also the (conditional) information matrix (associated with second order derivatives) under the normality assumption as in Liu and Polasek (1999), to develop likelihood ratio, Wald and misspecification tests for VAR-ARCH models. The plan of the paper is as follows: Section 2 reviews estimation results of the ML approach and Section 3 describes the likelihood ratio test, the Wald test and the White misspecification test. Section 4 concludes.

2 The VAR(k)-ARCH(q) model

Consider the following M dimensional VAR(k)-ARCH(q) model

$$y_t = \mu_t + u_t, \quad t = 1, \dots, T, \tag{1}$$

where $y_t = (y_{1t}, y_{2t}, \dots, y_{Mt})'$ is an $M \times 1$ vector of time series, u_t is an $M \times 1$ vector of error terms of a normal distribution with conditional mean $E(u_t|\psi_{t-1}) = 0$ with ψ_{t-1} indicating the information set until time $t-1$, $E(y_t|\psi_{t-1}) = \mu_t$, μ_t is the $M \times 1$ mean vector which follows a VAR process of order k (the mean equation)

$$\mu_t = b_0 + B_1 y_{t-1} + \dots + B_k y_{t-k}, \quad (2)$$

where b_0 is an $M \times 1$ vector and each B_i , $i = 1, \dots, k$, is an $M \times M$ matrix containing unknown AR parameters. The conditional variance of u_t is $V_t = E(u_t u_t' |\psi_{t-1})$ and is parameterized as (the variance equation)

$$\text{vech}V_t = a_0 + A_1 \text{vech}u_{t-1}u_{t-1}' + \dots + A_q \text{vech}u_{t-q}u_{t-q}', \quad (3)$$

where $\text{vech}V_t$ eliminates all supradiagonal elements of the matrix V_t and yields an $N \times 1$ vector, $N = M(M+1)/2$,

$$\text{vech}u_{t-j}u_{t-j}' = (u_{1t-j}^2, u_{2t-j}u_{1t-j}, u_{3t-j}u_{1t-j}, \dots, u_{Mt-j}^2)', \quad j = 1, \dots, q, \quad (4)$$

is an $N \times 1$ vector, a_0 is an $N \times 1$ vector of constants and A_j , $j = 1, \dots, q$, is an $N \times N$ matrix of unknown ARCH parameters such that $V_t > 0$ is positive definite.

We use the notation

$$\begin{aligned} \mu_t &= (x_t' \otimes I_M)\beta, \\ x_t &= (1, y_{t-1}', \dots, y_{t-k}')', \\ \beta &= \text{vec}(b_0, B_1, \dots, B_k), \\ \text{vech}V_t &= (z_t' \otimes I_N)\alpha, \\ z_t &= (1, \text{vech}'u_{t-1}u_{t-1}', \dots, \text{vech}'u_{t-q}u_{t-q}')', \\ \alpha &= \text{vec}(a_0, A_1, \dots, A_q), \\ \theta &= (\beta', \alpha')', \end{aligned}$$

where x_t is a $(1 + Mk) \times 1$ vector, z_t is a $(1 + Nq) \times 1$ vector, β is an $M(1 + Mk) \times 1$ vector of regression coefficients for the mean equation and α is an $N(1 + Nq) \times 1$ parameter vector for the variance equation.

The log-likelihood function under normality is defined as

$$L(\theta) = \sum_{t=1}^T L_t(\theta), \quad (5)$$

where

$$L_t(\theta) = -\frac{1}{2} \log V_t - \frac{1}{2} u_t' V_t^{-1} u_t.$$

The information matrix $F = F(\theta)$ can be given in two ways using the gradient vector $g(\theta) = \sum_{t=1}^T g_t(\theta)$ and the conditional information matrix $P_t = P_t(\theta)$

on ψ_{t-1} , respectively,

$$\mathbb{E}(gg') = F = \sum_{t=1}^T \mathbb{E}(P_t), \quad (6)$$

where

$$\begin{aligned} g_t(\theta) &= \partial L_t(\theta)/\partial\theta, \\ H_t(\theta) &= \partial L_t^2(\theta)/\partial\theta\partial\theta', \\ P_t(\theta) &= -\mathbb{E}(H_t|\psi_{t-1}). \end{aligned}$$

For relevant ideas and results, see Liu and Polasek (1999).

We partition the gradient vector $g = g(\theta)$ according to $\theta = (\beta', \alpha')'$ and use

$$\hat{g} = \sum_{t=1}^T g_t(\hat{\theta}), \quad (7)$$

where

$$\begin{aligned} g_t &= \begin{pmatrix} g_{\beta t} \\ g_{\alpha t} \end{pmatrix}, \\ g_{\beta t} &= \frac{1}{2}(\partial \text{vech} V_t / \partial \beta')' q_t + (\partial \mu_t / \partial \beta')' V_t^{-1} u_t \\ &= -Z_t' q_t + x_t \otimes V_t^{-1} u_t, \\ g_{\alpha t} &= \frac{1}{2}(\partial \text{vech} V_t / \partial \alpha')' q_t + (\partial \mu_t / \partial \alpha')' V_t^{-1} u_t \\ &= \frac{1}{2} z_t \otimes q_t, \\ Z_t &= \sum_{i=1}^q A_i D^+ (u_{t-i} x_{t-i}' \otimes I_M), \\ q_t &= D' D \text{vech}(V_t^{-1} u_t u_t' V_t^{-1} - V_t^{-1}), \end{aligned}$$

and where $g_{\beta t}$ is an $M(1 + Mk) \times 1$ vector, $g_{\alpha t}$ is an $N(1 + Nq) \times 1$ vector, Z_t is an $N \times M^2$ matrix, q_t is an $N \times 1$ vector, D is the $M^2 \times N$ duplication matrix such that $\text{vec} V = D \text{vech} V$, $\text{vec} V$ is the $M^2 \times 1$ vector transformed from V by stacking the columns of V one underneath the other and D^+ is the Moore-Penrose inverse of D , see Magnus and Neudecker (1999, Section 3.8).

We obtain the estimate of the information matrix based on the conditional information matrix evaluated at the ML location:

$$\hat{F}(\theta) = \sum_{t=1}^T P_t(\hat{\theta}), \quad (8)$$

where $P_t = P_t(\theta)$ is partitioned

$$\begin{aligned} P_t &= \begin{pmatrix} P_{\beta\beta t} & P_{\beta\alpha t} \\ P'_{\beta\alpha t} & P_{\alpha\alpha t} \end{pmatrix}, \\ P_{\beta\beta t} &= \frac{1}{2}(\partial \text{vech} V_t / \partial \beta')' W_t \partial \text{vech} V_t / \partial \beta' + (\partial \mu_t / \partial \beta')' V_t^{-1} \partial \mu_t / \partial \beta' \\ &= 2Z_t' W_t Z_t + x_t x_t' \otimes V_t^{-1}, \\ P_{\beta\alpha t} &= \frac{1}{2}(\partial \text{vech} V_t / \partial \beta')' W_t \partial \text{vech} V_t / \partial \alpha' + (\partial \mu_t / \partial \beta')' V_t^{-1} \partial \mu_t / \partial \alpha' \\ &= -Z_t' W_t (z_t' \otimes I_N), \\ P_{\alpha\beta t} &= P'_{\beta\alpha}, \\ P_{\alpha\alpha t} &= \frac{1}{2}(\partial \text{vech} V_t / \partial \alpha')' W_t \partial \text{vech} V_t / \partial \alpha' + (\partial \mu_t / \partial \alpha')' V_t^{-1} \partial \mu_t / \partial \alpha' \\ &= \frac{1}{2}(z_t \otimes I_N) W_t (z_t' \otimes I_N), \\ W_t &= D'(V_t^{-1} \otimes V_t^{-1}) D, \end{aligned}$$

$P_{\beta\beta t}$ is of order $M(1+Mk) \times M(1+Mk)$, $P_{\beta\alpha t}$ of $M(1+Mk) \times N(1+Nq)$, $P_{\alpha\alpha t}$ of $M(1+Mk) \times N(1+Nq)$, and W_t of $N \times N$.

The parameter estimates of $\theta = (\beta', \alpha')'$ are numerically obtained by iteration as in, e.g. Fomby et al. (1984, pp. 610-615):

$$\theta_{j+1} = \theta_j + \lambda_j F_j^{-1} g_j, \quad j = 1, 2, \dots, \quad (9)$$

where $\lambda_j \leq 1$ is a constant to improve convergence, j is the iteration index, g is given in (7) and F is in (8).

3 Testing procedure

3.1 Likelihood ratio test

Assume that the ML estimates $\hat{\theta} = (\hat{\beta}', \hat{\alpha}')'$ can be calculated for the full model and the restricted ML estimates $\tilde{\theta} = (\tilde{\beta}', \alpha_0')'$ for the model with the null hypothesis $H_0: \alpha = \alpha_0$, where α_0 consists only of the coefficients in the variance equation. Specifically, we are interested in $\alpha_0 = (a_0', a_1')'$, where a_0 is the vector of constants and $a_1 = \text{vec}(A_1, \dots, A_q) = 0$ with A_1, \dots, A_q being ARCH parameters. Then the likelihood ratio test statistic can be derived from the log-likelihood function (5) as

$$LR = 2[L(\hat{\beta}, \hat{\alpha}) - L(\tilde{\beta}, \alpha_0)] \sim \chi^2(r), \quad (10)$$

where LR follows asymptotically a chi-squared distribution with r degrees of freedom under H_0 and r is the dimension of α .

3.2 Wald test

Under certain regularity conditions, the distribution of the ML estimates $\hat{\theta} = (\hat{\beta}', \hat{\alpha}')'$ is approximately normal, see Gouriéroux (1997, Section 6.3.2). We then partition F according to P_t . We find under $H_0: \alpha = \alpha_0$ that the block submatrix, associated with α , of the asymptotic variance matrix is

$$(F_{\alpha\alpha} - F_{\alpha\beta}F_{\beta\beta}^{-1}F'_{\alpha\beta})^{-1}. \quad (11)$$

Similar to Gouriéroux (1997, Section 6.3.2), we may propose (8) with $\hat{\theta} = (\hat{\beta}', \hat{\alpha}')'$ as a consistent estimator of F . With such ML estimates, we can calculate the quadratic form and the Wald test statistic

$$W_\alpha = (\hat{\alpha} - \alpha_0)'(\hat{F}_{\alpha\alpha} - \hat{F}_{\alpha\beta}\hat{F}_{\beta\beta}^{-1}\hat{F}'_{\alpha\beta})(\hat{\alpha} - \alpha_0) \sim \chi^2(r), \quad (12)$$

where W_α follows asymptotically $\chi^2(r)$ and r is the dimension of α .

3.3 Misspecification test

Note that the White (1982, 1988) misspecification test is a test of the equality of the estimated information matrices based on the key relationship in (6) involving the gradient vector and the information matrix, respectively. We can then construct a misspecification test to test the null hypothesis $H_0: C'\text{vech}(F + gg') = 0$ against $H_1: C'\text{vech}(F + gg') \neq 0$, where C is a $m(m+1)/2 \times r$ selection matrix so as to test certain subsets of parameters (for example, AR parameters or ARCH parameters), m is the dimension of $\theta = (\beta', \alpha')$ and r is the rank of C . One advantage of this test is to use P_t in addition to g if P_t is available in closed form.

The test is computed via an artificial regression in which the constant unity 1 is regressed on the variables: \hat{g} and $C'\text{vech}(\hat{F} + \hat{g}\hat{g}')$ which can be obtained based on (7) and (8). Using the (adjusted) squared multiple correlation coefficient (determination ratio), the test statistic $TR^2 \sim \chi^2(r)$ follows a chi-squared distribution with r degrees of freedom approximately in large samples under H_0 .

4 Remarks

Based on the method of scoring, we present the exact ML estimates with estimated gradient \hat{g} and information matrix \hat{F} for the VAR-ARCH model in Section 2. This approach allows to apply the likelihood ratio, Wald and general misspecification tests as discussed in Section 3. Implementing the White misspecification test to the VAR-ARCH model implies the following

conclusion: If the null hypothesis is rejected, then there is statistical evidence that the maximum likelihood estimator for the model under investigation is inefficient and possibly inconsistent. The possible inconsistency can be further examined by the Hausman (1978) test for misspecification or Newey's (1985) conditional moment test.

References

- BERNDT, E., HALL, B., HALL, R. and HAUSMAN, J. (1974): Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement*, 3, 653-665.
- ENGLE, R.F. (1982): Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50, 987-1006.
- FOMBY, T.B., HILL, R.C. and JOHNSON, S.R. (1984): *Advanced Econometric Methods*. Springer-Verlag, New York.
- GOURIEROUX, C. (1997): *ARCH Models and Financial Applications*. Springer-Verlag, New York.
- HAUSMAN, J.A. (1978): Specification tests in econometrics, *Econometrica*, 46, 1251-1272.
- LIU, S. and POLASEK, W. (1999): Maximum likelihood estimation for the VAR-VARCH model: A new approach. In: U. Leopold-Wildburger, G. Feichtinger, and K.-P. Kistner (Eds.): *Modelling and Decisions in Economics, Essays in Honor of Franz Ferschl*. Physica-Verlag, Heidelberg, 99-113.
- MAGNUS, J.R. and NEUDECKER, H. (1999): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, revised edition. John Wiley and Sons, Chichester, UK.
- MATHSOFT (1996): *S+GARCH User's Manual, Version 1.0*. Data Analysis Products Division, MathSoft, Seattle, USA.
- MILLS, T.C. (1994): *The Econometric Modelling of Financial Time Series*. Cambridge University Press, Cambridge, UK.
- NEWHEY, W.K. (1985): Maximum likelihood specification testing and conditional moment tests. *Econometrica*, 53, 1047-1070.
- POLASEK, W. (1999): The BASEL package: A Bayesian Sampling Environment Language. University of Basel, Basel, Switzerland.
- WHITE, H. (1982): Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1-26.
- WHITE, H. (1988): White tests of misspecification. In: S. Kotz, N.L. Johnson and C.B. Read, ed. *Encyclopedia of Statistical sciences*, vol. 9, 594-596, John Wiley and Sons, New York, USA.

Goodness of Fit Measure based on Sample Isotone Regression of Mokken Double Monotonicity Model

Teresa Rivas Moya

Department of Basic Psychology

Psychobiology and Methodology of Behavioral Science, Málaga University

Campus de Teatinos s/n, 29071 Málaga, Spain

(e-mail: moya@uma.es)

Abstract. Based on concepts of Mokken Double Monotonicity model (1971, 1997) and Sample Isotone Regression (Barlow, Bartholomew, Bremmer & Brunk, 1972), a model goodness of fit measure is defined. It permits interpretation of the global deviation from Double Monotonicity in a set of dichotomous response items.

To this end, based on the order induced by the difficulty of the items, the disparity function associated with the proportion of positive and negative responses to pairs of items — given in the matrices \mathbf{P}_{11} and \mathbf{P}_{00} — is defined. In each matrix, the global deviation from Double Monotonicity is obtained as the sum of discrepancies between the proportions of responses observed on pairs of items and the disparities associated with these proportions.

1 Introduction

Mokken (1971, 1997) gives a non parametric Item Response Theory model for the analysis of dichotomous response items. It permits the ordering of the subjects of a sample and the set of items on a unidimensional scale.

A set of items is monotonly homogeneous (MH) if it satisfies unidimensionality, local stochastic independence, and for each item the item response function (IRF) is a monotonly nondecreasing function of the latent scale value. If a set of MH items also has IRFs that do not intersect, the set of items is a doubly monotone (DM) set (Molenaar, Debets, Sijtsma and Hemker, 1994). In the cited references, Mokken, Molenaar, Sijtsma, etc. have proposed methods to investigate whether a set of dichotomous or polytomous items is MH or DM. Molenaar et al. (1994) and, more recently, Molenaar and Sijtsma (1999) put forward a program which obtains important indices for the evaluation of the MH and DM models from responses of subjects to a set of items.

If a set of items is MH, Mokken proposes the study of DM analysing visually the \mathbf{P}_{11} and \mathbf{P}_{00} matrices containing the proportions of subjects giving positive (1, 1) and negative (0, 0) answers to each pair of items i, j . If the items are ordered in decreasing difficulty, an MH set of items is DM when columns and rows in matrix \mathbf{P}_{11} are monotonly nondecreasing, and columns

and rows in matrix \mathbf{P}_{00} are monotonely nonincreasing. Local deviations from these orders are considered violations of DM. This visual analysis is laborious, and the number of values which must be analysed greatly increases when the number of items is large. Molenaar et al. (1994) and Molenaar et al. (1999) suggest considering violations of DM if deviations from monotonicity nondecreasing in rows and columns of the \mathbf{P}_{11} matrix and deviations from monotonicity nonincreasing in rows and columns of the \mathbf{P}_{00} matrix are greater than a fixed value 0.02, 0.03 or other. But how many deviations, and of what size, can be admitted to ensure that a set of items is DM? That is to say, do many small deviations (0.01, 0.02, etc.) have the same effect as a few small deviations from monotonicity? Is the number of deviations obtained independent from the number of items? That is to say, would two deviations of the same magnitude have the same importance with four items as with twelve items, for example? This paper offers a measure which may provide the key to finding responses to these questions.

Based on the study of DM in \mathbf{P}_{11} and \mathbf{P}_{00} matrices, and on the concepts of Sample Isotonic Regression, a method is presented to ascertain if a set of items is DM. For this, disparities $\hat{p}_{ij}(0, 0)$ and $\hat{p}_{ij}(1, 1)$ associated with proportions $p_{ij}(0, 0)$ and $p_{ij}(1, 1)$, in pairs of items which do not satisfy DM, will be estimated in such a way that the set of items is DM when considering these disparities. To this end, the concepts of Cumulative Sum Diagram (CSD) and Greatest Convex Minorant (GCM) of Isotonic Regression (Barlow et al. 1972) are used, which was also adapted in Rivas (1989, 1998) to define disparity function in nonmetric Multidimensional Scaling. Disparities $\hat{p}_{ij}(0, 0)$ and $\hat{p}_{ij}(1, 1)$ are given as the slopes of GCM. In the following section, these are obtained to analyse the nondecreasing order in any row i of the matrix \mathbf{P}_{11} . Similarly, the same concepts can be used to analyse the order established in the columns of the matrix \mathbf{P}_{11} and in the rows and columns of the matrix \mathbf{P}_{00} . Finally, from the proportions observed and the estimated disparities, a measure of the global deviation from DM is obtained.

2 Disparity associated with the positive response proportion on pairs of items

Let $I = \{1, 2, \dots, n\}$ be a set of n items and let $\Omega_i = \{(i, j); i \leq i < j \leq n\}$ be the set of pairs of items — in row i — on which a total order can be established. Given matrix \mathbf{P}_{11} in Table 1, the difficulty of items in columns $j = 1, 2, \dots, n$ induces an order in each row of \mathbf{P}_{11} . Numbers in brackets, in Table 1, denote the nondecreasing order induced by the difficulty. This order can be considered a dissimilarity measure $\delta(i, j)$ between pairs of items i, j . Index of row i can be omitted because in each row it is constant. Then, in any row i , indices of columns $i+1 < i+2 < \dots < n$ or $\delta_{i+1} < \delta_{i+2} < \dots < \delta_n$ denote the dissimilarities between pairs of items.

Table 1. Observed proportions and dissimilarities in matrix \mathbf{P}_{11} .

i/j	i_1	i_2	i_3	i_4	\dots	i_n
i_1	-	p_{12} (2)	p_{13} (3)	p_{14} (4)	\dots	p_{1n} (n)
i_2		-	p_{23} (3)	p_{24} (4)	\dots	p_{2n} (n)
\vdots			\ddots		\vdots	
i_{n-1}				-		$p_{n-1,n}$ (n)
δ_j	δ_1	δ_2	δ_3	δ_4	\dots	δ_n

Let $\delta_i : \Omega_i \rightarrow R$ be the dissimilarity function. $\delta_{ij} = \delta(i, j)$ being the dissimilarity between each pair of items $(i, j) \in \Omega_i$. Then

$$(i, j) \leq (i, k) \text{ iff } \delta_{ij} \leq \delta_{ik} \Leftrightarrow \delta_j \leq \delta_k.$$

This order is estimated from the observed response proportions $p_i : \Omega_i \rightarrow R^+$; $p_{ij} = p_{ij}(1, 1)$ being the positive response proportion on pair of items (i, j) , then $(i, j) \leq (i, k) \Leftrightarrow p_{ij} \leq p_{ik}$.

If $p_{ij} \leq p_{ik} \forall j, k : 1, 2, \dots, n$, is not satisfied in each row i , the DM is violated. Then the disparity function $\hat{p}_i : \Omega_i \rightarrow R$ must be found. It must be found in such a way that if the proportions, p_{ij} , do not satisfy the nondecreasing monotone relation – in the same way as the dissimilarities between items – $\delta(i, j)$, they are substituted by the corresponding disparities \hat{p}_{ij} so that then the relation is nondecreasing monotone. That is, $(i, j) \leq (i, k) \Leftrightarrow \hat{p}_{ij} \leq \hat{p}_{ik}$. The disparities will be obtained as the isotonic regression of proportion function. For this, it is necessary to introduce the following concept of weight function.

Let $w_i : \Omega_i \rightarrow R^+$ be a weight function, $w_{ij} = w(i, j)$ being the *weight* associated with the pair of items (i, j) . If the difficulties of items are different then dissimilarities given in each row i will be different, and each pair of items is associated with the weight 1. Then, in row i , w_i is the identical 1 function. These weights w_{ij} will define the W_{ij} and \hat{W}_{ij} cumulative weights in the following section. W_{ij} and \hat{W}_{ij} will be the abscises of coordinates of CSD and GCM, respectively.

3 Isotonic regression of proportion function

Given $p_i : \Omega_i \rightarrow R^+$ proportion function and a weight function $w_i : \Omega_i \rightarrow R^+$, an isotonic function $\hat{p}_i : \Omega_i \rightarrow R^+$ is isotonic regression of p_i with weights w_i , in regard to the total order over Ω_i , if the following sum is minimized in the class of isotonic functions f :

$$\sum_{i < j} [p_i(i, j) - f_i(i, j)]^2 w_i(i, j).$$

A condition sufficient for the existence of an isotonic regression of a function p_i is that it is bounded (Barlow et al. (1972), pp. 29-30). Then, if $0 \leq p_{ij} \leq 1$, the isotonic regression \hat{p}_i of the proportion function exists and it is also bounded $0 \leq \hat{p}_{ij} \leq 1$. Then, the disparity function is given by the isotonic regression \hat{p}_i of proportion function p_i . Graphically, it is obtained as the function that associates each point (i, j) with the slope of the segment which has this point as endpoint. Then the disparities are these slopes. To obtain them, the following concepts of CSD of p_i and GCM associated with this CSD are given.

The CSD of function p_i is the set of points in the cartesian plane $\mathbf{P}_{00} = (0, 0)$ and $\mathbf{P}_{ij} = (W_{ij}, P_{ij})$ for all $(i, j) \in \Omega_i$, their coordinates being

$$W_{ij} = \sum_{q=1}^i \sum_{h=1}^j w_{qh} \text{ and } P_{ij} = \sum_{q=1}^i \sum_{h=1}^j p_{qh} w_{qh}.$$

The slope of CSD in \mathbf{P}_{ij} is the slope of segment joining the endpoint \mathbf{P}_{ij} to the endpoint of the previous segment \mathbf{P}_{ij-1} . Given the order ' \leq ' over Ω_i , the origin of points is given as $\mathbf{P}_{ij-1} = (W_{ij-1}, P_{ij-1})$ for all $(i, j) \in \Omega_i$. Then the slopes of segments joining \mathbf{P}_{ij-1} to \mathbf{P}_{ij} of CSD are given as

$$\hat{p}_{ij} = \frac{P_{ij} - P_{ij-1}}{W_{ij} - W_{ij-1}} \text{ being } i < j, j : i+1, \dots, n.$$

If $P_{in-1} > P_{in}$ and P_{ij} is the last value of row i that satisfies the monotony, then $P_{in+1} = P_{in} + p_{ij}$ is considered and $\hat{p}_{in} = (P_{in+1} - P_{in})/(W_{in+1} - W_{in})$.

The GCM of a set of convex functions is the supremum of all convex functions whose graphs lie below the CSD. Graphically, it is the path under which lies a taut string of minimum length, joining \mathbf{P}_{00} and $\mathbf{P}_{n-1,n}$ lying below the CSD. The GCM is obtained joining the $\hat{\mathbf{P}}_{ij} = (\hat{W}_{ij}, \hat{P}_{ij})$ whose coordinates are $\hat{W}_{ij} = W_{ij}$ and $\hat{P}_{ij} = \sum_{q=1}^i \sum_{h=2}^j \hat{p}_{qh} w_{qh}$.

The slope of GCM in $\hat{\mathbf{P}}_{ij}$ is the slope of segment with origin $\hat{\mathbf{P}}_{ij-1}$ and endpoint $\hat{\mathbf{P}}_{ij} : \hat{p}_{ij} = (\hat{P}_{ij} - \hat{P}_{ij-1})/(\hat{W}_{ij} - \hat{W}_{ij-1})$ being $i < j, j : i+1, \dots, n$ or, in the other case, $\hat{p}_{in} = (\hat{P}_{in+1} - \hat{P}_{in})/(\hat{W}_{in+1} - \hat{W}_{in})$.

4 Goodness of fit for Mokken Double Monotonicity model

The global deviation from monotonicity nondecreasing in the rows of matrix \mathbf{P}_{11} is given as

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^n |p_{ij} - \hat{p}_{ij}| \text{ being } 0 \leq \sum_{i=1}^{n-1} \sum_{j=i+1}^n |p_{ij} - \hat{p}_{ij}| \leq n(n-1)/2.$$

Similarly this deviation in the columns of \mathbf{P}_{11} is

$$\sum_{j=1}^{n-1} \sum_{i=j+1}^n |p_{ij} - \hat{p}_{ij}| \text{ being } 0 \leq \sum_{j=1}^{n-1} \sum_{i=j+1}^n |p_{ij} - \hat{p}_{ij}| \leq n(n-1)/2.$$

Then the global deviation from monotonicity nondecreasing in matrix \mathbf{P}_{11} is denoted by $D(\mathbf{P}_{11}, \Delta, n)$, Δ being the dissimilarity matrix, and is given as:

$$D(\mathbf{P}_{11}, \Delta, n) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n |p_{ij}(1,1) - \hat{p}_{ij}(1,1)|}{n(n-1)} + \\ \frac{\sum_{j=1}^{n-1} \sum_{i=j+1}^n |p_{ij}(1,1) - \hat{p}_{ij}(1,1)|}{n(n-1)}.$$

Similarly, the global deviation from monotonicity nonincreasing in matrix \mathbf{P}_{00} is given as

$$D(\mathbf{P}_{00}, \Delta, n) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n |p_{ij}(0,0) - \hat{p}_{ij}(0,0)|}{n(n-1)} + \\ \frac{\sum_{j=1}^{n-1} \sum_{i=j+1}^n |p_{ij}(0,0) - \hat{p}_{ij}(0,0)|}{n(n-1)}.$$

These measures of the deviation from monotonicity are bounded between 0 and 1. The set of items is DM if these measures are 0. The maximum deviation from DM, in relation to the order induced by the difficulty on the set of items, is given when these measures are 1.

5 Conclusions

From the preservation of order induced by the difficulty of the items, and applying the Sample Isotone Regression, a goodness of fit measure of DM model has been obtained. It permits quantification of the global deviation from DM. This measure assumes values between 0 and 1 for any set of items. Thus, a global measure of the extent in which the DM is not satisfied for a set of items is given. In addition, the extent to which several sets of items of different size do not satisfy DM can be compared. Finally, studies and applications using this measure would be necessary to find lower and upper boundaries which show the size (small, medium, large) of global deviation from DM.

This measure could be applied after the procedures which evaluate the DM in each item in regard to each one of the remaining items given in Molenaar et al. (1994) and Molenaar et al. (1999). Extensions of this measure can also be obtained to analyse deviations from DM with items of polytomous responses.

References

- BARLOW, R.E., BARTHOLOMEW,D.J., BREMNER, J.M. and BRUNK, H.D. (1972): *Statistical Inference under order restrictions.* John Wiley and Sons, London.
- MOKKEN, R.J. (1971): *A Theory and Procedure of Scale Analysis with Applications in Political Research.* Walter de Gruyter, Mouton, Berlin.
- MOKKEN, R.J. (1997): Nonparametric Models for Dichotomous Responses. In: W. J. van der Linden and R. K. Hambleton (Eds.): *Handbook of Modern Item Response Theory.* Springer, New York. 351–367.
- MOLENAAR, I.W., DEBETS, P., SIJTSMA, K. and HEMKER, B.T. (1994): *USER's Manual MSP4. A program for Mokken Scale Analysis for Polytomous Items.* Iec ProGAMMA, Groningen.
- MOLENAAR, I.W. and SIJTSMA, K. (1999): *USER's Manual MSP5 for Windows. A program for Mokken Scale Analysis for Polytomous Items.* Iec ProGAMMA, Groningen.
- RIVAS MOYA, T. (1989): Medida de bondad del ajuste del modelo de Escalamiento Multidimensional Euclideo. Unpublished Ph. D. Thesis. Málaga University.
- RIVAS MOYA, T. (1998): *Obtaining disparities as slopes of Greatest Minorant Convex.* In: A. Rizzi, M. Vichi and H.-H. Bock (Eds.): *Advances in Data Science and Classification* Springer, Berlin, 509–516.

Fuzzy Time Arrays and Dissimilarity Measures For Fuzzy Time Trajectories

Renato Coppi and Pierpaolo D'Urso

Dipartimento di Statistica, Probabilitá e Statistiche Applicate
 Universitá di Roma "La Sapienza", P.le A. Moro, 5, 00185, Roma, Italy
 (e-mail:coppir@pow2.sta.uniroma1.it)

Abstract. In this paper we define a fuzzy extension of a time array. The algebraic and geometric characteristics of the fuzzy time array are analyzed. Furthermore, considering the objects space \Re^{J+1} , where J is the number of variables and the remaining dimension is related to time, we suggest different dissimilarity measures for fuzzy time trajectories.

1 Fuzzy time arrays

Let $\mathbf{X} \equiv \{x_{ijt} = (c_{ijt}, r_{ijt}) : i = 1, I; j = 1, J; t = 1, T\}$ be the fuzzy time array (same objects \times same variables \times times) where i, j and t denote the objects, variables and times, respectively; $x_{ijt} = (c_{ijt}, r_{ijt})$ represents the symmetrical fuzzy variable j observed on the i object at time t , where c_{ijt} denotes the center (or mode) and r_{ijt} the left and right spread (or width). Usually a fuzzy variable $x_{ijt} = (c_{ijt}, r_{ijt})$ is defined by means of a triangular membership function (Zadeh, 1965; Zimmermann, 1991):

$$\mu(u_{ijt}) = \begin{cases} 1 - \frac{|c_{ijt} - u_{ijt}|}{r_{ijt}} & c_{ijt} - r_{ijt} \leq u_{ijt} \leq c_{ijt} + r_{ijt} \\ 0 & \text{otherwise} \end{cases}$$

where the fuzzy variable x_{ijt} is completely identified by the two parameters c_{ijt} (center) and r_{ijt} (left and right spread). If $r_{ijt} = 0$ then \mathbf{X} is the non-fuzzy time array (Coppi, D'Urso, 1999; D'Urso, Vichi, 1998). Considering the fuzzy time array \mathbf{X} we derive: $\mathbf{C} \equiv \{c_{ijt} : i = 1, I; j = 1, J; t = 1, T\}$ (centers time array) and $\mathbf{R} \equiv \{r_{ijt} : i = 1, I; j = 1, J; t = 1, T\}$ (spreads time array). If we combine suitably the indices I, J and T , we obtain from \mathbf{X} the following fuzzy supermatrices: $\mathbf{X} \equiv \{\mathbf{X}_i\}_{i=1,I}$, $\mathbf{X} \equiv \{\mathbf{X}_t\}_{t=1,T}$ and $\mathbf{X} \equiv \{\mathbf{X}_j\}_{j=1,J}$, where $\mathbf{X}_i \equiv \{x_{ijt} : j = 1, J; t = 1, T\}$, $\mathbf{X}_t \equiv \{x_{ijt} : i = 1, I; j = 1, J\}$ and $\mathbf{X}_j \equiv \{x_{ijt} : j = 1, I; t = 1, T\}$. Considering the centers and spreads time arrays, we can find the centers and spreads supermatrices respectively:

$\mathbf{C} \equiv \{\mathbf{C}_i\}_{i=1,I}$, $\mathbf{C} \equiv \{\mathbf{C}_t\}_{t=1,T}$, $\mathbf{C} \equiv \{\mathbf{C}_j\}_{j=1,J}$ and $\mathbf{R} \equiv \{\mathbf{R}_i\}_{i=1,I}$, $\mathbf{R} \equiv \{\mathbf{R}_t\}_{t=1,T}$, $\mathbf{R} \equiv \{\mathbf{R}_j\}_{j=1,J}$, where $\mathbf{C}_i \equiv \{c_{ijt} : j = 1, J; t = 1, T\}$, $\mathbf{C}_t \equiv \{c_{ijt} : i = 1, I; j = 1, J\}$, $\mathbf{C}_j \equiv \{c_{ijt} : i = 1, I; t = 1, T\}$, $\mathbf{R}_i \equiv \{r_{ijt} : j = 1, J; t = 1, T\}$, $\mathbf{R}_t \equiv \{r_{ijt} : i = 1, I; j = 1, J\}$ and $\mathbf{R}_j \equiv \{r_{ijt} : i = 1, I; t = 1, T\}$.

The introduction of the notion of “vagueness” in the treatment of multivariate time arrays broadens the scope of this type of data (including ill defined measurements) and of the methods for analyzing them (e.g. possibility of defining fuzzy trajectories, fuzzy classification of individuals, etc.). A hint to possible applications is provided in sec. 4.

2 The geometric representations of a fuzzy time array

Let \Re^{J+1} be the vectorial space, where the axes are referred to the J variables and time. In this space we represent each object i by means of the vectors, for each t : ${}_c\mathbf{u}_{it} = (c_{i1t}, \dots, c_{ijt}, \dots, c_{iJt}, t)^t$ and ${}_r\mathbf{u}_{it} = (r_{i1t}, \dots, r_{ijt}, \dots, r_{iJt}, t)^t$. Fixed t , the scatter of the points (matrices), ${}_fN_I(t) \equiv \{({}_c\mathbf{u}_{it}|{}_r\mathbf{u}_{it})\}_{i=1,I}$, represents the matrix \mathbf{X}_t . For each t , the scatters ${}_fN_I(t)$ are placed on hyperplanes parallel to the sub-space \Re^J (Figure 1).

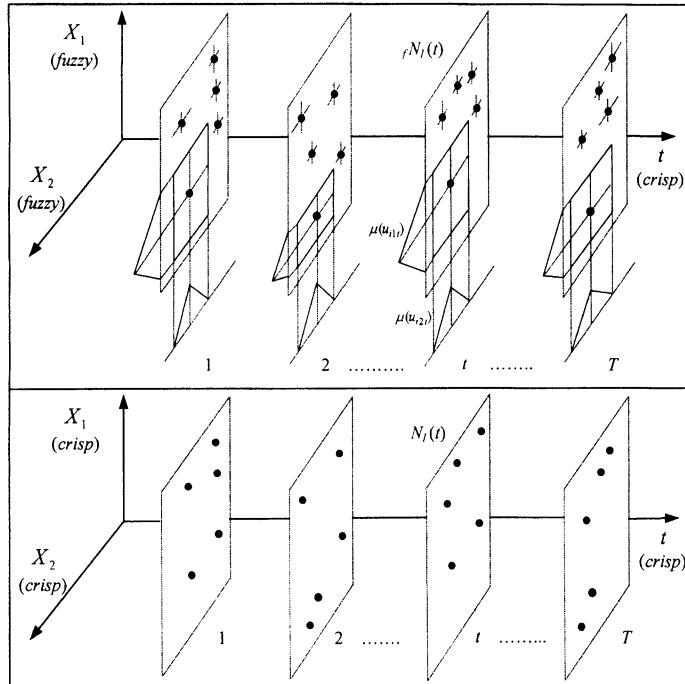


Fig. 1. Examples of fuzzy (${}_fN_I(t)$) and non-fuzzy scatters ($N_I(t)$)

Fixed i , ${}_fN_T(i) \equiv \{({}_c\mathbf{u}_{it}|{}_r\mathbf{u}_{it})\}_{t=1,T}$ represents \mathbf{X}_i . This scatter describes the *time trajectory* of object i across the time and $\left\{{}_fN_T(i) \equiv \{({}_c\mathbf{u}_{it}|{}_r\mathbf{u}_{it})\}_{t=1,T}\right\}_{i=1,I}$ represents the set of the time trajectories. Each time trajectory ${}_fN_T(i)$ crosses the T hyperplanes parallel to \Re^J (Figure 2).

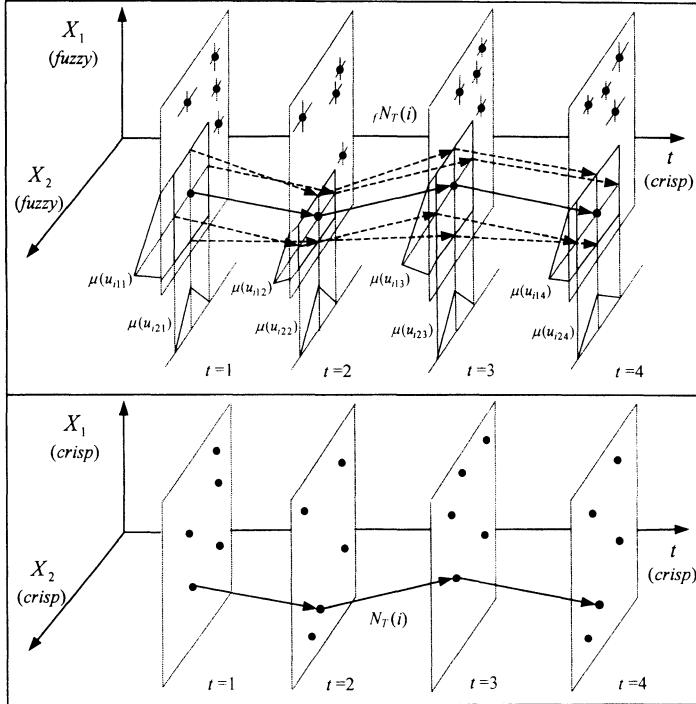


Fig. 2. Examples of fuzzy $fN_T(i)$ and non-fuzzy time trajectories ($N_T(i)$)

3 Dissimilarity measures between fuzzy time trajectories

Introducing an Euclidean metric in the space \Re^{J+1} , in this section we propose different dissimilarity measures for fuzzy synchronous (with variables observed at the same times) multivariate time trajectories with constant time support (with equidistant time intervals). In order to compare the fuzzy time trajectories taking into account their cross sectional and longitudinal characteristics we define dissimilarities between their position, velocity and acceleration (Carlier, 1986, 1991; Coppi, D'Urso, 1999; D'Urso, Vichi, 1998).

Remark 1. Before computing the following dissimilarity measures, the variables have to be standardized (Harshman, Lundy, 1984; Rizzi, Vichi, 1995).

Considering the centers time array \mathbf{C} , we define the following *center dissimilarity measures* between i -th and l -th fuzzy time trajectories:

$${}_1d_{ilc}^2 = \sum_{t=1}^T \sum_{j=1}^J ({}_1c_{ijt} - {}_1c_{ljt})^2 = \sum_{t=1}^T \| {}_1\mathbf{c}_{it} - {}_1\mathbf{c}_{lt} \|^2 \quad (\text{center-position dissimilarity}) \quad (1)$$

$${}_2d_{ilc}^2 = \sum_{t=2}^T \sum_{j=1}^J ({}_2c_{ijt} - {}_2c_{ljt})^2 = \sum_{t=2}^T \| {}_2\mathbf{c}_{it} - {}_2\mathbf{c}_{lt} \|^2 \quad (\text{center-velocity dissimilarity}) \quad (2)$$

$$3d_{ilc}^2 = \sum_{t=3}^T \sum_{j=1}^J ({}_3c_{ijt} - {}_3c_{ljt})^2 = \sum_{t=3}^T \| {}_3\mathbf{c}_{it} - {}_3\mathbf{c}_{lt} \|^2 \quad (\text{center- acceleration dissimilarity}) \quad (3)$$

$$d_{ilc}^2 = \sum_{p=1}^3 \alpha_p d_{ilc}^2, \quad \alpha_p \geq 0 \quad (\text{center-mixed dissimilarity}) \quad (4)$$

where, ${}_2\mathbf{c}_{it} = ({}_1\mathbf{c}_{it} - {}_1\mathbf{c}_{it-1})$ and ${}_3\mathbf{c}_{it} = \frac{1}{2}(2\mathbf{c}_{it} - 2\mathbf{c}_{it-1})$ are the velocity and acceleration (Coppi, D'Urso, 1999; D'Urso, Vichi, 1998) of the centers of i -th and l -th fuzzy time trajectories; ${}_1\mathbf{c}_{it} = ({}_1c_{i1t}, \dots, {}_1c_{ijt}, \dots, {}_1c_{iJt})'$, ${}_2\mathbf{c}_{it} = ({}_2c_{i1t}, \dots, {}_2c_{ijt}, \dots, {}_2c_{iJt})'$ and ${}_3\mathbf{c}_{it} = ({}_3c_{i1t}, \dots, {}_3c_{ijt}, \dots, {}_3c_{iJt})'$.

Considering the spreads time array \mathbf{R} , we define the following *spread dissimilarity measures* between i -th and l -th fuzzy time trajectories:

$${}_1d_{ilr}^2 = \sum_{t=1}^T \sum_{j=1}^J ({}_1r_{ijt} - {}_1r_{ljt})^2 = \sum_{t=1}^T \| {}_1\mathbf{r}_{it} - {}_1\mathbf{r}_{lt} \|^2 \quad (\text{spread-position dissimilarity}) \quad (5)$$

$${}_2d_{ilr}^2 = \sum_{t=2}^T \sum_{j=1}^J ({}_2r_{ijt} - {}_2r_{ljt})^2 = \sum_{t=2}^T \| {}_2\mathbf{r}_{it} - {}_2\mathbf{r}_{lt} \|^2 \quad (\text{spread-velocity dissimilarity}) \quad (6)$$

$${}_3d_{ilr}^2 = \sum_{t=3}^T \sum_{j=1}^J ({}_3r_{ijt} - {}_3r_{ljt})^2 = \sum_{t=3}^T \| {}_3\mathbf{r}_{it} - {}_3\mathbf{r}_{lt} \|^2 \quad (\text{spread- acceleration dissimilarity}) \quad (7)$$

$$d_{ilr}^2 = \sum_{p=1}^3 \alpha_p d_{ilr}^2, \quad \alpha_p \geq 0 \quad (\text{spread-mixed dissimilarity}) \quad (8)$$

where ${}_1\mathbf{r}_{it} = ({}_1\mathbf{r}_{it} - {}_1\mathbf{r}_{it-1})'$ and ${}_3\mathbf{r}_{it} = \frac{1}{2}(2\mathbf{r}_{it} - 2\mathbf{r}_{it-1})'$ are the velocity and acceleration of the spreads of i -th and l -th fuzzy time trajectories; ${}_1\mathbf{r}_{it} = ({}_1r_{i1t}, \dots, {}_1r_{ijt}, \dots, {}_1r_{iJt})'$, ${}_2\mathbf{r}_{it} = ({}_2r_{i1t}, \dots, {}_2r_{ijt}, \dots, {}_2r_{iJt})'$ and ${}_3\mathbf{r}_{it} = ({}_3r_{i1t}, \dots, {}_3r_{ijt}, \dots, {}_3r_{iJt})'$. Considering the different typologies of center and spread dissimilarities we obtain the following *fuzzy dissimilarity measures*:

$${}_1d_{il}^2 = {}_1\beta_1 d_{ilc}^2 + (1 - {}_1\beta_1) d_{ilr}^2 \quad (\text{fuzzy-position dissimilarity}) \quad (9)$$

$${}_2d_{il}^2 = {}_2\beta_2 d_{ilc}^2 + (1 - {}_2\beta_2) d_{ilr}^2 \quad (\text{fuzzy-velocity dissimilarity}) \quad (10)$$

$${}_3d_{il}^2 = {}_3\beta_3 d_{ilc}^2 + (1 - {}_3\beta_3) d_{ilr}^2 \quad (\text{fuzzy-acceleration dissimilarity}) \quad (11)$$

$$d_{il}^2 = \beta d_{ilc}^2 + (1 - \beta) d_{ilr}^2 \quad (\text{fuzzy-mixed dissimilarity}) \quad (12)$$

where ${}_1\beta, {}_2\beta, {}_3\beta, \beta \geq 0$.

There are different criteria to compute the weights of the dissimilarity measures (4) and (8)-(12). It is possible to define the weights subjectively or objectively. In the non-fuzzy case Carlier proposed a mixed dissimilarity in which the weights are determined subjectively (1986, 1991). An objective procedure has been suggested by D'Urso and Vichi (1998). This procedure can be extended to dissimilarity measures (4) and (8)-(12).

Remark 2. Before computing the dissimilarities (4), (8), (9), (10), (11) and (12), the dissimilarities (1) (2), (3), (5), (6) and (7) have to be normalized.

Remark 3. If we consider non-fuzzy time trajectories, then: ${}_1 d_{ilr}^2 = {}_2 d_{ilr}^2 = {}_3 d_{ilr}^2 = d_{ilr}^2 = 0$ and (9)-(12) coincide with the dissimilarities proposed by D'Urso and Vichi (1998).

Remark 4. For the dissimilarity measures (4), (8) and (12) we can consider the conic combination of the position component and one of the two longitudinal components (velocity or acceleration). If we consider the combination of the position and velocity dissimilarities between non-fuzzy time trajectories the *center-mixed dissimilarity measure* coincides with the mixed dissimilarity proposed by Carlier (1986, 1991).

4 Final remarks

In this paper we defined the fuzzy time array and proposed different fuzzy dissimilarity measures for fuzzy time trajectories. Other fuzzy extensions of dissimilarity measures for non-fuzzy time trajectories (Carlier, 1999; Coppi, D'Urso, 1999; Saporta, Lavellard, 1996) could be considered. This will be the object of further studies. Another perspective of research, in this context, is provided by the use of the proposed fuzzy dissimilarity measures in the framework of fuzzy classification of time trajectories. This methodological area has been extensively explored in D'Urso (1999), with reference to crisp time arrays. Many of the ideas discussed in this connection can be suitably generalized to fuzzy time arrays.

As to the practical utilization of the fuzzy approach in this field of data analysis, several potential examples might be mentioned, ranging from economical to psychological and engineering problems. One of these concerns the assessment of various features of different wines produced in a series of years (cfr. Hartigan, 1975). Each feature is originally assessed on a qualitative scale. To each qualitative judgement can be suitably associated a fuzzy number. Thus, a fuzzy time array is obtained, fuzzy dissimilarities between wines across the years can be computed and an appropriate classification system of wines can be set up, on the previously illustrated grounds. This application will be developed in a subsequent work.

References

- CARLIER, A. (1986): Factor Analysis of Evolution and Cluster Methods on Trajectories. In: F. De Antoni, N. Lauro, and A. Rizzi (Eds.): *Proceedings in Computational Statistics (COMPSTAT)*, Physica-Verlag, Heidelberg ,140-145.
- CARLIER, A. (1991): About Distances for Clustering Longitudinal Multivariate Data, *Proceedings of the Third Conference of the IFCS*, Edinburgh, Scotland.

- CARLIER, A. (1999): Distances Between Trajectories for Longitudinal Data, *Proceedings of the Conference of the VOC*, Leiden, The Netherlands.
- COPPI, R. and D'URSO, P. (1999): The Geometric Approach to the Comparison of Multivariate Time Trajectories, *Proceedings of the Conference of the CLADAG-SIS*, Rome, 5-6 July, 1999, 177-180.
- D'URSO, P. (1999): *Fuzzy Classification for Time Arrays*, PhD Thesis, University "La Sapienza", Rome, Italy (in Italian).
- D'URSO, P. and VICHI, M. (1998): Dissimilarities Between Trajectories of a Three-Way Longitudinal Data Set. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*, Springer, Heidelberg, 585-592.
- HARSHMAN, R. A. and LUNDY, M. E. (1984): Data Extended PARAFAC Model. In: H. G. Law, C. W. Snyder, Jr, J. A. Hattie and R. P. McDonald (Eds.): *Research Methods for Multimode Data Analysis*, Praeger, New York, 216-284.
- HARTIGAN, J.A. (1975): *Clustering Algorithms*, John Wiley & Sons, New York.
- RIZZI, A. and VICHI, M. (1995): Representations, Synthesis, Variability and Data Preprocessing of a Three-way Data Set, *Computational Statistics & Data Analysis*, 19, 203-222.
- SAPORTA, G. and LAVALLARD, F. (1996): *L'Analyse des Données Évolutives. Méthodes et Applications*. Éditions Technip, Paris.
- ZADEH, L. A. (1965): Fuzzy Sets, *Informat. Control*, 8, 338-353.
- ZIMMERMANN, H. J. (1991): *Fuzzy Set Theory and its Application*, Kluwer Academic Press, Dordrecht.

Three-Way Partial Correlation Measures

Donatella Vicari

Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma “La Sapienza”
(e-mail: vicari@pow2.sta.uniroma1.it)

Abstract. Analysis of linear relations between variables, given a third one, can be investigated for three-way three-mode data, by defining new measures of linear dependence between occasions. In this paper, two partial correlation coefficients between matrices are proposed. Their properties are analyzed, in particular with respect to the absence of conditional linear dependence.

1 Introduction

In complex phenomena, the analysis of relations between variables is of great importance. In the literature, several different measures of correlation between matrices have been proposed. They are syntheses of the linear relations between variables, which generalize the usual Pearson correlation coefficient to the three-way context (Cramèr and Nicewander, 1979; Vichi, 1989). Similarly, relations between sets of variables, controlling the influence of variables observed in different occasions, can be measured (Vicari, 1994).

In this paper, some partial correlation coefficients between matrices are considered. They synthesize the linear relations between sets of variables observed on the same units in two different occasions, when the influence of the other situations is controlled. Two partial correlation coefficients are proposed and their properties are analyzed, in particular with respect to the absence of conditional linear dependence.

2 Partial correlation between matrices

Let $\mathbf{X} = \{x_{ijh} : i = 1, \dots, n; j = 1, \dots, p; h = 1, \dots, r\}$ be a three-way data set, where x_{ijh} is the value of the j -th variable observed on the i -th unit, at the h -th occasion. The three-way array \mathbf{X} can be represented as a set of r two-way unit-variable matrices \mathbf{X}_h , containing all the information referred to the h -th occasion. Without loss of generality, \mathbf{X}_h can be centered and column standardized with unit variance. The most well-known correlation coefficient between matrices is, undoubtedly, RV (Escoufier, 1986):

$$RV(\mathbf{X}_h, \mathbf{X}_k) = \frac{\text{tr}(\mathbf{X}'_h \mathbf{X}_k \mathbf{X}'_k \mathbf{X}_h)}{[(\text{tr}(\mathbf{X}'_h \mathbf{X}_h \mathbf{X}'_h \mathbf{X}_h)(\text{tr}(\mathbf{X}'_k \mathbf{X}_k \mathbf{X}'_k \mathbf{X}_k))]^{1/2}} \quad (1)$$

which assumes values in $[0, 1]$. RV takes its maximum value when \mathbf{X}_h and \mathbf{X}_k are proportional or similar (equal up to an orthogonal transformation). In specific cases, the RV coefficient can be put in relation to well-known correlation coefficients (Escoufier, 1986), such as: linear correlation coefficient, multiple correlation coefficient, Pearson correlation ratio, chi-square contingency criterion, canonical correlation coefficient. In particular:

- a) if \mathbf{X}_h and \mathbf{X}_k are two vectors ($n \times 1$), then $RV(\mathbf{X}_h, \mathbf{X}_k) = r^2(\mathbf{X}_h, \mathbf{X}_k)$, where r is the usual Pearson linear correlation coefficient;
- b) if \mathbf{X}_h and \mathbf{X}_k are the centered dummy matrices associated to a pair of qualitative variables (with p_1 and p_2 levels, respectively) and weighted by the reciprocal marginal column totals, then:

$$RV(\mathbf{X}_h, \mathbf{X}_k) = T_{hk}^2 = \frac{\chi_{hk}^2}{n\sqrt{(p_1 - 1)(p_2 - 1)}},$$

where T_{hk} is the Tchuprov coefficient and χ_{hk}^2 is the usual measure of association between the variables under consideration.

In a three-way context, RV , and consequently all the derived measures, suffers the same limitations of the usual Pearson coefficient between variables, if the aim is to evaluate the correlation between occasions after the effect of the others has been removed. In fact, $RV(\mathbf{X}_h, \mathbf{X}_k)$ does not involve the other $r - 2$ occasions, for which data are available in this context.

Saporta (1976) defines the Tchuprov partial correlation coefficient between \mathbf{X}_h and \mathbf{X}_k , controlling \mathbf{X}_s :

$$T_{hk,s} = \frac{T_{hk} - T_{hs}T_{ks}}{[(1 - T_{hs}^2)(1 - T_{ks}^2)]^{1/2}}, \quad (2)$$

where the Tchuprov coefficient T_{hk} measures the association between variables h and k and \mathbf{X}_h , \mathbf{X}_k , and \mathbf{X}_s denote dummy matrices representing categorical variables.

Daudin (1979) analyzed the behavior of (2) with respect to the case of conditional independence between \mathbf{X}_h and \mathbf{X}_k , given \mathbf{X}_s , that is, he wanted to assess whether the following equivalence holds:

$$T_{hk,s} = 0 \Leftrightarrow (\mathbf{X}_h \perp \mathbf{X}_k | \mathbf{X}_s).$$

Daudin provided some counter-examples to show that, when variables have more than two categories (hence, the dummy matrices have more than two columns), the implication does not necessarily hold. In particular, the lack of the nullity of the coefficient, when \mathbf{X}_h and \mathbf{X}_k are conditionally independent seems to compromise the validity of its use, when the number of categories is more than two. Daudin did not analyze why the Tchuprov partial coefficient (2) is not appropriate, or which are the missing terms in its expression.

3 Partial correlation measures

Generally speaking, when a partial correlation coefficient between matrices is defined, the relevant question is the one that Daudin (1979) pointed out: assessing the behavior of the index when \mathbf{X}_h and \mathbf{X}_k are not correlated, controlling \mathbf{X}_s . Without loss of generality, let us consider three matrices \mathbf{X} , \mathbf{Y} , and \mathbf{Z} : for $r > 3$ the same considerations hold. To evaluate the linear relations between \mathbf{X} and \mathbf{Y} , controlling \mathbf{Z} , let us consider the linear regressions of \mathbf{Y} and \mathbf{X} , respectively, on \mathbf{Z} and the corresponding least squares solutions:

$$\begin{aligned}\mathbf{Y} &= \mathbf{ZB} + \mathbf{E}_1 \quad \text{and} \quad \mathbf{X} = \mathbf{ZC} + \mathbf{E}_2 \\ \hat{\mathbf{B}} &= (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \quad \text{and} \quad \hat{\mathbf{C}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}.\end{aligned}$$

By premultiplying by \mathbf{Z} , it holds:

$$\mathbf{Z}\hat{\mathbf{B}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} = \mathbf{H}_{\mathbf{Z}}\mathbf{Y} \quad \text{and} \quad \mathbf{Z}\hat{\mathbf{C}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} = \mathbf{H}_{\mathbf{Z}}\mathbf{X},$$

where $\mathbf{H}_{\mathbf{Z}}$ is the projection matrix. Residual matrices from the regressions can be written:

$$\mathbf{X}^* = \mathbf{X} - \mathbf{H}_{\mathbf{Z}}\mathbf{X} \quad \text{and} \quad \mathbf{Y}^* = \mathbf{Y} - \mathbf{H}_{\mathbf{Z}}\mathbf{Y}.$$

They represent those portions of the original data matrices that have no dependence on the values of the variables in \mathbf{Z} . Hence, similarly to the univariate case (Draper and Smith, 1966), where partial correlations are determined as correlations between residuals from linear regressions, a partial correlation coefficient between matrices can be defined by involving the starred matrices:

$$P_{XY.Z} = \frac{\|\mathbf{X}^{*\prime}\mathbf{Y}^*\|}{(\|\mathbf{X}^{*\prime}\mathbf{X}^*\| \|\mathbf{Y}^{*\prime}\mathbf{Y}^*\|)^{1/2}}, \quad (3)$$

where $\|\cdot\|$ stands for any matrix norm.

3.1 Two partial correlation coefficients based on matrix norms

We first consider using the spectral norm in expression (3):

$$\|\mathbf{X}^{*\prime}\mathbf{Y}^*\|_s = \sqrt{\lambda_{\max}},$$

where λ_{\max} is the largest eigenvalue of $\mathbf{X}^{*\prime}\mathbf{Y}^*\mathbf{Y}^{*\prime}\mathbf{X}^*$. In this case, the coefficient is symmetric, takes values in $[0, 1]$, assuming its maximum when the two matrices are proportional or similar (equal up to an orthogonal transformation). The largest eigenvalue λ_{\max} is zero, and the spectral norm is also zero, if and only if the cross-product matrix is null; hence, $P_{XY.Z}$ is zero when linear correlation between \mathbf{X} and \mathbf{Y} , controlling \mathbf{Z} , does not exist:

$$\mathbf{X}^{*\prime}\mathbf{Y}^* = \mathbf{0} \Leftrightarrow P_{XY.Z} = 0.$$

Note that the coefficient is indeterminate only in the irrelevant case when all cross-product matrices involved are null.

Next, we consider using the Schur (or Frobenius) norm in (3). We then obtain what we call the partial *RV* coefficient:

$$RV(\mathbf{X}\mathbf{Y}, \mathbf{Z}) = \frac{\text{tr}(\mathbf{X}^{*\prime}\mathbf{Y}^*\mathbf{Y}^{*\prime}\mathbf{X}^*)}{\sqrt{\text{tr}(\mathbf{X}^{*\prime}\mathbf{X}^*\mathbf{X}^{*\prime}\mathbf{X}^*)\text{tr}(\mathbf{Y}^{*\prime}\mathbf{Y}^*\mathbf{Y}^{*\prime}\mathbf{Y}^*)}},$$

where the numerator can be written:

$$\text{tr}(\mathbf{X}^{*\prime}\mathbf{Y}^*\mathbf{Y}^{*\prime}\mathbf{X}^*) = \text{tr}[(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{H}_Z\mathbf{Y})(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{H}_Z\mathbf{Y})']. \quad (4)$$

It is interesting to note that, in case each matrix contains only one column, the partial *RV* coefficient reduces to the square of the partial correlation coefficient.

Now, let us consider the special case where \mathbf{X} , \mathbf{Y} , and \mathbf{Z} are dummy matrices associated to three categorical variables. The numerator of a relative coefficient of association between the categorical variables X and Y , controlling a third one Z , can be derived from (4):

$$\sum_i \sum_j \left(n_{ij} - \sum_k \frac{n_{i.k}n_{.jk}}{n_{..k}} \right)^2, \quad (5)$$

where n_{ij} is the marginal total referred to categories i and j , respectively of X and Y .

The expression in (5) is zero if and only if:

$$\sum_k \frac{n_{i.k}n_{.jk}}{n_{..k}} = n_{ij} \quad \forall i, j = 1, \dots, p,$$

which is just the association frequency in case of conditional independence.

3.2 A partial correlation coefficient based on determinants

Now let us consider substituting the matrix norm in (3) by the determinant:

$$|\mathbf{X}^{*\prime}\mathbf{Y}^*| = |(\mathbf{X} - \mathbf{H}_Z\mathbf{X})'(\mathbf{Y} - \mathbf{H}_Z\mathbf{Y})| = |\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{H}_Z\mathbf{Y}|.$$

In this case, for column-standardized matrices, we have

$$n^{-1}|\mathbf{X}^{*\prime}\mathbf{Y}^*| = |\mathbf{R}_{XY} - \mathbf{R}_{XZ}\mathbf{R}_{ZZ}^{-1}\mathbf{R}_{ZY}|, \quad (6)$$

where \mathbf{R}_{hh} is the usual $p \times p$ correlation matrix *inside* the h -th occasion and \mathbf{R}_{hk} is the $p \times p$ correlation matrix *between* the occasions h and k .

The expression in (6) equals the numerator of a partial correlation coefficient between matrices, proposed by Vicari (1994), following a different approach.

Let us define the block matrix $\mathbf{V} = \{\mathbf{R}_{hk} : h, k = 1, \dots, p\}$, where \mathbf{R}_{hk} is the nonsingular $p \times p$ correlation matrix between occasions h and k . The cofactor \mathbf{v}^{hk} of the submatrix \mathbf{R}_{hk} of \mathbf{V} is: $\mathbf{v}^{hk} = (-1)^{h+k} \text{minor}(\mathbf{R}_{hk})$, where $\text{minor}(\mathbf{R}_{hk})$ denotes the determinant of the submatrix obtained from \mathbf{V} , by deleting the p rows and p columns containing \mathbf{R}_{hk} . Hence, the usual expression of the partial correlation coefficient between variables can be generalized:

$$C_{hk,s} = \frac{-\mathbf{v}^{hk}}{(\mathbf{v}^{hh}\mathbf{v}^{kk})^{1/2}}. \quad (7)$$

When \mathbf{X}_h , \mathbf{X}_k , and \mathbf{X}_s are, respectively, \mathbf{X} , \mathbf{Y} , and \mathbf{Z} , the term $-\mathbf{v}^{hk}$ coincides just with (6), up to the determinant $|\mathbf{R}_{ZZ}|$ which cancels out when the denominator is considered. The coefficient (7), which is symmetric and takes values in $[-1, 1]$, assumes the endpoints when \mathbf{X}_h and \mathbf{X}_k are proportional or similar, conditionally to \mathbf{X}_s . Furthermore, when each matrix has only one column ($p = 1$), $C_{hk,s}$ is just the usual partial correlation coefficient between variables. Therefore, the partial correlation coefficient between occasions C_{XYZ} , even introduced in a different way, can be derived analogously to the univariate case.

The coefficient is zero, when linear correlation between the starred matrices does not exist:

$$\mathbf{X}^{*\prime}\mathbf{Y}^* = \mathbf{0} \Rightarrow C_{hk,s} = 0,$$

which excludes the worst situation as in Daudin. The determinant of $\mathbf{X}^{*\prime}\mathbf{Y}^*$ can be zero even when a perfect linear relation exists inside one of the two matrices, but in this case C_{XYZ} is indeterminate.

4 Conclusions

Partial correlation coefficients between matrices synthesize linear correlations between sets of variables, after removing the effect of the variables observed in different situations, but related to the same (for instance, economic, social, environmental, biological) context.

The definition of such indices needs to take into account the complexity of the data by considering, in particular, their behavior in absence of linear conditional dependence.

Two partial correlation coefficients are derived, similarly to the univariate case. Such coefficients, based on matrix norms and determinants, are zero when linear dependence between matrices does not exist, controlling the influence of the other matrices.

Starting from the partial RV coefficient, a coefficient of conditional association between variables can be defined, which is zero if and only if the two variables are conditionally independent and excludes the incoherence as in Daudin (1979).

References

- DAUDIN, J.J. (1979): Coefficient de Tschuprow partiel et independence conditionnelle. *Statistique et Analyse des Données*, 3, 55–58.
- DRAPER, N.R. and SMITH, H. (1966): *Applied Regression Analysis*. Wiley.
- CRAMÈR, E.M. and NICEWANDER, W.A. (1979): Some symmetric, invariant measures of multivariate association. *Psychometrika*, 44, 43–54.
- ESCOUFIER, Y. (1986): A propos du choix des variables en analyse des données. *Metron*, 44, 31–47.
- SAPORTA, G. (1976): Quelques applications des opérateurs d'Escoufier au traitement des variables qualitatives. *Bulletin de l'Association des Statisticiens Universitaires*, 38–46.
- VICARI, D. (1994): Misure di correlazione parziale generalizzata. *Statistica*, 54, 491–499.
- VICHI, M. (1989): La connessione e la correlazione tra due matrici dei dati componenti una matrice a tre indici. *Statistica*, 49, 225–243.

Statistical Models for Social Networks

Stanley Wasserman¹ and Philippa Pattison²

¹ Department of Psychology and Department of Statistics
Beckman Institute for Advanced Science and Technology
University of Illinois
603 East Daniel Street, Champaign, Illinois, 61820, USA
(email: stanwass@uiuc.edu; fax: 1-217-244-5876)

² School of Behavioural Science
Department of Psychology
University of Melbourne
Parkville, Victoria, 3052, Australia
(email: pattison@psych.unimelb.edu.au)

Abstract. Recent developments in statistical models for social networks reflect an increasing theoretical focus in the social and behavioral sciences on the interdependence of social actors in dynamic, network-based social settings (e.g., Abbott, 1997; White, 1992, 1995). As a result, a growing importance has been accorded the problem of modeling the dynamic and complex interdependencies among network ties and the actions of the individuals whom they link. Included in this problem is the identification of cohesive subgroups, or classifications of the individuals. The early focus of statistical network modeling on the mathematical and statistical properties of Bernoulli and dyad-independent random graph distributions has now been replaced by efforts to construct theoretically and empirically plausible parametric models for structural network phenomena and their changes over time.

1 Introduction

In most of the literature on statistical models for networks, the set $N = \{1, 2, \dots, n\}$ of network nodes is regarded as fixed and the edges, or ties, between nodes are assumed to be random. The tie linking node i to node j ($i, j \in N$) may be denoted by the random variable X_{ij} . In the simplest case of binary valued random variables, X_{ij} takes the value 1 if the tie is present and 0 otherwise. Other cases of interest include:

1. *valued* networks, where X_{ij} is assumed to take values in the set $\{0, 1, \dots, C - 1\}$;
2. *multiple relational* or *multivariate* networks, where the variable X_{ijk} represents the possible tie of type k from node i to node j (with $k \in R = \{1, 2, \dots, r\}$, a fixed set of *types of tie*); and
3. *time-dependent* networks, where X_{ijt} represents the tie from node i to node j at time t .

In each case, network ties may be *directed* (e.g., X_{ij} and X_{ji} are distinct random variables) or nondirected (e.g., X_{ij} and X_{ji} are not distinguished).

The $n \times n$ array $\mathbf{X} = [X_{ij}]$ of random variables can be regarded as the adjacency matrix of a *random (directed) graph* on N ; the state space of all possible realizations of these arrays may be denoted by Ω_n . The array $\mathbf{x} = [x_{ij}]$ denotes a realization of \mathbf{X} , with $x_{ij} = 1$ if there is an observed tie from node i to node j , and $x_{ij} = 0$ otherwise. In some cases, models may also refer to variables measured on actor attributes: for instance, the m^{th} attribute $Z_i^{[m]}$ of actor i gives rise to a random vector $\mathbf{Z}^{[m]}$ with realization $\mathbf{z}^{[m]}$.

2 Statistical models

2.1 Bernoulli graphs

A basic statistical model for a (directed) graph assumes a Bernoulli distribution, in which each edge, or tie, is statistically independent of all others and governed by a theoretical probability P_{ij} . In addition to edge independence, simplified versions also assume equal probabilities across ties; other versions allow the logs of probabilities to depend on structural parameters (Frank & Nowicki, 1993). These distributions have often been used as models (Bollobas, 1995), but are of questionable utility due to the independence assumption. Bernoulli graph distributions are described at length by Erdős and Rényi (1960), Erdős (1959, 1960), and Gilbert (1959), and have been used extensively by Frank (1977, 1980, 1981, 1989).

2.2 Dyadic structure in networks

Statistical models for social network phenomena have been developed from their edge-independent beginnings in a number of major ways. In one important series of developments, the p_1 model introduced by Holland and Leinhardt (1981), and developed by Wasserman and Fienberg (1981), recognized the theoretical and empirical importance of dyadic structure in social networks, that is, of the interdependence of the variables X_{ij} and X_{ji} . This class of Bernoulli dyad distributions and their generalization to valued (Wasserman & Galaskiewicz, 1984; Wasserman & Iacobucci, 1986), multivariate (Fienberg, Meyer & Wasserman, 1985; Wasserman, 1987) and time-dependent (Iacobucci & Wasserman, 1988; Wasserman, 1980; Wasserman & Iacobucci, 1988) forms, gave parametric expression to ideas of reciprocity and exchange in dyads and their development over time.

The p_1 model assumes that each dyad (X_{ij}, X_{ji}) is independent of every other and, in a commonly constrained form, specifies: $P(\mathbf{X} = \mathbf{x}) =$

$$\prod_{i < j} \exp\left\{ \lambda_{ij} + \theta\left(\sum_{i \neq j} x_{ij}\right) + \rho\left(\sum_{i < j} x_{ij}x_{ji}\right) + \alpha_i\left(\sum_j x_{ij}\right) + \beta_j\left(\sum_i x_{ij}\right) \right\}$$

where θ is a density parameter; ρ is a reciprocity parameter; the parameters α_i and β_j reflect individual differences in expansiveness and popularity; and λ_{ij} ensures that probabilities for each dyad sum to 1.

An important elaboration of the p_1 model permits the dependence of density, popularity and expansiveness parameters on a partition of the nodes into classes or subsets that can either be specified *a priori* (e.g., from hypotheses about social position) or can be identified *a posteriori* from observed patterns of node interrelationships (e.g., Anderson, Wasserman & Faust, 1992; Holland, Laskey & Leinhardt, 1983; Wang & Wong, 1986; Wasserman & Faust, 1994). The resulting *stochastic blockmodels* represent hypotheses about the interdependence of social positions and the patterning of network ties (as originally elaborated in the notion of *blockmodel* by White, Boorman & Breiger, 1976). A potentially valuable variant of these models regards the allocation of nodes to blocks as a latent classification; Snijders and Nowicki (1997) have provided Bayesian estimation methods for the case of two latent classes. Another valuable extension of the p_1 model is the so-called p_2 model (Lazega & van Duijn, 1997; van Duijn & Snijders, 1997). Like the p_1 model, the p_2 model assumes dyad-independence, but it adds the possibility of arbitrary node and dyad covariates for the parameters of the p_1 model, and models the unexplained parts of the popularity and expansiveness parameters as random rather than fixed effects.

2.3 Null models for networks

Despite the useful parameterization of dyadic effects associated with the p_1 model and its generalizations, the assumption of dyadic independence has attracted sustained criticism. Thus, another series of developments has been motivated by the problem of assessing the degree and nature of departures from simple structural assumptions like dyadic independence. Following a general approach exemplified by early work of Rapoport (e.g., Rapoport, 1949), a number of *conditional uniform* random graph distributions were introduced as null models for exploring the structural features of social networks (see Wasserman & Faust, 1994, for a historical account). These distributions, denoted by $U|Q$, are defined over subsets Q of the state space Ω_n of directed graphs, and assign equal probability to each member of Q . The subset Q is usually chosen to have some specified set of properties (e.g., a fixed number of mutual, asymmetric and null dyads, as in the $U|MAN$ distribution of Holland & Leinhardt, 1975). When Q is equal to Ω_n the distribution is referred to as the *uniform* (di)graph distribution, and is equivalent to a Bernoulli distribution with homogeneous tie probabilities. Enumeration of the members of Q and simulation of $U|Q$ is often straightforward (e.g., see Wasserman & Pattison, in press), although certain cases, such as the distribution that is conditional on the indegree ($\sum_j x_{ji}$) and outdegree ($\sum_j x_{ij}$) of each node i in the network, require more sophisticated approaches (Snijders, 1991; Wasserman, 1977). A typical application of these distributions is to assess whether the occurrence of certain higher-order (e.g., triadic) features in an observed network is unusual, given the assumption that the data arose

from a uniform distribution that is conditional on plausible lower-order (e.g., dyadic) features (e.g., Holland & Leinhardt, 1975).

This general approach has also been developed for the analysis of multivariate social networks. The best-known example is probably the Quadratic Assignment Procedure or Model (*QAP*) for networks (Hubert & Baker, 1978; see also Katz & Powell, 1953, 1957). In this case, the association between two graphs defined on the same set of nodes is assessed using a uniform multivariate graph distribution that is conditional on the unlabelled graph structure of each univariate graph. In a more general framework for the evaluation of structural properties in multivariate graphs, Pattison, Wasserman, Robins & Kanfer (in press) also introduced various multigraph distributions associated with stochastic blockmodel hypotheses.

2.4 Extra-dyadic local structure in networks

A significant step in the development of parametric statistical models for social networks was taken by Frank and Strauss (1986) with the introduction of the class of *Markov random graphs*. This class of models permitted the parameterization of extra-dyadic local structural forms and so allowed a more explicit link between some important network conceptualizations and statistical network models. Frank and Strauss (1986) adapted research originally developed for modeling spatial dependencies among observations (Besag, 1974; Mantel, 1967) to graphs. They observed that the Hammersley-Clifford theorem provides a general probability distribution for \mathbf{X} from a specification of which pairs (X_{ij}, X_{kl}) of edge random variables are conditionally dependent, given the values of all other random variables.

Define a *dependence graph* \mathbf{D} with node set $N(\mathbf{D}) = \{(X_{ij} : i, j \in N, i \neq j\}$ and edge set $E(\mathbf{D}) = \{(X_{ij}, X_{kl}) : X_{ij}$ and X_{kl} are assumed to be conditionally dependent, given the rest of $\mathbf{X}\}$. Frank and Strauss used \mathbf{D} to obtain a model for $\Pr(\mathbf{X} = \mathbf{x})$, termed p^* by Wasserman and Pattison (1996), in terms of parameters and sub-structures corresponding to cliques of \mathbf{D} . The model has the form

$$\Pr(\mathbf{X} = \mathbf{x}) = p^*(\mathbf{x}) = \left(\frac{1}{c}\right) \exp \left\{ \sum_{P \subseteq N(\mathbf{D})} \alpha_P z_P(\mathbf{x}) \right\}$$

where:

1. the summation is over all cliques P of \mathbf{D} (with a *clique* of \mathbf{D} defined as a nonempty subset P of $N(\mathbf{D})$ such that $|P| = 1$ or $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ for all $X_{ij}, X_{kl} \in P$);
2. $z_P(\mathbf{x}) = \prod_{X_{ij} \in P} x_{ij}$ is the (observed) *network statistic* corresponding to the clique P of \mathbf{D} ; and
3. $c = \sum_{\mathbf{x}} \exp\{\sum_P \alpha_P z_P(\mathbf{x})\}$ is a *normalizing* quantity.

The specification of a dependence graph raises a deep theoretical issue for network analysis: what constitutes an appropriate set of dependence assumptions? Frank and Strauss (1986) proposed a Markov assumption, in which $(X_{ij}, X_{kl}) \in E(\mathbf{D})$ whenever $\{i, j\} \cap \{k, l\} \neq \emptyset$. This assumption implies that the occurrence of a network tie from one node to another is conditionally dependent on the presence or absence of other ties in a *local neighborhood* of the tie. A Markovian local neighborhood for X_{ij} comprises all possible ties involving i and/or j . Any network tie is seen both as dependent on its neighboring ties and as participating in the neighborhood of other ties, so the resulting model embodies the assumption that social networks have a local self-organizing quality. In the case of Markovian neighborhoods, cliques in the dependence graph correspond to triadic and star-like network configurations. A homogeneity assumption, that model parameters do not depend on the specific identities of nodes but only on the network configurations to which they correspond, yields a model that expresses the probability of the network in terms of propensities for subgraphs of triadic and star-like structures to occur.

p^* random graph models permit the parameterization of many important ideas about local structure in univariate social networks, including transitivity (Holland & Leinhardt, 1970), local clustering (Wasserman & Faust, 1994), degree variability, and centralization (Wasserman & Pattison, 1996). Valued (Robins, Pattison & Wasserman, 1999) and multivariate (Pattison & Wasserman, 1999) generalizations also lead to parameterizations of substantively interesting multirelational concepts, such as those associated with balance and clusterability (Cartwright & Harary, 1979; Davis, 1967; 1979), generalized transitivity and exchange (Lazega & Pattison, 1999), and the strength of weak ties (Granovetter, 1971).

2.5 Prospects for structural models

The p^* family of random graph models poses many interesting questions, both substantive and statistical. A major empirical question is the sufficiency of these local structural descriptions: do they suffice, or do they need to accommodate other characterizations of local dependence, more global considerations, other types of constraints (e.g., temporal and spatial), or variability in local structural dependencies? On the statistical side, major questions are raised by the complexity and rich parameterization of models that are deemed to be theoretically plausible. For instance, does a Markov chain Monte Carlo maximum likelihood approach to parameter estimation (Crouch & Wasserman, 1998; Geyer & Thompson, 1992) offer a viable alternative to approximate methods such as pseudolikelihood estimation (Strauss & Ikeda, 1990)? In fact, just how good is pseudolikelihood estimation? How can these distributions best be simulated, and what are the convergence properties of various approaches, such as the Metropolis-Hastings algorithm (Strauss, 1986)?

3 Dynamic models

Empirically and theoretically defensible parametric models for networks constitute one challenge for statistical modelling, but an arguably more significant challenge is to develop models for the emergence of network phenomena, including the evolution of networks and the unfolding of individual actions (e.g., voting, attitude change, decision-making) and interpersonal transactions (e.g., patterns of communication or interpersonal exchange) in the context of longer-standing relational ties (Doreian & Stokman, 1998). Early attempts to model the evolution of networks in either discrete (Katz & Proctor, 1959; Iacobucci & Wasserman, 1988) or continuous time (e.g., Leenders, 1996a, 1996b; Wasserman, 1979, 1980) assumed dyad-independence and Markov processes in time.

A step towards continuous time Markov chain models for network evolution that relaxes the assumption of dyad-independence has been taken by Snijders (1996), Snijders and van Duijn (1997), and van de Bunt, van Duijn, and Snijders (1995, 1999). Their models incorporate both random change and change arising from individuals' attempts to optimize features of their local network environment (such as the extent to which reciprocity and balance prevail) and use the computer package *SIENA* for fitting. A tension function is proposed to represent the function that actors wish to minimize and Robbins-Monro procedures are used to estimate free parameters. The approach illustrates the potentially valuable role of simulation techniques for models that make empirically plausible assumptions; clearly, such methods provide a promising focus for future development.

The value of simulation has also been demonstrated by Watts and Strogatz (1998) and Watts (1999) in their analysis of the so-called *small world phenomenon* (the tendency for individuals to be linked by short acquaintanceship network paths). They considered a probabilistic model for network change in which the probability of a new tie was determined, in part, by existing patterns of neighboring ties (a clustering component) and, in part, by a small random component. They used simulations to establish that only a small random component was needed for the simulated network to be likely to possess the small world property.

4 Models bridging analytic levels

A recurring theme among social theorists is the need for models that integrate the dynamic, interdependent, interpersonal and cultural contexts in which social action occurs (e.g., Emirbayer, 1997; Emirbayer & Goodwin, 1995; Lindenberg, 1997). The network modelling challenge posed by such framing of the nature of social processes is the development of models that represent interdependencies across several analytic categories (e.g., individuals, relational ties, settings, groups).

Snijder's (1996) model for network change represents one step in this direction; the class of social influence models represents another (Friedkin, 1998). These models are exemplified by the network effects model for the m^{th} attribute variable:

$$\mathbf{z}^{[m]} = \alpha \mathbf{x} \mathbf{z}^{[m]} + \mathbf{y} \beta + \epsilon$$

where α is a network effects parameter, \mathbf{y} is a matrix of exogenous attribute variables with corresponding parameter vector β , and ϵ is a vector of residuals (e.g., Doreian, 1982; Marsden & Friedkin, 1994).

A further step in the direction of modelling interdependent node attributes and network ties has been to generalize the dependence graph approach to p^* network models (Section 2.4) so as to include both node attribute variables ($\mathbf{Z}^{[m]}$) and relational variables (\mathbf{X}). This step is facilitated if *directed* dependencies are permitted in the dependence graph (i.e., dependencies in which one random variable is assumed to be conditionally dependent on a second, but the second not on the first). Robins, Pattison, and Elliott (in press) drew on graphical modelling techniques (Lauritzen, 1996) to construct models with directed dependencies among network ties and individual attributes. If the dependencies are directed from network ties to node attributes, models for *social influence* are obtained; with dependencies oriented from node attributes to network ties, *social selection* models result. Indeed, this generalization introduces substantial flexibility in model construction, with potential applications to joint influence and selection models, discrete time models for network change, network path models, and group-level attribute models.

The dependence approach of Section 2.4 has also been used to construct principled models for more general multi-way relational data structures representing interdependent analytic categories. Examples to date have included bipartite (Faust & Skvoretz, 1999) and k -partite (Pattison, Mische, & Robins, 1998) graphs and illustrate how this general distribution has the capacity for a systematic and coherent approach to a variety of major network modelling challenges.

5 Other future directions

In addition to some of the specific future prospects already noted, a more general expectation is that the current classes of models is likely to be the foundation of substantial empirical and model-building activity in the near future. From their early beginnings, statistical approaches to network modeling have now reached a point where sustained interaction with empirical studies is likely to yield major insights. A particularly important benefit of such interaction is likely to be empirical research that allows the development of greatly improved models for processes occurring on networks, with important applications in many fields of social and behavioral science, including epidemiology, organizational studies, social psychology, and politics.

6 Acknowledgements

Research supported by grants from the National Science Foundation to the University of Illinois, and the Australian Research Council to the University of Melbourne.

References

- ABBOTT, A. (1997): Of time and space: the contemporary relevance of the Chicago School. *Social Forces*, **75**, 1149-1182.
- ANDERSON, C., WASSERMAN, S., & FAUST, K. (1992): Building stochastic blockmodels. *Social Networks*, **14**, 137-161.
- BESAG, J.E. (1974): Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**, 96-127.
- BOLLOBAS, B. (1985): Random Graphs. London: Academic Press. van de Bunt, G., van Duijn, M., & Snijders, T. A. B. (1995): Friendship networks and rational choice. In M. G. Everett & K. Rennolls (eds.), *Proceedings of the International Conference on Social Networks, Volume I*, London, 6th-10th July, 1995. Greenwich: University of Greenwich Press.
- CARTWRIGHT, D., & HARARY, F. (1979): Balance and clusterability: An overview. In P. W. Holland & S. Leinhardt. (eds.), *Perspectives on Social Network Research*, pages 25-50. New York: Academic Press.
- CROUCH, B., & WASSERMAN, S. (1998): Fitting p^* : Monte Carlo maximum likelihood estimation. Paper presented at International Conference on Social Networks, Sitges, Spain, May 28-31.
- DAVIS, J. A. (1967): Clustering and structural balance in graphs. *Human Relations*, **20**, 181-187.
- DAVIS, J.A. (1979): The Davis/Holland/Leinhardt studies: An overview. In P. W. Holland & S. Leinhardt. (eds.), *Perspectives on Social Network Research*, pages 51-62. New York: Academic Press.
- DOREIAN, P. (1982): Maximum likelihood methods for linear models. *Sociological Methods & Research*, **10**, 243-269.
- DOREIAN, P., & STOKMAN, F. (1997): *Evolution of Social Networks*. Amsterdam: Gordon & Breach.
- EMIRBAYER, M. (1997): Manifesto for a relational sociology. *American Journal of Sociology*, **103**, 281-317.
- EMIRBAYER, M., & Goodwin, J. (1994): Network analysis, culture, and the problem of agency. *American Journal of Sociology*, **99**, 1411-1454.
- ERDÖS, P. (1959): Graph theory and probability, I. *Canadian Journal of Mathematics*, **11**, 34-38.
- ERDÖS, P. (1961): Graph theory and probability, II. *Canadian Journal of Mathematics*, **13**, 346-352.
- ERDÖS, P., & RÉNYI, A. (1960): On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **5**, 17-61.
- FAUST, K., & SKVORETZ, J. (1999): Logit models for affiliation networks. In M. Becker, & M. Sobel (eds.), *Sociological Methodology 1999*, pages 253-280. New York: Basil Blackwell.

- FIENBERG, S. E., MEYER, M. M., & WASSERMAN, S. (1985): Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, **80**, 51-67.
- FIENBERG, S., & WASSERMAN, S. (1981): Categorical data analysis of single sociometric relations. In S. Leinhardt (ed.), *Sociological Methodology 1981*, pages 156-192. San Francisco: Jossey-Bass.
- FRANK, O. (1977): Estimation of graph totals. *Scandinavian Journal of Statistics*, **4**, 81-89.
- FRANK, O. (1980): Sampling and inference in a population graph. *International Statistical Review*, **48**, 33-41.
- FRANK, O. (1981): A survey of statistical methods for graph analysis. In S. Leinhardt (ed.), *Sociological Methodology 1981*, pages 110-155. San Francisco: Jossey-Bass.
- FRANK, O. (1989): Random graph mixtures. *Annals of the New York Academy of Science*. 576: *Graph Theory and its Applications*, pages 192-199. New York: East and West.
- FRANK, O., & NOWICKI, K. (1993): Exploratory statistical analysis of networks. In J. Gimbel, J. W. Kennedy, & L. V. Quintas (eds.), *Quo Vadis Graph Theory? A Source Book for Challenges and Directions*. Amsterdam: North-Holland. (also *Annals of Discrete Mathematics*, **55**, 349-366.)
- FRANK, O., & STRAUSS, D. (1986): Markov graphs. *Journal of the American Statistical Association*, **81**, 832-842.
- FRIEDKIN, N. (1998): *A Structural Theory of Social Influence*. New York: Cambridge University Press.
- GEYER, C., & THOMPSON, E. (1992): Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B*, **54**, 657-699.
- GILBERT, E. N. (1959): Random graphs. *Annals of Mathematical Statistics*, **30**, 1141-1144.
- GRANOVETTER, M. (1973): The strength of weak ties. *American Journal of Sociology*, **78**, 1360-1380.
- HOLLAND, P. W., LASKEY, K. B., & LEINHARDT, S. (1983): Stochastic blockmodels: some first steps. *Social Networks*, **5**, 109-137.
- HOLLAND, P. W., & LEINHARDT, S. (1970): A method for detecting structure in sociometric data. *American Journal of Sociology*, **70**, 492-513.
- HOLLAND, P. W., & LEINHARDT, S. (1975): The statistical analysis of local structure in social networks. In D. R. Heise (ed.), *Sociological Methodology 1976*, pp. 1-45. San Francisco: Jossey-Bass.
- HOLLAND, P. W., & LEINHARDT, S. (1981): An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, **76**, 33-50.
- HUBERT, L. J., & BAKER, F. B. (1978): Evaluating the conformity of sociometric measurements. *Psychometrika*, **43**, 31-41.
- IACOBucci, D., & WASSERMAN, S. (1988): A general framework for the statistical analysis of sequential dyadic interaction data. *Psychological Bulletin*, **103**, 379-390.
- KATZ, L., & POWELL, J. H. (1953): A proposed index of conformity of one sociometric measurement to another. *Psychometrika*, **18**, 249-256.

- KATZ, L., & POWELL, J. H. (1957): Probability distributions of random variables associated with a structure of the sample space of sociometric investigations. *Annals of Mathematical Statistics*, **28**, 442-448.
- KATZ, L., & PROCTOR, C.H. (1959). The concept of configuration of interpersonal relations in a group as a time-dependent stochastic process. *Psychometrika*, **24**, 317-327.
- LAURITZEN, S. (1996): *Graphical Models*. Oxford: Oxford University Press.
- LAZEGA, E. and VAN DUIJN, M. (1997): Position in formal structure, personal characteristics and choices of advisors in a law firm: A logistic regression model for dyadic network data. *Social Networks*, **19**, 375-397.
- LAZEGA, E.,& PATTISON, P. (1999): Multiplexity, generalized exchange and co-operation in organizations. *Social Networks*, **21**, 67-90.
- LEENDERS, R. (1995): Models for network dynamics: A Markovian framework. *Journal of Mathematical Sociology*, **20**, 1-21.
- LEENDERS, R. (1996): Evolution of friendship and best friendship choices. *Journal of Mathematical Sociology*, **21**, 133-148.
- LINDENBERG, S. (1997): Grounding groups in theory: functional, cognitive and structural interdependencies. *Advances in Group Processes*, **14**, 281-331.
- MANTEL, N. (1967): The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209-220.
- MARSDEN, P., & FRIEDKIN, N. (1994): Network studies of social influence. In S. Wasserman & J. Galaskiewicz (eds.), *Advances in Social Network Analysis* (pages 3-25). Thousand Oaks, CA: Sage.
- PATTISON, P., MISCHE, A., & ROBINS, G.L. (1998): The plurality of social relations: k -partite representations of interdependent social forms. Keynote address, Conference on Ordinal and Symbolic Data Analysis, Amherst, Sept. 28-30.
- PATTISON, P. E., & WASSERMAN, S. (1999): Logit models and logistic regressions for social networks, II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, **52**, 169-194.
- PATTISON, P., WASSERMAN, S., ROBINS, G.L., & KANFER, A.M. (in press): Statistical evaluation of algebraic constraints for social networks. *Journal of Mathematical Psychology*.
- RAPOPORT, A. (1949): Outline of a probabilistic approach to animal sociology, I. *Bulletin of Mathematical Biophysics*, **11**, 183-196.
- ROBINS, G.L., PATTISON, P., & ELLIOTT, P. (in press): Network models for social influence processes. *Psychometrika*.
- ROBINS, G.L., PATTISON, P., & WASSERMAN, S. (1999): Logit models and logistic regressions for social networks, III. Valued relations. *Psychometrika*, **64**, 371-394.
- SNIJDERS, T. A. B. (1991): Enumeration and simulation methods for 0-1 matrices with given marginals. *Psychometrika*, **56**, 397-417.
- SNIJDERS, T. A. B. (1996): Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, **21**, 149-172.
- SNIJDERS, T. A. B., & NOWICKI, K. (1997): Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, **14**, 75-100.
- SNIJDERS, T.A.B., & VAN DUIJN, M.A.J. (1997). Simulation for statistical inference in dynamic network models. In R. Conte, R, Hegselmann, & P. Terna (eds.), *Simulating Social Phenomena* (pages 493-512). Berlin: Springer-Verlag.

- STRAUSS, D. (1986): On a general class of models for interaction. *SIAM Review*, **28**, 513-527.
- STRAUSS, D., & IKEDA, M. (1990): Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, **85**, 204-212.
- VAN DE BUNT, G., VAN DUIJN, M., & SNIJDERS, T. A. B. (1999): Friendship networks through time: an actor-oriented dynamic statistical network model. *Computational and Mathematical Organization Theory*, **5**, 167-192.
- VAN DUIJN, M., & SNIJDERS, T. A. B. (1997): p_2 : a random effects model for directed graphs. Unpublished manuscript.
- WANG, Y. Y., & WONG, G. Y. (1987): Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, **82**, 8-19.
- WASSERMAN, S. (1977): Random directed graph distributions and the triad census in social networks. *Journal of Mathematical Sociology*, **5**, 61-86.
- WASSERMAN, S. (1979): A stochastic model for directed graphs with transition rates determined by reciprocity. In K. Schuessler (ed.) *Sociological Methodology 1980*, pages 392-412. San Francisco: Jossey-Bass.
- WASSERMAN, S. (1980): Analyzing social networks as stochastic processes. *Journal of the American Statistical Association*, **75**, 280-294.
- WASSERMAN, S. (1987): Conformity of two sociometric relations. *Psychometrika*, **52**, 3-18.
- WASSERMAN, S., & FAUST, K. (1994): *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press.
- WASSERMAN, S., & GALASKIEWICZ, J. (1984): Some generalizations of p_1 : External constraints, interactions, and non-binary relations. *Social Networks*, **6**, 177-192.
- WASSERMAN, S., & IACOBUCCI, D. (1986): Statistical analysis of discrete relational data. *British Journal of Mathematical and Statistical Psychology*, **39**, 41-64.
- WASSERMAN, S., & IACOBUCCI, D. (1988): Sequential social network data. *Psychometrika*, **53**, 262-282.
- WASSERMAN, S., & PATTISON, P. E. (1996): Logit models and logistic regressions for social networks, I. An introduction to Markov random graphs and p^* . *Psychometrika*, **60**, 401-425.
- WASSERMAN, S., & PATTISON, P. E. (in press): *Multivariate Random Graph Distributions*. Springer Lecture Note Series in Statistics.
- WATTS, D. (1999): Networks, dynamics, and the small-world phenomenon. *American Journal of Sociology*, **105**, 493-527.
- WATTS, D., & STROGATZ, S. (1998): Collective dynamics of 'small-world' networks. *Nature*, **393**, 440-442.
- WHITE, H. C. (1992): *Identity and Control*. Princeton, NJ: Princeton University Press.
- WHITE, H. C. (1995): Network switchings and Bayesian forks: reconstructing the social and behavioral sciences. *Social Research*, **62**, 1035-1063.
- WHITE, H. C., BOORMAN, S., & BREIGER, R. L. (1976): Social structure from multiple networks, I. Blockmodels of roles and positions. *American Journal of Sociology*, **81**, 730-780.

Application of Simulated Annealing in some Multidimensional Scaling Problems

Javier Trejos¹, William Castillo², Jorge González³, and Mario Villalobos⁴

¹ PIMAD-CIMPA, School of Mathematics
University of Costa Rica, 2060 San José, Costa Rica
(e-mail: jtrejos@cariari.ucr.ac.cr)

² Same address, (e-mail: wcastill@cariari.ucr.ac.cr)

³ Same address, (e-mail: jgonzale@cariari.ucr.ac.cr)

⁴ Same address, (e-mail: mvillalo@emate.ucr.ac.cr)

Abstract. We apply simulated annealing as a combinatorial optimization heuristic in some multidimensional scaling (MDS) contexts for the minimization of Stress: metric MDS, MDS with restrictions in the configuration and INDSCAL parameter estimation. The application of this technique is based on a discretization of the representation space by a grid. Results obtained are compared to those of usual well-known algorithms and are shown to be better in most of the cases.

1 Introduction

Given a set of dissimilarities between n objects, multidimensional scaling (MDS) intends to find a representation of the objects in a low dimensional vector space \mathbb{R}^p , such that the dissimilarities are represented by Euclidean distances in \mathbb{R}^p . There are many MDS methods for different contexts, depending on some additional information on the dissimilarities or the objects. For instance, if the dissimilarities are Euclidean then Classical Scaling can be applied by the diagonalization of an inner product matrix and the representation space has a dimension equal to the number of non-zero eigenvalues. In metric scaling, values of the dissimilarities are important and the Stress

$$\sigma(X) = \sum_{i < j} [\delta_{ij} - d_{ij}(X)]^2$$

is to be minimized, where (δ_{ij}) is the $n \times n$ matrix of dissimilarities, X is the $n \times p$ matrix of representation in \mathbb{R}^p and $d_{ij}(X)$ are the Euclidean distances in \mathbb{R}^p between objects. In nonmetric MDS only dissimilarity ranks are important. For classical, metric and nonmetric MDS there are well known methods in the literature (see Borg and Groenen (1997)).

One can add some information to the dissimilarities. For instance, there can be some restrictions, linear or not, on the configuration, given by a specialist. In this case, restrictions should be added to the criterion. In case that dissimilarities are given in several tables, indexed for example by time or different judges, then there are m dissimilarity tables (δ_{ij}^k) , each of dimension

$n \times n$, with $k \in \{1, \dots, m\}$. Depending on the way to analyze these tables, one can state different kinds of methods, such as INDSCAL or IDIOSCAL.

Generally, MDS methods are based on optimization techniques and heuristics. This is the case for Kruskal's method, SMACOF, tunneling (see Groenen and Heiser (1996)) or distance smoothing (Groenen et al. (1999)) in metric MDS, CANDECOMP (Carroll and Chang (1970)) or SYMPRES (Ten Berge et al. (1993)) in the estimation of parameters for INDSCAL, and the procedure used in MDS with restrictions in the configuration based on SMACOF. Random optimization techniques, such as simulated annealing (SA) and genetic algorithms, can also be used as heuristics of optimization; they have the important property of asymptotic convergence to a global minimum using a Markov chain modelling. Machmouchi (1992) has applied SA in non-metric MDS for the search of an initial configuration. We have applied SA in metric MDS, MDS with restrictions on the configuration and INDSCAL. In some other data analysis problems we obtained good results by the use of SA (Trijos et al. (1998), Trijos and Castillo (2000)). De Soete et al. (1988) applied SA to unidimensional scaling (that is, when $p = 1$), with poor results, probably because the length of the associated Markov chains was not enough; also, Hubert and Arabie (1986) applied dynamic programming in a method that guarantees to reach the global minimum but that can handle only small data sets, and Groenen (1993) used tabu search, with the same results as methods based on permutations, such as LOPI.

SA is based on the annealing technique of mechanical statistics for construction of "pure" crystals: the material is heated at a very high temperature, and then cooled very slowly. Annealing has the feature that energy can be higher even if the temperature decreases. In combinatorial optimization SA is a general metaheuristic based on the Metropolis' rule: a new state of the combinatorial optimization problem is accepted if the cost function σ decreases, otherwise it is accepted with probability $\exp(-\Delta\sigma/c_t)$, where c_t is an external parameter that plays the role of temperature and controls the acceptance of new states that do not improve the criterion. The algorithm is modelled with Markov chains and asymptotic convergence is proved if some conditions are fulfilled. A finite-time implementation of SA requires that these asymptotic convergence conditions are satisfied, as well as the definition of a *cooling program*: initial and final values of c_t , definition of the (slow) decreasing of c_t and the length of the Markov chains associated with each value of c_t . Moreover, it should use a simplified computation of $\Delta\sigma$. See Aarts and Korst (1989) for full details on SA.

2 Metric multidimensional scaling

Our application of SA in metric MDS, called SAMSCAL, is based on a discretization of the representation space \mathbb{R}^p by a grid of width $h \in \mathbb{R}$. A state is a set of n points in the grid and a move is defined by the choice (uni-

formly) at random of a point and a direction. That is, the neighborhood of state $I = x_1^I, \dots, x_n^I$, with $x_i^I \in (h\mathbb{Z})^p$, is the set of $2np$ configurations of n points such that one of the points is perturbed in h in one of its coordinates. The Metropolis rule is applied for the acceptation of these new states and the usual SA algorithm, with an appropriate cooling program.

We have applied the SAMSCAL algorithm on many examples, for comparing its performance with respect to the best known methods. We define the *attraction rate* τ of a local minimum of Stress as the percentage of cases in which this local minimum is found when using many random starts. To facilitate comparison of our results to those elsewhere in the literature, we report *normalized Stress*, σ/η_δ^2 (with η_δ^2 the sum of the squared dissimilarities), because at a local minimum it is always between zero and one and it is equal to the square of Kruskal's Stress-1 (see Borg and Groenen (1997)). We denote by σ_{opt} the best value of σ found with the program SAMSCAL and $\#$ the number of times that the algorithm was applied. All applications reported here are for representations in \mathbb{R}^2 .

First, we consider a set of 4 points such that $\delta_{ij} = 0$ if $i = j$ and $\delta_{ij} = 1$ otherwise. For $\# = 80$ we found $\sigma_{opt}/\eta_\delta^2 = 0.028595479$ with $\tau = 100\%$, while the same value was found with SMACOF by De Leeuw (1988).

Second, we consider 9 points in \mathbb{R}^2 that form a squared grid with 3 points equally spaced in each side, and we compute the usual Euclidian distance between the 9 points. For $\# = 300$, we found $\sigma_{opt}/\eta_\delta^2 = 0.0$ with $\tau = 95.47\%$, the same best value found by Groenen (1993) but with $\tau = 72.0\%$.

We also considered the colas data set (as it appears in Groenen (1993), p. 135) that crosses 10 beverages and contains the similarities between each couple of cola brands, according to an experiment with 38 students. We found $\sigma_{opt}/\eta_\delta^2 = 0.0367837933$ and $\tau = 17.74\%$ for $\# = 310$, while Groenen et al. (1995) found a better value of Stress 0.03678052, which is improved to 0.03678033 in Groenen et al. (1999) using distance smoothing. We suppose that the error in SAMSCAL is due to the discretization of \mathbb{R}^2 . However, it should be noted that the mean error that we found is 0.0394416, while Groenen et al. (1995) found 0.04145104, 0.04070994 and 0.04113443, with SMACOF, relaxed SMACOF and the KYST version of Kruskal's method, respectively. Mathar (1995) found the same value as Groenen et al. (1995) but with $\tau = 100\%$; the genetic algorithm uses a crossover based on convex combinations of representations and SMACOF descent after a number of iterations.

The final example, is the one considered in Mathar and Žilinskas (1993), which consists on the Euclidean distances between 10 points generated randomly in $[-1, 1]^5$. Using the same data set, we found $\sigma_{opt}/\eta_\delta^2 = 0.036026404$ with $\tau = 37.27\%$ and a mean of 0.0401566, while Mathar and Žilinskas (1993) found $\sigma_{opt}/\eta_\delta^2 = 0.036162536$ with $\tau = 15.6\%$ and a mean of 0.0417546.

3 Multidimensional scaling with restrictions on the configuration

In MDS it is possible to add to the knowledge on the similarities between the objects, supplementary restrictions that are desired on the searched configuration with MDS. These restrictions arise from important considerations on the properties that the data must have.

MDS with Restrictions on the configuration (MDSR) aims to minimize the Stress $\sigma(X)$ subject to some restrictions on the matrix X . De Leeuw and Heiser (1980) propose an algorithm based on SMACOF for MDSR. We apply SA using the same discretization as in metric MDS, in the case of linear restrictions, that is $X = YC$, where Y is a given $n \times q$ matrix and C is a $q \times p$ matrix to be determined.

Another utility of MDSR is that it can help to confirm hypothesis that are supposed to be satisfied by the data, which can be verified if the MDS and the MDSR Stresses are equal. This can be illustrated using the “Facial Expressions” data table given by Borg and Groenen (1997), that contains the observed similarities between 13 face expressions and as restrictions some theoretical ranks for emotional messages given by specialists. We found the same Stress value of 0.045 using SA and SMACOF, for C given by the transposed vectors $(1.036, 0.275, 0.233)$ and $(-0.886, 0.391, 0.589)$. In some other data sets generated at random, we also obtained the same results with both methods.

4 Parameter estimation of INDSCAL

Given m matrices B_1, \dots, B_m of approximate scalar products, the INDSCAL model proposed by Carroll and Chang (1970), estimates an $n \times p$ matrix X and m diagonal matrices W_1, \dots, W_m of nonnegative weights, such that the following function f is minimum:

$$f(X, W_1, \dots, W_m) = \sum_{k=1}^m \|B_k - XW_kX^t\|^2.$$

In the literature, several procedures have been proposed for the minimization of f . Carroll and Chang (1970) proposed the CANDECOMP algorithm, which consists in the minimization of the function:

$$g(X, Y, W_1, \dots, W_m) = \sum_{k=1}^m \|B_k - XW_kY^t\|^2$$

using alternating least squares. This procedure has three disadvantages: finding the global minimum is not guaranteed, it can find negative weights and a diagonal matrix D such that $X = YD$ may not exist (this is known as the

symmetry problem), which is a necessary condition for the minimizing matrix X of g to also be a minimizing matrix of f . To overcome these last two difficulties, Ten Berge et al. (1993) proposed an algorithm called SYMPRES, obtaining the same results as CANDECOMP on two sets of 100 matrices.

The use of SA and a discretization of the representation space (as in metric MDS), which we call ssINDS, solves the symmetry problem and the non-negativeness of the weights. Let X be a $n \times p$ configuration and W a $m \times p$ matrix whose rows are nonnegative weights of INDSCAL model. A state is a $(n+m) \times p$ matrix that has the rows of X over those of W . The current state is perturbed by the random choice of X or W (uniformly), the random choice of a direction, and the random choice of a row in X (with probability $1/n$) or in W (with probability $1/m$).

We compared ssINDS and CANDECOMP on matrices generated as in Ten Berge et al. (1993). We denote the m matrices of size $n \times n$ as $n \times n \times m$. We generated at random 20 sets of matrices $3 \times 3 \times 3$ and 20 sets of matrices $6 \times 6 \times 9$. Each matrix B_k was constructed by generating a matrix A_k with uniform random entries in $[-1, 1]$, computing $B_k = A_k A_k^t$, and verifying that it is positive definite. The fitting quality is measured as the percentage of the sum of squares in the data. That is,

$$perc = 1 - \frac{\sum_k \sum_{i \leq j} (B_{kij} - X_i W_k X_j^t)^2}{\sum_k \sum_{i \leq j} B_{kij}^2}.$$

ssINDS and CANDECOMP were executed 10 times for each set of matrices and the best solution was chosen accordingly with the greater value of $perc$. Then it was calculated the mean of these 20 sets of matrices. The mean fitting $perc$ values obtained with ssINDS are 90.73% for a $3 \times 3 \times 3$ matrix and 49.27% for a $6 \times 6 \times 9$ matrix, while for CANDECOMP they are 88.60% and 49.97%, respectively.

5 Concluding remarks

The formulation of multidimensional scaling algorithms with SA can give interesting results, comparable to those of known methods, and sometimes better. The errors due to the discretization can be overcome with thinner grids. A full comparison with other methods should be made, as well as a deep investigation on the dependence on the SA parameters, specially on the length of the Markov chains associated to each value of the control parameter; indeed, a small value can lead to suboptimal solutions. The SA method in the estimation of INDSCAL parameter solves the problem of non-negative weights. This method can also be easily adapted to the use of dissimilarity matrices without considering approximate scalar products, by the minimization of a least-squares Stress function that compares the k -th dissimilarity matrix, δ_k , and $d(X, W_k)$, the distance matrix between the rows of X , with weights defined by W_k .

References

- AARTS, E. and KORST, J. (1989): *Simulated Annealing and Boltzmann Machines. A Stochastic Approach to Combinatorial Optimization and Neural Computing*. John Wiley & Sons, Chichester.
- BORG, I. and GROENEN, P.J.F. (1997): *Modern Multidimensional Scaling*. Springer, New York.
- CARROLL, J.D. and CHANG, J.J. (1970): Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart-Young decomposition. *Psychometrika*, 35, 283–319.
- DE LEEUW, J. and HEISER, W. (1980): Multidimensional scaling with restrictions on the configuration. In: P.R. Krishnaiah (Ed.): *Multivariate Analysis*, North-Holland, Amsterdam, 501–522.
- DE LEEUW, J. (1988): Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, 5, 163–180.
- DE SOETE, G., HUBERT, L. and ARABIE, P. (1988): On the use of simulated annealing for combinatorial data analysis. In: W. Gaul and M. Schader (Eds.): *Data Expert, Knowledge and Decisions*. Springer, Berlin, 329–340.
- GROENEN, P.J.F. (1993): *The Majorization Approach to Multidimensional Scaling*. DSWO Press, Leiden.
- GROENEN, P.J.F. and HEISER, W.J. (1996): The tunneling method for global optimization in multidimensional scaling. *Psychometrika*, 61, 529–550.
- GROENEN, P.J.F.; HEISER, W.J. and MEULMAN, J.J. (in press): Global optimization in least-squares multidimensional scaling by distance smoothing. To appear in *Journal of Classification*.
- GROENEN, P.J.F., MATHAR, R. and HEISER, W.J. (1995): The majorization approach to multidimensional scaling for Minkowski distances. *Journal of Classification*, 12, 3–19.
- HUBERT, L. and ARABIE, P. (1986): Unidimensional scaling and combinatorial optimization. In: J. De Leeuw, W.J. Heiser, J. Meulman and F. Critchley (Eds.): *Multidimensional Data Analysis*. DSWO Press, Leiden, 181–196.
- MACHMOUCHI, M. (1992): *Contributions à la Mise en Œuvre des Méthodes d'Analyse des Données de Dissimilarité*. Thèse de Doctorat, Université de Grenoble II.
- MATHAR, R. (1995): A genetic algorithm for multidimensional scaling. Internal Report, RWTH, Aachen.
- MATHAR, R. and ŽILINSKAS, A. (1993): On global optimization in multidimensional scaling. *Acta Aplicandae Mathematicae*, 33, 109–118.
- TEN BERGE, J.M.F., KIERS, H.A.L. and KRIJNEN, W.P. (1993): Computational solutions for the problem of negative saliences and nonsymmetry in INDSCAL. *Journal of Classification*, 10, 115–124.
- TREJOS, J., MURILLO, A. and PIZA, E. (1998): Global stochastic optimization for partitioning. In: A. Rizzi, M. Vichi and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer, Heidelberg, 185–190.
- TREJOS, J. and CASTILLO, W. (in press): Simulated annealing optimization for two-mode partitioning. To appear in: W. Gaul and R. Decker (Eds.): *Classification and Information Processing at the Turn of the Millennium*. Springer, Heidelberg.

Data Analysis Based on Minimal Closed Subsets

S. Bonnevay and C. Largeron-Leteno

Université Claude Bernard Lyon1,
L.A.S.S. - bât. 101, 43 bd du 11/11/18,
69622 Villeurbanne Cedex, France
(e-mail: bonnevay@univ-lyon1.fr largeron@univ-st-etienne.fr)

Abstract. The aim of this paper is to provide a framework which enables us to treat structural analysis problems. This framework is based on pretopological theory. We apply the concepts of pseudoclosure and minimal closed subsets to bring out the structural information. In order to illustrate our method, an application to co-authorships of publications between French geographical areas is displayed.

1 Introduction

Pretopological closed subsets are the main actors of this paper. Indeed, the goal of our study is to extract structural information of a space E according to a pretopological pseudoclosure (see section 2.1). This pseudoclosure is defined from the connections existing between elements of the population E . As in topology, pretopological closed subsets can be defined on E . Some of these closed subsets, called minimal and elementary ones, are used to highlight groups of homogenous elements. The structural method first computes smallest closed subsets, then those which contain them, and so on, until structural analysis of the entire population has been completed (see section 2.2).

In this paper, we present some data of co-authorships of publications between French departments which are French administrative geographical areas. These data are used to apply our method to the space E of the French departments. The aim of this study is to show relations between departments according to co-authorships of scientific publications.

We first develop the pretopological concepts and the structural method in section 2. Then we expose results of this structuring process on a part of co-authorships of publications data in section 3.

2 Mathematical tools and method

2.1 Pretopological concepts

In our study and in many other problems of discrete nature spaces, we need to endow a non-metric space with a topological structure. In this way, the mathematical theory used is called Pretopology. It is an extension of topology

which has been developed in view to study discrete problems (Auray 1982, Bonnevay 1999, Duru 1980, Lamure 1987, Largeron 1997). Like Čech and Dikranjan, in their topology books (Čech 1966, Dikranjan 1995), axiomatic of topology is weakened, in particular idempotency of closure operators isn't assumed. In this section, a small introduction of these pretopological notions follows. A complete description can be found in Belmandt 1993.

Let E be a finite non empty set. Let $\mathcal{P}(E)$ be the family of subsets of E . According to such discrete system problems, some relations exist between elements of E . These relations are not necessary metric ones, they can describe some notions of influence, intensity or anything else. From each significant relation r_i , a notion of neighborhood $B_i(x)$ can be defined for each x of E :

$$B_i(x) = \{y \in E \mid xr_iy\} \cup \{x\}$$

And then, from these $B_i(x)$, we can define the family $V(x)$ of the neighborhood of x by:

$$V(x) = \{V \in \mathcal{P}(E) \mid \exists i, B_i(x) \subset V\}$$

From this family $V(x)$, which is a prefilter of subsets of E , a mapping $a(\cdot)$ can be defined from $\mathcal{P}(E)$ to $\mathcal{P}(E)$ such as:

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E \mid \forall V \in V(x), V \cap A \neq \emptyset\} \text{ or also}$$

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E \mid \forall i, B_i(x) \cap A \neq \emptyset\}$$

The mapping $a(\cdot)$ is called a *pseudoclosure* on E . The space E endowed with a pseudoclosure $a(\cdot)$ is called a *V-pretopological space*. It's noted (E, a) .

Indeed, a V-pretopological space is also defined by:

- $a(\emptyset) = \emptyset$,
- $\forall A \in \mathcal{P}(E), A \subset a(A)$,
- $\forall A, B \in \mathcal{P}(E), A \subseteq B \implies a(A) \subseteq a(B)$.

There exists many other types of pretopological spaces according to other properties of $a(\cdot)$ (Belmandt 1993).

One of the main differences between this pseudoclosure and a topological closure is that the set $a(a(A))$ is not always equal to $a(A)$. According to this property, it's possible to apply a mapping $a(\cdot)$ on a set A , in sequence:

$$A \subseteq a(A) \subseteq a^2(A) \subseteq a^3(A) \subseteq \dots$$

These successive spreadings of A can model expansions corresponding to different types of phenomena (dilation, propagation, influence, ...).

Let (E, a) be a pretopological space, a part A of E for which $a(A) = A$ is called a *closed subset* of E .

Let \mathcal{F} be the family of closed subsets of E . We also define the closure $F(A)$ of a subset A of E by:

$$F(A) = \bigcap \{F \in \mathcal{F} \mid A \subset F\}$$

It's easy to prove that in a finite space E , $F(A)$ can be obtained by successive pseudoclosure $a(\cdot)$ in the sense that:

$$\exists k \leq \|E\| \mid F(A) = a^k(A)$$

In this work, we use two particular kinds of closed subsets of E which are called:

- *minimal closed subsets* of E , noted \mathcal{F}_m :

$$\mathcal{F}_m = \{F \in \mathcal{F}, \forall G \in \mathcal{F} \text{ such that } G \subset F \text{ and } G \neq F\}$$

- elementary closed subsets of E , noted \mathcal{F}_e :

$$\mathcal{F}_e = \{F(x), x \in E\}$$

We can prove that there exist only 3 kinds of relations between elements of \mathcal{F}_e : either the elementary closed subsets are separated, either one is included in the other one, or their intersection contains elements whose elementary closure is included in this intersection:

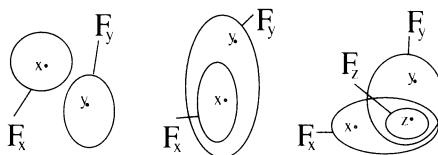


Fig. 1. Relations between elementary closed subsets.

One of the most important properties we use in this paper is that:

$$F \in \mathcal{F}_m \Leftrightarrow F \in \mathcal{F}_e \text{ and } F \text{ is minimal in terms of inclusion in } \mathcal{F}_e. \quad (1)$$

The proof of this property can be found in Belmandt 1993 or Bonnevay 1999. It means that it's not worth searching the minimal closed subsets in the whole parts of E , it is sufficient to find them in \mathcal{F}_e .

2.2 Method

The aim of the method is to bring out the structure of a space E according to the pseudoclosure $a(\cdot)$ defined on E . In this way, closed subsets are of particular interest within the context of structural analysis. They enable a representation of the homogenous subsets of E in regard to the pseudoclosure retained. Indeed, there are significant connections in regard to $a(\cdot)$ between elements of a closed subset F , and there are no significant connections between these elements and those outside of F . Thus, in view to extract structural information of a space E , we display relations between some closed subsets with the structural algorithm described in Bonnevay 1999.

This algorithm is divided into 3 parts:

- It builds the set of elementary closed subsets \mathcal{F}_e .
- Then, it determines the minimal closed subsets \mathcal{F}_m ; $F \in \mathcal{F}_m$ means that it's the most homogeneous subset of E in regard to $a(\cdot)$. According to the property (1) these closed subsets are found in \mathcal{F}_e .
- At last, a recursive process displays an inclusion relation of the sets \mathcal{F}_e by means of successive enlargements from each minimal closed subset.

3 Co-authorships of publications structural analysis

In order to illustrate our method, we present an example where the elements are French departments : $E = \{1, \dots, n\}$, with $n=89$ (Massard 1999,

Autant 1999). We study co-authorships of scientific publications between these departments. The data come from OST (Observatoire des Sciences et Techniques) and are extracted from the Science Citation Index. For each year and each technological field (chemistry, biology, ...) we have a matrix $C = [c_{xy}]_{x,y \in \{1, \dots, n\}}$, where c_{xy} gives the number of co-authored publications written by at least one author belonging to department x and at least one author belonging to department y .

The number of co-authorings in which at least one author of department x has participated is noted $c_x = \sum_{y=1, \dots, n} c_{xy}$. The total number of co-authorships for the whole country is $c..$, defined by $c.. = \sum_{x=1, \dots, n} c_{x..}$.

There are of course many manners in which a pretopological structure on E can be constructed from C . Here, we present two examples.

A pseudoclosure $a(A)$ of a subset A of E can be defined by:

$$\forall A \in \mathcal{P}(E), a(A) = \{x \in E \mid B(x) \cap A \neq \emptyset\} \quad (2)$$

where $B(x)$ is the set of departments with which x mainly publish.

$$\text{For example, } B(x) = \{y \in E \mid c_{xy} = \max\{c_{xz}, z \neq x\}\} \cup \{x\} \quad (3)$$

$$\text{or } B(x) = \{y \in E \mid c_{xy} > (c_x \cdot c_y)/c..\} \cup \{x\} \quad (4)$$

Results given below, have been obtained with (3). In that case, the structuring process provides well known results of graph theory when the connection between the elements is the relation of the successors. For all elements x of a subset $E' = \{26, 31, 35, 37, 38, 47, 56, 61, 64, 70, 73, 75, 94\}$ of E , the following table gives the set $B(x)$, the pseudoclosure $a(x)$ and the closure $F(x)$ (minimal closed subsets are designated by an asterisk *).

x	Example 1		Example 2	
	$B(x)$	$a(x)$	$F(x)$	$F(x)$
26	{26, 38}	{26}	{26}* {26, 38, 47, 73}	{26}* {26, 38, 47}
31	{31, 75}	{31, 64, 70}	{31, 61, 64, 70}	{31, 70}
35	{35, 75}	{35, 56}	{35, 56}	{35, 56}
37	{37, 75}	{37, 61}	{37, 61}	{37, 61}
38	{38, 75}	{26, 38, 47, 73}	{26, 38, 47, 70, 73}	{26, 38, 47}
47	{38, 47, 75}	{47}	{47}* {26, 38, 47, 70, 73}	{47}* {26, 38, 47}
56	{35, 56}	{56}	{56}* {26, 38, 47, 70, 73}	{56}* {26, 38, 47}
61	{37, 61, 64}	{61}	{61}* {26, 38, 47, 70, 73}	{61}* {26, 38, 47}
64	{31, 64}	{61, 64}	{61, 64} {26, 38, 47, 70, 73}	{61, 64} {26, 38, 47}
70	{31, 70, 73}	{70}	{70}* {26, 38, 47, 70, 73}	{70}* {26, 38, 47}
73	{38}	{70, 73}	{70, 73}	{70, 73}
75	{75, 94}	{31, 35, 37, 38, 47, 75, 94}	E'	{26, 47, 75, 94}
94	{75, 94}	{75, 94}	E'	{94}

By definition (3) of B , a department x belongs to the pseudoclosure $a(A)$ of a set of departments A , if and only if x has mainly published with at least a department of A . $F(x)$ is the set of departments which have mainly published either with x directly or with other elements which have mainly published with x directly or not.

The next step aims at defining the inclusion relation on the set of elementary closed subsets (\mathcal{F}_e, \subset) . An extract of results obtained on publications in biology (1997) is illustrated in the figure 2. An elementary closed subset

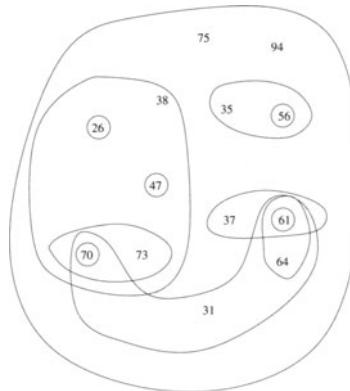


Fig. 2. Extract of result given by the algorithm for Biology in 1997.

$F(x)$ reduced to a one-element set $\{x\}$ corresponds to a department such that $F(x) = a(\{x\}) = \{x\}$. This means there does not exist any department, except x itself, which has mainly published with x . This refers for example to elements 26, 56, 61, 70, An element y , such that $y \in F(x)$, with $x \in E - \{y\}$, corresponds to a department y which has connection with x , either directly, or indirectly through other departments. In other words, y has mainly published either with x directly or with other elements which have mainly published with x directly or not. For example, department 61 has mainly published with departments 37 and 64, so 61 belongs to $F(37)$ and $F(64)$. On the other hand, department 70 has mainly published directly with 73 and not directly with 38. Indeed, $70 \in a(73)$, $70 \notin a(38)$, but $70 \in F(38)$ and $73 \in a(38)$.

The pseudoclosure $a(.)$ can also be defined by:

$$a(A) = \{x \in E \mid (\sum_{y \in A} c_{xy}/c_{x.}) > \alpha\} \cup A \quad (5)$$

with α a positive real number.

In this case, an element x belongs to the pseudoclosure $a(A)$ of a set A of departments, if and only if its number of co-authorings with all the elements of A relatively to its total number of co-authorings is greater than α . So the pseudoclosure defined by (5) enable us to formulate not only connection between two elements but also connection between an element and a set. Obviously, the algorithm can be applied to treat this new structural analysis problem even if the connection is not the relation of successors defined in graph theory. The closure $F(x)$ given in the table for some x in E' , have been computed with $\alpha = 1/3$. We can note that in this example we have $26 \notin a(75)$, $26 \notin a(47)$, $26 \notin a(94)$, but $26 \in a(\{75, 47, 94\})$.

4 Conclusion

This paper presents a general framework for problems encountered in population structural analysis. The method is based on pretopological concepts of pseudoclosure and minimal closed subsets. In the case described in the example with the pseudoclosure defined by (3), it provides results could be given by graph theory. Indeed, the set of elementary closed subsets \mathcal{F}_e corresponds to the transitive closure of the graph of which the set of successors of all elements x of E is $B(x)$. One obvious advantage of this method is that it can be applied in the same way, when the connections that exist between elements are different, as shown in the second example. It's just advisable to change the pseudoclosure definition which cannot be modeled by graph theory. Moreover, that application to co-authoring publications shows it can be useful to have indicators or statistical tests on the one hand to summarized the structural properties of the population E , and on the other hand to compare two pretopological structures induced on E . Future works are planned in that way.

References

- AURAY J-P. (1982): Contribution l'tude des structures pauvres. *Thse d'tat*, Universit Lyon1, France.
- AUTANT C., MASSARD N. (1999): Economtrie des externalits technologiques locales et gographie de l'innovation: une analyse critique. *Programme CNRS Les enjeux conomiques de l'innovation - Les cahiers de l'innovation n 99025*, Paris.
- BELMANDT Z.T. (1993): Manuel de prtopologie et ses applications. *Hermes*, Paris.
- BONNEVAY S., LAMURE M., LARGERON-LETO NO C., NICOLOYANNIS N. (1999): A pretopology approach for structuring data in non-metric spaces. *Electronic Notes in discrete Mathematics*.
- ČECH E. (1966): Topological spaces, *Publishing House of the Czechoslovak Academy of Sciences*, London, 893 pages.
- DIKRANJAN D., THOLEN W. (1995): Categorical Struture of Closure Operators, *Kluwer Academic Publishers*, Boston.
- DURU G. (1980): Contribution l'tude des systmes complexes dans les Sciences Humaines. *Thse d'tat*, Universit Lyon1, France.
- LAMURE M. (1987): Espaces abstraits et reconnaissance des formes. Application au traitement des images digitales. *PhD thesis*, Université Claude Bernard Lyon1, France, November.
- LARGERON-LETO NO C., BONNEVAY S. (1997): Une mthode de structuration par recherche de ferms minimaux - application la modlisation de flux de migrations intervilles. *SFC'97*, Lyon, France, 111-118.
- MASSARD N., LARGERON-LETO NO C. (1999): Externalits et cooprations scientifiques une tude de la structure gographique des co-publications en France. *XXXVI colloque de l'ASRDLF 'Innovation et conomie rgionale'*, Hyers, France, September.

A Robust Method for Multivariate Regression

Stefan Van Aelst¹, Katrien Van Driesssen² and Peter J. Rousseeuw³

¹ Research Assistant with the FWO, Belgium,
University of Antwerp, Dept of Mathematics and Computer Science,
Universiteitsplein 1, 2610 Wilrijk, Belgium.
(e-mail: Stefan.VanAelst@uia.ua.ac.be)

² University of Antwerp, Faculty of Applied Economics,
Prinsstraat 13, 2000 Antwerp, Belgium.
(e-mail: Katrien.VanDriesssen@ufsia.ac.be)

³ University of Antwerp, Dept of Mathematics and Computer Science,
Universiteitsplein 1, 2610 Wilrijk, Belgium.
(e-mail: Peter.Rousseeuw@uia.ua.ac.be)

Abstract. We introduce a new method for multivariate regression based on robust estimation of the location and scatter matrix of the joint response and explanatory variables. The resulting method has good equivariance properties and the same breakdown value as the initial estimator for location and scatter. We also derive a general expression for the influence function at elliptical distributions. We compute asymptotic variances and compare them to finite-sample efficiencies obtained by simulation.

1 Introduction

It is well known that classical multiple regression is extremely sensitive to outliers in the data. This problem also holds in the case of multivariate regression. Therefore, we aim to construct a robust method for multivariate regression which has bounded influence function and positive breakdown value.

Suppose we have p predictors $\mathbf{x} = (x_1, \dots, x_p)^t$ and q response variables $\mathbf{y} = (y_1, \dots, y_q)^t$, then the multivariate regression model is given by $\mathbf{y} = \mathcal{B}^t \mathbf{x} + \boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ where \mathcal{B} is the $(p \times q)$ slope matrix, $\boldsymbol{\alpha}$ is the q dimensional intercept vector, and the errors $\boldsymbol{\varepsilon}$ are i.i.d. according to $N_q(\mathbf{0}, \Sigma_{\boldsymbol{\varepsilon}})$. Let us write the mean $\boldsymbol{\mu}$ and covariance matrix Σ of the joint (\mathbf{x}, \mathbf{y}) variables as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{x}} \\ \boldsymbol{\mu}_{\mathbf{y}} \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{\mathbf{xx}} & \Sigma_{\mathbf{xy}} \\ \Sigma_{\mathbf{yx}} & \Sigma_{\mathbf{yy}} \end{pmatrix}.$$

In the classical case the empirical mean and covariance can be used to estimate $\boldsymbol{\mu}$ and Σ , yielding the least squares estimator given by

$$\hat{\mathcal{B}} = \hat{\Sigma}_{\mathbf{xx}}^{-1} \hat{\Sigma}_{\mathbf{xy}} \tag{1}$$

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\mu}}_{\mathbf{y}} - \hat{\mathcal{B}}^t \hat{\boldsymbol{\mu}}_{\mathbf{x}} \tag{2}$$

$$\hat{\boldsymbol{\varepsilon}} = \hat{\Sigma}_{\mathbf{yy}} - \hat{\mathcal{B}}^t \hat{\Sigma}_{\mathbf{xx}} \hat{\mathcal{B}}. \tag{3}$$

We now propose to use robust estimators for the center μ and covariance matrix Σ of the joint (\mathbf{x}, \mathbf{y}) variables. We will show that replacing the classical estimates by robust estimates for μ and Σ in expressions (1) and (2) yields a robust multivariate regression method. Note that in the special case of $q = 1$ we obtain the multiple regression case. Section 2 shows that the resulting estimator has the equivariance properties which we expect from a multivariate regression method. In Section 3 we investigate the robustness properties of this method. In Section 4 we use the minimum covariance determinant estimator (Rousseeuw (1984), (1985)) to estimate the location and scatter. As an example we consider the case of simple regression. Finally, Section 5 gives some conclusions.

2 Equivariance properties

First we generalize regression, scale and affine equivariance to multivariate regression estimators. Denote $T(\mathbf{x}, \mathbf{y}) = (\hat{\beta}^t, \hat{\alpha})^t$. The estimator T is called *regression equivariant* if

$$T(\mathbf{x}, \mathbf{y} + D^t \mathbf{x} + \mathbf{w}) = T(\mathbf{x}, \mathbf{y}) + (D^t, \mathbf{w})^t \quad (4)$$

where D is any $(p \times q)$ matrix and \mathbf{w} is any vector with q components. The estimator T is said to be *y-affine equivariant* if

$$T(\mathbf{x}, C\mathbf{y} + \mathbf{d}) = T(\mathbf{x}, \mathbf{y}) C^t + (O_{pq}^t, \mathbf{d})^t \quad (5)$$

where C is any nonsingular $(q \times q)$ matrix, \mathbf{d} is any vector with q components and O_{pq} is a $(p \times q)$ matrix containing zeros. One says the estimator T is *x-affine equivariant* if

$$T(A\mathbf{x} + \mathbf{v}, \mathbf{y}) = (\hat{\beta}^t A^{-1}, \hat{\alpha} - \hat{\beta}^t A^{-1}\mathbf{v})^t \quad (6)$$

for any nonsingular $(p \times p)$ matrix A and any column vector $\mathbf{v} \in \mathbb{R}^p$.

If the initial estimator $(\hat{\mu}, \hat{\Sigma})$ is an affine equivariant estimator of location and scatter, then the estimator $T = (\hat{\beta}^t, \hat{\alpha})^t$ given by (1) and (2) is regression, y-affine, and x-affine equivariant

3 Robustness properties

The breakdown value of an estimator T is the smallest fraction of observations (or probability mass) that needs to be replaced to carry T beyond all bounds (see Hampel et al. 1986). From (1) and (2) it follows that the estimator $T = (\hat{\beta}^t, \hat{\alpha})^t$ has the same breakdown value as the initial estimator for location and scatter.

The influence function of an estimator T at a distribution H measures the effect on T of an infinitesimal contamination at a single point (Hampel

et al. 1986). If we denote the point mass at $\mathbf{z} = (\mathbf{x}^t, \mathbf{y}^t)^t$ by $\Delta_{\mathbf{z}}$ and write $H_{\varepsilon} = (1 - \varepsilon)H + \varepsilon\Delta_{\mathbf{z}}$ then the influence function is given by

$$IF(\mathbf{z}, T, H) = \lim_{\varepsilon \downarrow 0} \frac{T(H_{\varepsilon}) - T(H)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} T(H_{\varepsilon})|_{\varepsilon=0}$$

Let us consider elliptical distributions $H_{\boldsymbol{\mu}, \Sigma}$ with density

$$h_{\boldsymbol{\mu}, \Sigma}(\mathbf{z}) = \frac{g((\mathbf{z} - \boldsymbol{\mu})^t \Sigma^{-1}(\mathbf{z} - \boldsymbol{\mu}))}{\sqrt{\det(\Sigma)}} \quad (7)$$

with $\boldsymbol{\mu} \in I\!\!R^p$ and Σ a positive definite matrix of size p . We assume the function g to have a strictly negative derivative, so that $H_{\boldsymbol{\mu}, \Sigma}$ is unimodal. From (1), (2), and (3) it follows that if the initial estimator $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ is Fisher-consistent at elliptical distributions then $\hat{\mathcal{B}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\Sigma}_{\varepsilon}$ are Fisher-consistent estimators of \mathcal{B} , $\boldsymbol{\alpha}$ and Σ_{ε} at elliptical distributions. Note that it suffices to compute the influence function at spherical distributions $H_{0,I}$ where I is the identity matrix. The influence function at elliptical distributions then follows from the equivariance properties in Section 2. If $(\hat{\boldsymbol{\mu}}, \hat{\Sigma})$ is Fisher-consistent, then the influence functions of $\hat{\mathcal{B}}$, $\hat{\boldsymbol{\alpha}}$ and $\hat{\Sigma}_{\varepsilon}$ at spherical distributions $H_{0,I}$ equal

$$IF(\mathbf{z}, \hat{\mathcal{B}}, H_{0,I}) = IF(\mathbf{z}, \hat{\Sigma}_{\mathbf{xy}}, H_{0,I}) \quad (8)$$

$$IF(\mathbf{z}, \hat{\boldsymbol{\alpha}}, H_{0,I}) = IF(\mathbf{z}, \hat{\boldsymbol{\mu}}_{\mathbf{y}}, H_{0,I}) \quad (9)$$

$$IF(\mathbf{z}, \hat{\Sigma}_{\varepsilon}, H_{0,I}) = IF(\mathbf{z}, \hat{\Sigma}_{\mathbf{yy}}, H_{0,I}) \quad (10)$$

Therefore, our proposal to use a positive-breakdown estimator with bounded influence function for the location $\boldsymbol{\mu}$ and scatter Σ in expressions (1), (2), and (3) yields a multivariate regression method with positive breakdown and bounded influence function.

4 Regression based on MCD

As robust estimator for the center $\boldsymbol{\mu}$ and covariance matrix Σ we propose to use the minimum covariance determinant estimator (MCD), and we call the resulting method *MCD regression*. Fix $0.5 \leq \alpha < 1$. The $MCD(\alpha)$ looks for the subset containing α percent of the data with a covariance matrix that has the smallest determinant. The estimates for the center and scatter are then the mean and a multiple of the covariance matrix of the optimal subset. The $MCD(\alpha)$ is a robust estimator of multivariate location and scatter with breakdown value approximately $1 - \alpha$, and it also has a bounded influence function (Croux and Haesbroeck (1999)). Two common choices for α are $\alpha = 0.5$ which yields the highest possible breakdown value, and $\alpha = 0.75$ which gives a better compromise between efficiency and breakdown. Recently Rousseeuw and Van Driessen (1999) constructed a fast algorithm to compute

the MCD. This algorithm makes the MCD very useful for analyzing large data sets.

For example, let us consider the case of simple regression ($p = 1, q = 1$). Figure 1 shows the influence functions at the bivariate gaussian distribution $H = N_2(\mathbf{0}, I)$ for slope \hat{b} , intercept \hat{a} and error scale $\hat{\sigma}^2$ of the MCD regression. Note that these influence functions are bounded.

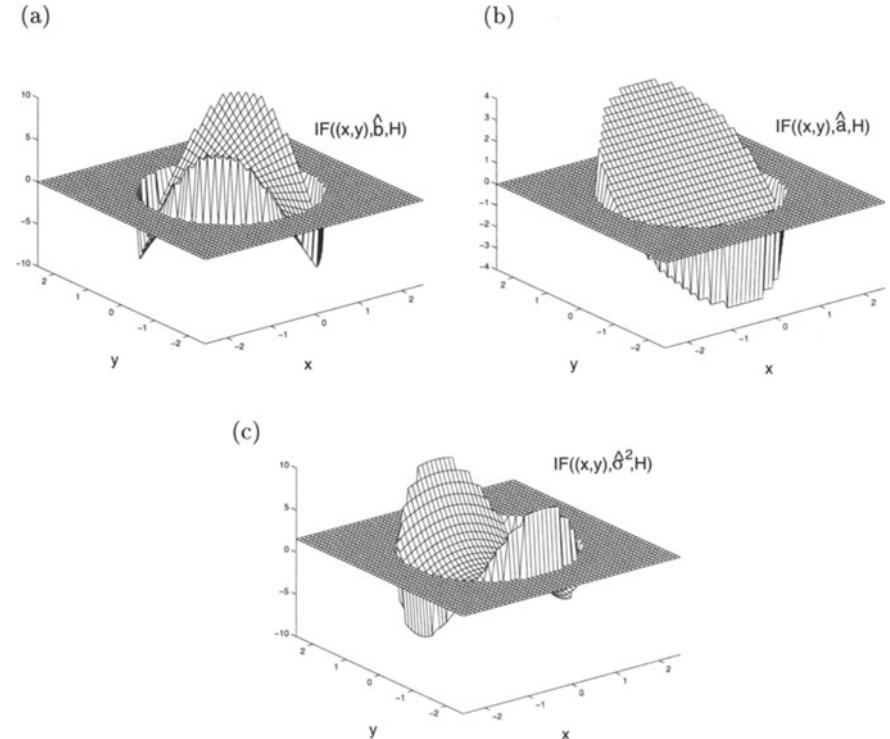


Fig. 1. Influence functions of (a) slope, (b) intercept, and (c) error scale of the MCD regression in the simple regression case.

Butler et al. (1993) showed that the MCD is asymptotically normal, therefore asymptotic variances can be computed from the influence functions. For the slope we use the following expression

$$ASV(\hat{b}, H) = \int IF^2(\mathbf{z}, \hat{b}, H) dH(\mathbf{z}).$$

The asymptotic relative efficiency (ARE) of \hat{b} relative to the least squares slope \hat{b}_{LS} is then given by

$$ARE(\hat{b}, H) = \frac{ASV(\hat{b}_{LS}, H)}{ASV(\hat{b}, H)}. \quad (11)$$

Similar expressions hold for the intercept \hat{a} and the error scale $\hat{\sigma}^2$. In the case $\alpha = 0.5$ and using (11) we obtain 3.3% for the ARE of the slope. For the intercept we find 15.3%, and the ARE of the error scale equals 6.2%. In the case $\alpha = 0.75$ the ARE of the slope becomes 16.3%, for the intercept we now obtain 40.3%, and for the error scale this yields 26.2%.

We want to compare these asymptotic results with finite-sample efficiencies. Therefore we performed the following simulations. For various sample sizes n we generated m data sets of size n from the bivariate gaussian distribution. For each data set $Z^{(j)}, j = 1, \dots, m$ we computed the slope $\hat{b}^{(j)}$, intercept $\hat{a}^{(j)}$ and error scale $\hat{\sigma}_{(j)}^2$. To measure the variance of slope and intercept we use the mean squared error (MSE). The MSE of the slope is given by

$$\text{MSE}(\hat{b}) = n \text{ave}_j(\hat{b}^{(j)})^2$$

since the true value of the slope equals 0. The corresponding finite-sample efficiency is given by $1/\text{MSE}(\hat{b})$. Analogously we find the finite-sample efficiency of the intercept. To measure the accuracy of the error term, we use the standardized variance (Bickel and Lehmann 1976) defined as

$$\text{StVar}(\hat{\sigma}^2) = \frac{n \text{var}_j(\hat{\sigma}_{(j)}^2)}{[\text{ave}_j(\hat{\sigma}_{(j)}^2)]^2}. \quad (12)$$

The corresponding finite-sample efficiency is then given by $2/\text{StVar}(\hat{\sigma}^2)$ since the Fisher information is 2. The results of our simulations are given in Table 1. Note that the finite-sample efficiencies converge well to the corresponding asymptotic efficiencies which are listed under $n = \infty$.

		$n = 100$	$n = 300$	$n = 500$	$n = \infty$
$\alpha = 0.75$	slope	0.17	0.18	0.17	0.163
	intercept	0.39	0.41	0.36	0.403
	error scale	0.24	0.27	0.25	0.262
$\alpha = 0.5$	slope	0.04	0.03	0.03	0.033
	intercept	0.14	0.15	0.15	0.153
	error scale	0.07	0.07	0.07	0.062

Table 1. Finite-sample efficiencies of the slope, intercept and error scale of MCD regression. The number of replications was $m = 500$.

The efficiency of the MCD can be increased by reweighting the $\text{MCD}(\alpha)$ estimates of μ and Σ . If we use the reweighted MCD as initial estimator for the multivariate regression, then the asymptotic efficiencies of the slope, intercept and error scale increase. From Table 2 we see that reweighting the MCD also increases the corresponding finite-sample efficiencies.

		$n = 100$	$n = 300$	$n = 500$	$n = \infty$
$\alpha = 0.75$	slope	0.49	0.56	0.67	0.637
	intercept	0.79	0.81	0.82	0.874
	error scale	0.56	0.61	0.57	0.599
$\alpha = 0.5$	slope	0.25	0.35	0.37	0.401
	intercept	0.55	0.73	0.75	0.847
	error scale	0.37	0.42	0.43	0.455

Table 2. Finite-sample efficiencies of the slope, intercept and error scale of regression based on the reweighted MCD. The number of replications was $m = 500$.

5 Conclusions

Least squares multivariate regression is sensitive to outliers in the data set. We have shown that a robust method is obtained when substituting robust estimates for location and scatter in the classical expressions for the slope, intercept and error scale. When using the MCD as initial estimator for the location and scatter we obtain a positive breakdown, bounded influence method. For the case of simple regression we computed the efficiency of the MCD regression method and performed simulations to compare the asymptotic results with finite-sample efficiencies. We also saw that the efficiency of the MCD regression can be markedly increased by using the reweighted MCD.

References

- BICKEL, P.J. and LEHMANN, E.L. (1976): Descriptive Statistics for Nonparametric Models III: Dispersion. *The Annals of Statistics*, 4, 1139–1159.
- BUTLER, R.W., DAVIES, P.L., and JHUN, M. (1993): Asymptotics for the Minimum Covariance Determinant Estimator. *The Annals of Statistics*, 21, 1385–1400.
- CROUX, C. and HAESBROECK, G. (1999): Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis*, 71, 161–190.
- HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J., and STAHEL, W.A. (1986): *Robust Statistics: the Approach based on Influence Functions*. John Wiley, New York.
- ROUSSEEUW, P.J. (1984): Least Median of Squares Regression. *Journal of the American Statistical Association*, 79, 871–880.
- ROUSSEEUW, P.J. (1985): Multivariate Estimation with High Breakdown Point. In: W. Grossmann, G. Pflug, I. Vincze and W. Wertz (Eds.): *Mathematical Statistics and Applications*, Vol. B. Reidel, Dordrecht, 283–297.
- ROUSSEEUW, P.J. and VAN DRIESSEN, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, 212–223.

Robust Methods for Complex Data Structures

Ursula Gather¹, Claudia Becker², and Sonja Kuhnt³

¹ Department of Statistics, University of Dortmund,
D-44221 Dortmund, Germany
(e-mail: gather@statistik.uni-dortmund.de)

² (e-mail: cbecker@statistik.uni-dortmund.de)

³ (e-mail: kuhnt@statistik.uni-dortmund.de)

Abstract. It is well known that the results of classical statistical methods may be affected by model deviations as for example the occurrence of outliers. The more complex a data structure or statistical procedure, the more complicated might be the mechanism of how outliers influence the analysis. The impact of spurious observations becomes less transparent with growing complexity of models and methods. Detailed sensitivity analyses and the development of suitable robustness concepts are therefore necessary.

1 Introduction

The statistical analysis of a dataset can be affected in various ways by deviations from the assumed model. This leads to a need for robust statistical procedures. Developing such procedures becomes harder and even sometimes impossible with increasing complexity of the data structure. Here, we concentrate on the case of model deviations given by outliers in a dataset. For example, when estimating the expectation of a univariate normal sample, we can understand the effects of outliers just by visualizing what happens. This becomes clearly more difficult with increasing dimension of the data. The easy looking task of identifying all outliers in a normal sample becomes rather complicated when dealing with 10-dimensional data, say. If, additionally, the data are of a more complex structure, like the one given by a contingency table, it is even more difficult to determine the influence of outliers. This is also the case if not only the data structure is complex, but the statistical procedure is complex, too. As an example, think of a compound procedure which is carried out in two steps. Here, an outlier may affect the first step of the procedure and the result of the first step affects the analysis in the second step. At the same time, the same outlier may also directly affect the second step of the procedure.

Such examples show that we need detailed analyses of the influence outliers may have when dealing with complex data structures and complex statistical procedures. For this purpose, we need criteria describing the impact of outliers on the analysis. A criterion measuring the influence of outliers or extreme values on an estimator is the finite-sample breakdown point (Donoho

and Huber (1983)). Estimators with high breakdown points do not produce arbitrarily bad results in the presence of a certain amount of extreme outliers and are robust in this respect. Taking the finite-sample breakdown point as a starting and worst case criterion, we can think of transferring its concept also to more complex situations. In the following sections, we discuss the development of robust methods with respect to breakdown criteria for some selected complex data structures. Section 2 focuses on multivariate simultaneous outlier identification rules in the case of a multivariate normal null model. In Section 3, we deal with structured data given by contingency tables. Finally, we discuss robust methods and possible breakdown concepts for a compound procedure in the context of dimension reduction.

2 Outliers in multivariate data

Consider a sample of size N from a p -dimensional normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Based on the idea of Davies and Gather (1993) we define the α_N outlier region of $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as the set

$$\text{out}(\alpha_N, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) > \chi_{p;1-\alpha_N}^2\}.$$

We take $\alpha_N = 1 - (1 - \alpha)^{1/N}$ for some given value of $\alpha \in (0, 1)$, adjusting the size of the outlier region to the sample size N (see Gather and Becker (1997), Becker and Gather (1999) for details). Now assume that we wish to detect all observations in a sample $\underline{X}_N = (\mathbf{X}_1, \dots, \mathbf{X}_N)$ which are outliers in the sense that they lie in the (unknown) outlier region $\text{out}(\alpha_N, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. To identify such outliers, we define and use a so-called α_N outlier identifier (which estimates the outlier region):

$$\text{OR}(\underline{X}_N, \alpha_N) = \{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \mathbf{m})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{m}) \geq c\},$$

where $\mathbf{S} = \mathbf{S}(\underline{X}_N) \in PDS(p)$ estimates $\boldsymbol{\Sigma}$ ($PDS(p) = \{\mathbf{S} \in \mathbb{R}^{p \times p} : \mathbf{S}$ positive definite and symmetric}), $\mathbf{m} = \mathbf{m}(\underline{X}_N) \in \mathbb{R}^p$ estimates $\boldsymbol{\mu}$, and $c = c(p, N, \alpha_N) \in \mathbb{R}$, $c \geq 0$ denotes a normalizing constant. Any $\mathbf{x} \in \text{OR}(\underline{X}_N, \alpha_N)$ is identified as an α_N outlier with respect to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. As a normalizing condition to determine c , we use

$$P(\mathbf{X}_i \in \mathbb{R}^p \setminus \text{OR}(\underline{X}_N, \alpha_N), i = 1, \dots, N) = 1 - \alpha,$$

where in this case $\mathbf{X}_1, \dots, \mathbf{X}_N$ are i.i.d. according to $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. This means we claim that with probability $1 - \alpha$ in a sample of size N from the p -variate normal, no observation will be identified as an outlier.

Breakdown of such an outlier identification procedure can for example occur if non-outliers are declared as arbitrarily large outliers (swamping breakdown) or if arbitrarily large outliers are not identified at all (masking breakdown). Thus, the idea of the finite-sample breakdown point of an estimator

can be transferred to the situation of outlier identification by defining masking and swamping breakdown points of identification rules. For the masking case,

$$\varepsilon^M(\mathbf{OR}) = \frac{k^M}{n + k^M},$$

denotes the so-called masking breakdown point of **OR**. The interpretation is that in a sample of size N , a number k^M of observations can be placed as outliers in such a way that they cause a masking breakdown of the identification procedure **OR**, where $N = n + k^M$, and that k^M is the minimum number of outliers to achieve this.

As shown in Becker and Gather (1999), for every affine equivariant identifier **OR**, we have

$$\frac{k}{N} \leq \varepsilon^M(\mathbf{OR}) \leq \frac{K}{N},$$

where $\varepsilon^*(\mathbf{m}) = k_1/N$ and $\varepsilon^*(\mathbf{S}) = k_2/N$ denote the finite-sample breakdown points of \mathbf{m} and \mathbf{S} with $k_i < N/2$, $i = 1, 2$, and $k = \min\{k_1, k_2\}$, $K = \max\{k_1, k_2\}$. Thus, the masking breakdown point of **OR** and the finite-sample breakdown point of the estimators involved in the identification procedure are strongly related. Hence robust multivariate simultaneous outlier identification procedures as introduced above should be based on robust estimators of location and scatter with high breakdown points to avoid masking breakdown of the identifier. An example of such an identification procedure would be the MCD identifier based on the robust MCD estimators (Rousseeuw (1985), Rousseeuw and van Driessen (1999), Becker and Gather (1999)).

From the findings presented above we see that the concept of the finite-sample breakdown point can be immediately applied to the context of outlier identification procedures in multivariate unstructured samples. In structured situations, like for example given by contingency tables, the situation appears to be slightly different.

3 Robust methods for the analysis of contingency tables

Contingency tables represent samples from distributions of categorical (i.e. finite discrete) vectors. Let such a vector consist of T variables with possible outcomes $\{1, \dots, i_t\}$, $t \in \{1, \dots, T\}$. For each possible combination $\mathbf{i} \in \mathcal{I} = \otimes_{t=1}^T \{1, \dots, i_t\}$ of outcomes, a cell with entry $x_{\mathbf{i}}$ means that combination \mathbf{i} occurs $x_{\mathbf{i}}$ times in the sample. The contingency table then consists of all frequencies $x_{\mathbf{i}}$, $\mathbf{i} \in \mathcal{I}$, which can also be seen as a vector $\mathbf{x} = (x_{\mathbf{i}})_{\mathbf{i} \in \mathcal{I}}$. Assuming a loglinear Poisson model (see Bishop, Fienberg and Holland (1975)) the corresponding random counts $X_{\mathbf{i}}$ are independent Poisson variables with expected values $m_{\mathbf{i}}$, $\mathbf{m} = (m_{\mathbf{i}})_{\mathbf{i} \in \mathcal{I}} \in \mathcal{M}$ with $\mathcal{M} \subseteq \mathbb{R}^{|\mathcal{I}|}$ as the model parameter space. Here, an α_I -outlier region of the respective Poisson distribution

$Poi(m_i)$ can be defined for each cell $i \in \mathcal{I}$ of the table:

$$out(\alpha_I, Poi(m_i)) = \{x \in \mathbb{N} : poi(x, m_i) < \delta(\alpha_I)\},$$

(where \mathbb{N} denotes the set of all positive integers and $poi(\cdot, m_i)$ the probability density function of the $Poi(m_i)$ distribution), with

$$\delta(\alpha_I) = \sup\{\delta > 0 : \sum_{x \in \mathbb{N}} poi(x, m_i) \mathbf{1}_{[0, \delta]}(poi(x, m_i)) \leq \alpha_I\}.$$

The adjustment $\alpha_I = 1 - (1 - \alpha)^{1/I}$ is made according to the number of cells in the table ($|\mathcal{I}| = I$). A cell count x_i is then considered as an outlier, if it lies in the outlier region $out(\alpha_I, Poi(m_i))$, which is again unknown and must be estimated from the data.

Let $\hat{\mathbf{m}}(\mathbf{X})$ be an estimator for the expectation vector \mathbf{m} of \mathbf{X} , with $\mathbf{m} = (m_i)_{i \in \mathcal{I}}$ and $\mathbf{X} = (X_i)_{i \in \mathcal{I}}$. A simultaneous outlier identifier can then be defined by

$$\mathbf{OI}(x_i; \mathbf{x}, \alpha_I) = \mathbf{1}_{out(c(\alpha_I), Poi(\hat{m}_i(\mathbf{x})))}(x_i) \quad i \in \mathcal{I}, \mathbf{x} \in \mathbb{N}^I,$$

with $c(\alpha_I)$ a given constant. The interpretation is that if $\mathbf{OI}(x_i; \mathbf{x}, \alpha_I) = 1$ the count x_i is identified as an outlier. The constant $c(\alpha_I)$ should be chosen in a way that the identifier fulfills the condition

$$P_{\otimes_{i \in \mathcal{I}} Poi(m_i)}(\mathbf{OI}(x_i; \mathbf{x}, \alpha_I) = 0) \quad \forall i \in \mathcal{I} \geq 1 - \alpha \quad \forall \mathbf{m} \in \mathcal{M}.$$

The concepts of breakdown for estimators and outlier identifiers presented in section 2 can be adjusted to the situation of contingency tables. However, for the finite-sample breakdown point, we have to note here, that \mathbf{x} takes values in \mathbb{N}^I , that the number of observations is fixed by I , and that $m_i \in]0, \infty[, i \in \mathcal{I}$. We therefore define

$$\varepsilon^*(\mathbf{x}, \hat{\mathbf{m}}) = \frac{k^*}{I},$$

where

$$k^* = \min\{k : \sup_{i \in \mathcal{I}, \mathbf{x}^k} \hat{m}_i(\mathbf{x}^k) = \infty \text{ or } \inf_{i \in \mathcal{I}, \mathbf{x}^k} \hat{m}_i(\mathbf{x}^k) = 0, k \in \{1, \dots, I\}\},$$

with \mathbf{x}^k constructed by replacing within \mathbf{x} a subset of k components by arbitrary positive integers.

In the same way, the masking breakdown point has to be defined by the minimum number k^M of cell counts, which have to be replaced by outliers to cause masking breakdown:

$$\varepsilon^M(\mathbf{x}, \mathbf{OI}) = \frac{k^M}{I}.$$

It can be shown that

$$\varepsilon^M(\mathbf{x}, \mathbf{OI}) \geq \varepsilon^*(\mathbf{x}, \hat{\mathbf{m}}).$$

An outlier identifier **OI** with a high masking breakdown point can therefore be obtained by using an estimator with a high finite-sample breakdown point.

The considered loglinear Poisson models describe complex associations between the categorical variables. The variables X_i are independent, but not identically distributed, and replacing one subset of the cell counts may be more effective in disturbing an estimator/identifier than replacing another subset. In this sense the above discussed breakdown concepts take the most pessimistic viewpoint and alternative concepts considering the pattern of the outliers in the table may be more appropriate. Some ideas in this direction will be presented. For example a stochastic breakdown function following Donoho and Huber (1983) can be defined. This function gives for each possible number of replaced observations the probability that they are arranged in a pattern causing breakdown.

From these investigations for the data structure given by a contingency table, it can be seen that a modification of well-known robustness concepts like the breakdown concept to such situations seems appropriate and possible. The following section deals with the case where also the procedure under consideration is of a more complex type.

4 Robustness for compound procedures: the example of sliced inverse regression

Statistical methods for analysing highdimensional and complex data structures often combine several procedures. This is the case for hybrid methods from different disciplines (like e.g. from computer science and statistics) as well as for statistical procedures composed of several parts. Developing robustness concepts for such compound procedures means investigating the performance of rather complex methods under model deviations. Concepts like the breakdown point cannot be transferred into this framework immediately.

As an example, consider the method Sliced Inverse Regression (SIR) for reducing the dimensionality in regression problems (Li (1991)). Here, we are in the regression setting $Y = g(\mathbf{X}) + \varepsilon$, where Y, \mathbf{X} are \mathbb{R} - and \mathbb{R}^d -valued random variables, respectively, and d may be rather large. The idea for dimension reduction is that there exists some lowdimensional space $\mathcal{B} = \text{span}[\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K]$, $K \ll d$, and some link funktion f such that to investigate the relation between Y and \mathbf{X} it suffices to consider \mathcal{B} because we may write $Y = f(\boldsymbol{\beta}_1^T \mathbf{X}, \dots, \boldsymbol{\beta}_K^T \mathbf{X}, \varepsilon)$. The aim of SIR is to estimate \mathcal{B} , and this is done in a five-step procedure consisting of standardization, "slicing", estimating slice locations, estimating the covariance of the slice locations, and performing a principal component analysis.

Thus, a first approach to study the breakdown of SIR could be to consider the breakdown of the estimators used in each step of the procedure and to define the breakdown point of the whole procedure as the minimum of the breakdown points of all estimators. But it turns out that this approach is not appropriate. For example, the first step of SIR consists in standardizing the observations, where a location and a covariance estimator are used. Detailed analysis of the performance of SIR shows that the location estimate in this step does not influence the outcome of the procedure. It should therefore not be taken into account when defining the breakdown point of the procedure. Altogether more refined concepts are needed to describe the breakdown and robustness of SIR and the same is true for other statistical compound procedures as well.

References

- BECKER, C. and GATHER, U. (1999): The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association*, 94, 947–955.
- BISHOP, Y.M.M., FIENBERG, S.E. and HOLLAND, P.W. (1975): *Discrete multivariate analysis*. The MIT Press Cambridge, Massachusetts.
- DAVIES, P.L. and GATHER, U. (1993): The Identification of Multiple Outliers. Invited paper with discussion and rejoinder. *Journal of the American Statistical Association*, 88, 782–801.
- DONOHO, D.L. and HUBER, P.J. (1983): The Notion of Breakdown Point. In: P.J. Bickel, K.A. Doksum and J.L. Hodges Jr. (Eds.): *A Festschrift for Erich L. Lehmann*. Wadsworth, Belmont, CA, 157–184.
- GATHER, U. and BECKER, C. (1997): Outlier Identification and Robust Methods. In: G.S. Maddala and C.R. Rao (Eds.): *Handbook of Statistics, Vol. 15: Robust Inference*. Elsevier, Amsterdam, 123–143.
- LI, K.-C. (1991): Sliced Inverse Regression for Dimension Reduction (with discussion). *Journal of the American Statistical Association*, 86, 316–342.
- ROUSSEEUW, P.J. (1985): Multivariate Estimation with High Breakdown Point. In: W. Grossmann, G. Pflug, I. Vincze and W. Wertz (Eds.): *Mathematical Statistics and Applications*. Reidel, Dordrecht, 283–297.
- ROUSSEEUW, P.J. and VAN DRIESSEN, K. (1999): A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, 212–223.

Robust Methods for Canonical Correlation Analysis

Catherine Dehon¹, Peter Filzmoser², and Christophe Croux¹

¹ ECARES, CP139, Université Libre de Bruxelles,
Av. F.D. Roosevelt 50, B-1050 Bruxelles, Belgium
(e-mail: cdehon@ulb.ac.be)

² Department of Statistics, Probability Theory, and Actuarial Mathematics,
Vienna University of Technology,
Wiedner Hauptstr. 8-10, A-1040 Vienna, Austria

Abstract. Canonical correlation analysis studies associations between two sets of random variables. Its standard computation is based on sample covariance matrices, which are however very sensitive to outlying observations. In this note we introduce, discuss and compare four different ways for performing a robust canonical correlation analysis. One method uses robust estimators of the involved covariance matrices, another one uses the signs of the observations, a third approach is based on projection pursuit, and finally an alternating regression algorithm for canonical analysis is proposed.

1 Introduction

The aim of Canonical Correlation Analysis (CCA) is to identify and quantify the relations between a p -dimensional random variable \mathbf{X} and a q -dimensional random variable \mathbf{Y} . Herefore we look for linear combinations $a^\top \mathbf{X}$ and $b^\top \mathbf{Y}$ of the original variables having maximal correlation. Expressed in mathematical terms, CCA seeks for vectors $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ such that

$$(\alpha, \beta) = \underset{a, b}{\operatorname{argmax}} |\operatorname{Corr}(a^\top \mathbf{X}, b^\top \mathbf{Y})|. \quad (1)$$

The resulting univariate variables $U = \alpha^\top \mathbf{X}$ and $V = \beta^\top \mathbf{Y}$ are then called the *canonical variates* and can be used for dimension reduction and graphical display. Note that the vectors α and β are only determined up to a constant factor by definition (1). To define them uniquely one adds a normalization constraint requiring that both U and V have unit variance or alternatively that α and β have unit norm. The first canonical correlation ρ is now defined as the absolute value of the correlation between the two canonical variates, which equals the maximum attained in (1).

Higher order canonical variates and correlations are defined as in (1), but now under the additional restriction that a canonical variate of order k , with $1 < k \leq \min(p, q)$, should be uncorrelated with all canonical variates of lower order. Due to space limitations, we restrict attention to a first order canonical analysis.

The above CCA problem (1) has a fairly simple solution (see e.g. Johnson and Wichern, 1998, Chapter 10). Denote by Σ the population covariance matrix of the random variable $\mathbf{Z} = (\mathbf{X}^\top, \mathbf{Y}^\top)^\top$. We decompose Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

The vectors α and β are now the eigenvectors corresponding to the largest eigenvalue of the matrices

$$\Sigma_{xx}^{-1} \Sigma_{xy} \Sigma_{yy}^{-1} \Sigma_{yx} \quad \text{and} \quad \Sigma_{yy}^{-1} \Sigma_{yx} \Sigma_{xx}^{-1} \Sigma_{xy}. \quad (2)$$

Both of the above matrices have the same positive eigenvalues and the largest one equals the first canonical correlation.

For estimating the unknowns α , β , and ρ one typically computes the sample covariance matrix $\hat{\Sigma}$ out of a sample z_1, \dots, z_n , with $z_i = (x_i^\top, y_i^\top)^\top \in \mathbb{R}^p \times \mathbb{R}^q$. Computing eigenvectors/values of the empirical counterparts of the matrices in (2) results then immediately in estimates of the canonical variates and correlations.

Since the classical estimation of a covariance matrix is very vulnerable with respect to outlying observations, also the eigenvalues and -vectors based on these covariance estimations will be very sensitive, as was shown in the context of CCA by Romanazzi (1992). In Section 2 of this paper four different robustifications of CCA are proposed and discussed. They are compared by means of a modest simulation study in Section 3. A more comprehensive study of the different approaches is part of current research of the authors.

2 Robust approaches to CCA

2.1 Using robust covariance matrix estimators

The obvious way for robustifying CCA is to robustly estimate Σ and to compute eigenvectors/values from the estimated version of (2) in the usual way. Some theoretical results for this method have been obtained by Croux and Dehon (1999). As robust covariance estimators one could use M-estimators as in Karel (1991), but it is known that these estimators have poor robustness properties in higher dimensions. A more appropriate choice for $\hat{\Sigma}$ seems to be the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985). The MCD estimator is obtained by looking for that subset of size h of the data which has the smallest value of the determinant of the empirical covariance matrix computed from it. Typically, $h \approx n/2$ or $3n/4$. The resulting estimator is then nothing else but the covariance matrix computed over that optimal subset. An efficient algorithm for computing the MCD estimator has been proposed by Rousseeuw and Van Driessen (1999), and it

was shown by Croux and Haesbroeck (1999) that the MCD has reasonable efficiency properties.

Robust covariance matrix estimators can routinely be used in multivariate statistics. Filzmoser (1999) applied them for robust factor analysis of geostatistical data.

2.2 Using the signs of the data

Instead of working with robust covariance matrices, one could mitigate the influence of outliers directly by replacing the observations by their “signs” or “ranks” and compute an ordinary covariance matrix out of these signs and ranks (cfr. Visuri, Koivunen, and Oja, 1999). Signs were already used for principal components analysis by Locantore et al. (1999).

In order to compute the sign of an observation x_i (and similarly for y_i), we first need to compute the spatial median m_X of the data cloud x_1, \dots, x_n . The spatial median is defined as

$$m_X = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^n \|x_i - \mu\|,$$

with $\|\cdot\|$ the Euclidean norm. The sign of x_i is then the projection of x_i on a unit sphere centered at m_X :

$$S(x_i) = \frac{x_i - m_X}{\|x_i - m_X\|}.$$

The rank of an observation does not require a location estimate and is defined as

$$R(x_i) = \frac{1}{n} \sum_{j=1}^n \frac{x_i - x_j}{\|x_i - x_j\|}.$$

Note that both signs and ranks are vector valued. In this abstract, we prefer to work with signs instead of ranks. The computation of all ranks asks for $O(n^2)$ computation time, while the signs can be computed in $O(n)$ time, given that very fast iterative algorithms to compute m_X and m_Y exist.

2.3 Using Projection Pursuit

A Projection Pursuit (PP) approach starts from the initial definition (1) of CCA. Instead of maximizing $|\operatorname{Corr}(a^\top \mathbf{X}, b^\top \mathbf{Y})|$, other correlation indices could be maximized. Since this index only needs to be defined between univariate variables, a simple Spearman correlation can be taken.

In practice, it is not obvious how to find the vectors α and β maximizing (the absolute value of) this correlation index. A simple and fairly good approximation is obtained by restricting the search to the finite set

$\{(a_i, b_j) | 1 \leq i, j \leq n\}$, where a_i is the normed vector $x_i - \mu_X$, μ_X a robust location measure of the X -population (e.g. the spatial median defined in Section 2.2 or the coordinatewise median) and b_j the normed $y_j - \mu_Y$.

Although that the non-parametric nature of the Spearman correlation is very appealing, the computational complexity of $O(n^3 \log n)$ may become prohibitive for bigger sample sizes.

2.4 Using robust alternating regressions

The alternating regressing technique originates from Wold (1966) and received renewed interest over the last few years (e.g. Gabriel, 1998). As already proposed by Wold (1966), it can also be used to estimate canonical variates. Its use is motivated by the observation that, for a given α ,

$$\beta = \underset{b}{\operatorname{argmax}} |\operatorname{Corr}(\alpha^\top \mathbf{X}, b^\top \mathbf{Y})|.$$

But then it follows from standard results on multiple regression, that β is proportional to the regression coefficient b in the model

$$\alpha^\top X = b^\top Y + \gamma_1 + \varepsilon_1. \quad (3)$$

In the same way, for a given β , the optimal α equals (upto a scalar term) the parameter a in the regression model

$$b^\top Y = a^\top X + \gamma_2 + \varepsilon_2. \quad (4)$$

Start now with an initial value α_0 . (This can for example be obtained by performing a robust principal components analysis on the data matrix formed by x_1, \dots, x_n .) According to (3), a first β_1 is then obtained by regressing the univariate $\alpha_0^\top X$ on the Y variables. Afterward, using (4), an updated α_1 is obtained by regressing $\beta_1^\top Y$ on X . This procedure is then iterated until convergence. Note that the estimated regression coefficients are normalized in each step and that an estimator of ρ is obtained by computing a bivariate robust correlation between the canonical variates.

The regression estimators in the above scheme need to be robust. Since they are computed several times, a fast, reliable estimator should be chosen. We propose to use a weighted L_1 -estimator, as was motivated by Croux and Filzmoser (1998) in an application of alternating regressions to factor analysis.

3 Simulation

In this section the proposed methods of Section 2 are compared with the classical CCA method by a modest simulation study. For $m = 1, \dots, M = 200$ simulations, we generated data $x_1^m, \dots, x_n^m \in \mathbb{R}^p$ and $y_1^m, \dots, y_n^m \in \mathbb{R}^q$ from

a specified $N(0, \Sigma)$ distribution. We chose sample size $n = 50$, and $p = 2$, $q = 3$. The estimated parameters are denoted by $\hat{\rho}^m$, $\hat{\alpha}^m$, and $\hat{\beta}^m$ and are compared with the “true” parameters ρ , α and β which were derived from the known population covariance matrix Σ . Therefore the following measures of mean squared error (MSE) are computed:

$$\text{MSE}(\hat{\rho}) = \frac{1}{M} \sum_{m=1}^M (\phi(\hat{\rho}^m) - \phi(\rho))^2, \quad (5)$$

where $\phi(\rho) = \tanh^{-1}(\rho)$ is the Fisher transformation of ρ (which is classically applied to turn a distribution of correlation coefficients towards normality), and for the canonical variates

$$\text{MSE}(\hat{\alpha}) = \frac{1}{M} \sum_{m=1}^M \cos^{-1} \left(\frac{|\alpha^\top \hat{\alpha}^m|}{\|\hat{\alpha}^m\| \cdot \|\alpha\|} \right), \quad (6)$$

and similarly for $\text{MSE}(\hat{\beta})$. The measure (6) is the average value of the positive angles between the vectors $\hat{\alpha}^m$ and α . The use of angles makes the MSE invariant to the choice of the normalization constraint for the canonical variates.

Since we are also interested in the behaviour of the methods when outliers are present, we generated in a second experiment the data as before, but we added to the first 5 generated observations x_1^m, \dots, x_5^m randomly generated noise from $N(0, 50I_p)$. This results in a level of 10% contamination. The MSEs for the estimators outlined in Section 2 are summarized in Table 1 for both experiments.

Table 1. Simulated MSEs for the canonical correlations and covariates estimators using classical, MCD-based, sign-based, PP-based, and alternating regression based CCA at normal data sets (*clean*) and contaminated data sets (*outliers*) of size $n = 50$, where $p = 2$ and $q = 3$.

	MSE($\hat{\rho}$)		MSE($\hat{\alpha}$)		MSE($\hat{\beta}$)	
	clean	outliers	clean	outliers	clean	outliers
Classical	0.02	1.33	0.14	0.99	0.28	0.92
MCD-based	0.06	0.05	0.23	0.22	0.40	0.40
Sign-based	0.12	0.25	0.18	0.19	0.32	0.40
PP-based	0.03	0.16	0.21	0.27	0.35	0.51
Regression	0.06	0.05	0.20	0.22	0.34	0.37

We clearly see that on the clean data the classical CCA is the most precise, but the robust procedures are not so far behind. On the other hand, the

classical method looses its optimality in the contaminated case where it is outperformed by the robust methods. Distinguishing between the 4 robust approaches on the basis of this simulation experiment is much more difficult. One could say that the MCD-based and the alternating regression based method give the better results for the estimation of the canonical correlation.

As a final conclusion, making a choice between the four available robust procedures is quite difficult. It seems that software availability and computational complexity are the more determining factors.

References

- CROUX, C. and DEHON, C. (1999): Robust Canonical Correlations using High Breakdown Scatter Matrices. Preprint, Université Libre de Bruxelles.
- CROUX, C., and FILZMOSER, P. (1998): A Robust Biplot Representation of Two-way Tables. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer-Verlag, Berlin, 355–361.
- CROUX, C. and HAESBROECK, G. (1999): Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis*, 71, 161–190.
- FILZMOSER, P. (1999): Robust Principal Component and Factor Analysis in the Geostatistical Treatment of Environmental Data. *Environmetrics*, 10, 363–375.
- GABRIEL, K.R. (1998): Generalized Bilinear Regression. *Biometrika*, 85, 186–196.
- JOHNSON, R.A. and WICHERN, D.W. (1998): *Applied Multivariate Statistical Analysis*. Fourth Edition, Prentice Hall, New Jersey.
- KARNEL, G. (1991): Robust Canonical Correlation and Correspondence Analysis. *The Frontiers of Statistical Scientific Theory & Industrial Applications*, 335–354.
- LOCANTORE, N., MARRON, J.S., SIMPSON, D.G., TRIPOLI, N., ZHANG, J.T., and COHEN, K.L. (1999): Robust Principal Components for Functional Data. *Test*, 8, 1–73.
- ROMANAZZI, M. (1992): Influence in Canonical Correlation Analysis. *Psychometrika*, 57, 237–259.
- ROUSSEEUW, P.J. (1985): Multivariate Estimation with High Breakdown Point. In: W. Grossmann et al. (Eds.): *Mathematical Statistics and Applications*, Vol. B. Reidel, Dordrecht, 283–297.
- ROUSSEEUW, P.J., and VAN DRIESSEN, K. (1999): A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41, 212–223.
- VISURI, S., KOIVUNEN, V., and OJA, H. (1999): Sign and Rank Correlation Matrices. *Journal of Statistical Planning and Inference*. To appear.
- WOLD, H. (1966). Nonlinear Estimation by Iterative Least Squares Procedures. In: F.N. David (Ed.): *A Festschrift for J. Neyman*. Wiley and Sons, New York, 411–444.

Part IV

Data Science

From Data Analysis to Data Science

Noboru Ohsumi

The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku Tokyo 106-8569, Japan
(e-mail: ohsumi@ism.ac.jp)

Abstract. This paper discusses the significance of the term “data science” to the Japanese Classification Society (JCS) and the international relevance of JCS’s research. In 1992, the author argued the urgency of the need to grasp the concept “data science”. Despite the emergence of concepts such as data mining, this issue has not been addressed. Discussion will emphasize the history of methods of data analysis proposed by J. Tukey. The interaction between Japan and, particularly, France in the development of data analysis will be emphasized.

1 The research interchange between Japan and France

Because of differences in cultures and researchers’ approaches, globalization of the field of statistical science and data analysis remains a future prospect. At the risk of being accused of making an arbitrary interpretation, the author asserts that Japanese researchers looked to France and Germany in the field of mathematics, and toward the UK and the USA in the field of statistical science. The field of data analysis was a rare exception, where Japanese and French researchers collaborated. To our regret, the history of these interchanges is not widely known among statistical science researchers.

One such exchange was between Professor Matusita of the Institute of Statistical Mathematics (ISM) and the late Professor Dugué of the Institute of Statistics at the University of Paris VI. Through their shared interests in traditional mathematical statistics, especially multivariate analysis, they organized the Japanese–French Scientific Seminar on “Data Analytic Methods for Analysing Measurement Datasets”. This “bridging seminar” marked the beginning of research exchanges between the researchers in both countries and the subsequent development of data analysis in both countries. It also marked the beginning of an enduring collaboration of Japanese researchers with French researcher Professor J.-P. Benzécri and promising young data analysts of Benzécri’s school, including Lebart, Roux, and Jambu.

A group of Japanese researchers led by C. Hayashi was at the hub of this field in Japan at the time, and had achieved considerable advances in data analysis research. Researchers at the ISM knew only partially of concurrent developments in France. It was known, for example, that development of a method similar to Hayashi’s Quantification Methods, Type III, had stimulated progress in data analysis in France. However, nothing was known of the work of the “phantom researcher” Professor J.-P. Benzécri.

Such contact included seminars and special lectures at the ISM, the JCS, and other places in Japan to introduce the French philosophies to Japanese researchers. The “*analyse des données*” introduced by Roux was astonishingly new and stimulating to Japanese researchers, as were correspondence analysis (CA) and automatic classification. Roux made an immense contribution to clarifying the similarity of the mathematics between CA and the Type III Quantification Method. Many notable achievements, such as Hayashi’s quantification methods and Akaike’s information criterion, developed in quick succession. The ISM played an essential part in offering opportunities to put these new theories into practice. It was particularly noteworthy that much of the research based on social surveys in Japan, such as the Survey of Japanese National Character, was undertaken using the Type III Quantification Method.

After Roux visited Japan, Professor L. Lebart, an authority in data analysis and social survey research, was invited to participate in a project that involved Japanese and French researchers from the Japan Society for the Promotion of Science (JSPS), the Centre National de la Recherche Scientifique (CNRS), and the ISM conducting a survey of international attitudes to the “Japanese and French national characters”.

2 Later international research interchanges

By 1983, the SFC had started in France, and the Classification Society of North America (CSNA), the Gesellschaft für Klassifikation in Germany, and the British Classification Society (BCS) had been organized. In the same year, Hayashi, together with some researchers who were international members of the CSNA and BCS, founded the Japanese Classification Society. Membership in the International Federation of Classification Societies (IFCS) was gained and, through the great effort of H. Bock and others, the JCS was in the fortunate position of being able to host the Fifth IFCS-96 Conference.

Japanese researchers also promoted international interchange through large conferences such as the meetings of the International Statistical Institute (ISI) and International Biometric Society (IBS) held in Tokyo in the 1980s, which attracted such researchers as Y. Escoufier, J.-P. Nakache, Bouroche, J. Gower, A. Rizzi, and N. Lauro.

The period between 1979 and 1985 marked an important period of close research exchanges between Japan and many European countries.

3 “Analyse des Données” and “Deta Kaiseki”

It is most important to discuss the similarities and differences in the approaches to data analysis between Japan and France with regard to the Quantification Method, especially CA. It is important to emphasize that we agree on the need to develop, through practice, research on the theory

and application of data analysis into a new “data science”. Hayashi’s Quantification Methods comprise several methods, from Type I to Type VI. In particular, Type III coincides with CA. Hayashi proposed this method in 1952. Underpinning Hayashi’s methods was the concept of scaling methods, by which the other methods were unified and discussed. Benzécri’s CA (AFC: Analyse Factorielle des Correspondances) appeared about 1962 (Benzécri, 1982). How well it was accepted and what applications were developed from it goes without saying. Benzécri and his school developed elaborate and varied theories of CA and related methodologies. Moreover, considerable research was conducted on automatic classification by many researchers, including Diday, Jambu, Lerman, and Roux. To our regret, however, the “barrier of language” prevented Japanese researchers from gaining true recognition for their achievements. The jargon used in research on the “*analyse des données*” made things even more difficult. Although there has been some improvement, we are still in much the same situation.

In Japan, the term “deta kaiseki” (data analysis) was often misunderstood. The Japanese language used in these papers prevented these achievements from becoming known to international researchers. However, because of publications in Japanese by Lebart and Ohsumi (1994), Japanese researchers are now able to obtain more results of research in France and in other countries. Differences in language, thought and culture make most Japanese researchers more interested in research in English-speaking countries, which presents a great problem for us to solve in the present time. Books in English by Greenacre (1984) and Jambu (1983) are read by many Japanese researchers and students. Those who are interested in *analyse des données* are increasing in number.

Two important results should be remembered in the history of research interchange. In the past, Japanese–French Scientific Seminars were arranged. The first meeting was held at the ISM in Tokyo in 1987 and attracted 180 researchers—an unexpectedly large number. The second meeting was in Montpellier University II in France in 1992. Fewer researchers participated, but the outcome of this meeting was significant: the term “data science” appeared for the first time, and was subsequently used in the preface of a conference publication (“Data Science and Its Applications— La Science des Données et ses Applications”: Escoufier et al., 1995).

Researchers in Japan do not all share the same understanding of the concept “data science”. The Japan Statistical Society held special sessions on data science at its annual meetings in 1996 and 1997, and drew much interest. However, in the opinion of most researchers, they did not go beyond the general framework of statistical modelling or traditional statistical analysis. One organizer was heard to criticize Japanese researchers for using other researchers’ data without paying any attention to the most important problem of data acquisition. What, then, is our “data science”?

What I mean by “data science” includes the most essential studies and concepts on *how to gather data*, including *how to design experiments in data gathering*, and *how to analyse the collected data*. These are the fundamental ways to obtain meaningful findings from many events. How data are gathered is the key to defining the relevant information and making it easy to understand and analyse. In my opinion, this viewpoint on the meaning of data science is fundamentally different from data mining (DM) and knowledge discovery (KD). These concepts are not of practical use because they neglect the problems of “data acquisition” and its practice.

4 Relationship to IFCS: Changing from a linear to a spatial perspective

Japan’s foreign relations in the field of data science developed from initial research exchanges with France. The relation was at first a linear one, but more extensive relations followed through foundation of the IFCS, as did exchanges with many other countries. The IFCS was founded in 1983 to federate the classification societies from many countries. The First IFCS International Conference, held at Aachen in Germany in 1987 (organized by Professor Bock), deserves special mention for being the first meeting held by the federation of BCS, CSNA, GfKI, JCS, SFC, and SIS. Japan hosted the Fifth IFCS-96 Conference; this was the culmination of 20 years of international research exchange. In this context, the association between Japan and France may have undergone a marked change from a linear to a spatial relationship.

5 Toward Data Science: as prospects in Data Analysis

The Japanese song “A canary that has forgotten singing” describes the current trend in the field of the data analysis. It appears researchers are seeking mathematical methodologies without considering “what data analysis is” and “what the data acquisition should be”. Were we not seeking for a different world of statistical science and data analysis?

Owing to qualitative and quantitative changes in data, it is, indeed, becoming increasingly difficult to grasp all aspects of a dataset in explaining various phenomena. Therefore, new techniques, such as DM, KD, complexity, and neural networks, are being proposed. However, the potential of these methods to solve any of these problems is questionable.

We now have to deal with not only extremely complicated analyses but also greatly altered data. Their characteristics could be categorized as follows:

1. A dataset collected with a definite purpose of explaining phenomena on the basis of statistically appropriate design of experiments or sampling procedures; for example, social survey data including opinion surveys or

- attitude surveys. The data acquisition process is transparent, traceable, or reproducible.
2. Laboratorial measurement data gathered with measuring tools or devices. The data include various kinds of measurement units, such as environmental indexes and health indexes, as well as datasets acquired through actual measurements.
 3. Data that accumulate gradually in the database by an information processing technique. The purpose or intention of the accumulation cannot be clearly demonstrated. These data include POS data, banking and credit data, and basic financial and personnel databases of corporations.
 4. A new qualitative kind of dataset and its database. In particular, textual or non-numerical data extracted from open-ended responses or free format answers and collected systematically. For example, textual data gathered almost automatically through Internet researches, telephony-marketing researches, and call centres for customers.
 5. An aggregated dataset generating spontaneously and accumulating automatically in the electronic data collection environments, and its database or data warehouse. A mass of data, the importance of which is not readily known, but which is managed by the high-tech database with a view to extracting some meaning in the future.

When it comes to analysing these datasets, people discuss DM and related techniques. However, the important questions to answer are: what dataset is necessary to explicate a certain phenomenon, why is it necessary, how to design its acquisition, and how difficult the whole process is. This is more important than the dataset itself. Books on DM do contain terms such as “data preparation”, “getting the data”, “sampling procedures”, and “data auditing”, but there is an assumption that the dataset is given and the procedure may start with analysis. Fiddling with a dataset once it is collected is merely a self-contented play of data handling. As noted, there are many possible ways to acquire a dataset. Taking this into consideration, one should ask what data analysis should be. To come to the point: I have discussed the paradigm through which we should discuss the concept of data science. Unfortunately, although it is such a basic and fundamental concept. I doubt whether data analysts have been well aware of its importance.

A decline in statistical science was brought to our attention long ago. Nevertheless, no marked improvement has been made. No university in Japan has a department of statistics. In the field of statistical science and data science, we have only one specialized research institute, ISM. Our only statistical science course in graduate school is also at the ISM. Recently in Japan, there has been a great deal of discussion over the guidelines for improving scientific research. In the fields of computational science and informatics, many have thought it necessary to examine how to advance research, and many research projects from other countries are being introduced for the purpose of comparison or benchmarking. Models drawn from large-scale national research

centres in European countries, such as INRIA, the organization of CNRS, the Max Planck Gesellschaft, and MPI-Institut für Informatik have drawn considerable interest. In the field of data analysis, a large-scale National Institute of Informatics was partially commenced in 1999, and is planned as part of a structural reorganization program.

At present, however, there is no clear direction for change; we must determine that direction, each re-examining and revitalizing our separate attitudes. For that purpose, we might have to seek collaboration with other fields, or even consider the possibility of re-organization and integration. We might have to abandon such terms as statistical science or data analysis or similar concepts, and choose, for example, "data science" as a new paradigm. We do believe that such a concept can help to guide and foster a fruitful and expanding relationship among many countries in the future. We very much hope this new age of "data science" will come to fruition, and that what we have achieved in the history of "data analysis" will be of enduring benefit to the coming science and to future research.

Acknowledgements

I wish to thank all staff and members of the Organization Committee of the IFCS-2000 Conference for giving me the opportunity to present this report, and to the researchers from each country belonging to the IFCS, for making great efforts toward the development of data science.

References

- DIDAY, E., LEBART, L., PAGES, J.-P. and TOMASSONE, R. (Eds.) (1979): *Data Analysis and Informatics*. North-Holland, Amsterdam.
- DIDAY, E., JAMBU, M., LEBART, L., PAGES, J.-P. and TOMASSONE, R. (Eds.) (1983): *Data Analysis and Informatics III*. North-Holland, Amsterdam.
- DIDAY, E. and others. (Eds.) (1986): *Data Analysis and Informatics IV*. North-Holland, Amsterdam.
- ESCOUFIER, Y., HAYASHI, C., FFICHET, B., DIDAY, E., LEBART, L., OHSUMI, N. and BABA, Y. (Eds.) (1995): *Data Science and its Applications* (La science des données et ses applications). Academic Press, Tokyo.
- GREENACRE, M. J. (1984): *Theory and Applications of Correspondence Analysis*. Academic Press, Boston.
- HAYASHI, C., DIDAY, E., JAMBU, M. and OHSUMI, N. (Eds.) (1988): *Recent Developments in Clustering and Data Analysis* (Developpements récents en classification automatique et analyse des données). Academic Press, Boston.
- HAYASHI, C., OHSUMI, N., BOCK, H.-H. and others (Eds.) (1997): *Data Science, Classification and Related Methods*. Springer-Verlag, Tokyo.
- JAMBU, M. and LEBEAUX, M.-O. (1983): *Cluster Analysis and Data analysis*. North-Holland, Amsterdam.
- OHSUMI, N., LEBART, L. and others (1994): *Multivariate Descriptive Statistical Analysis* (in Japanese). JUSE Press, Ltd., Tokyo.

Evaluation of Data Quality and Data Analysis

Chikio Hayashi

The Institute of Statistical Mathematics
Sakuragaoka Birijian 304, 15-8 Sakuragaoka,
Shibuya-ku, Tokyo 150-0031, Japan
(e-mail: kazue@med.Teikyo-u.ac.jp)

Abstract. The practical evaluation of data quality is discussed. Such evaluations are essential if we intend to carry out useful data analyses. Here we treat this problem in the context of cross-societal (comparative social) surveys.

1 Introduction

Data quality evaluation is crucial to data analysis if we are to draw out useful and relevant information. Analyses of low-quality data never bears fruit; however, data analytic methods can be refined. In spite of the importance of this issue in actual data mining and data analysis, I am forced to ask why this problem cannot be discussed at its most essential level. Perhaps it is a matter of the laborious practical work involved or the otherwise plodding pace of research. Indeed, data quality evaluation is rarely addressed because in academic circles it is regarded as unsophisticated. In the present paper, I dare to take up this problem, regarding it as one very important to data science. Here we address two problems in cross-societal sample surveys. The first concerns the translation of questionnaires. Suppose that a question written in Language A is translated into Language B. Then the question in Language B is translated back into Language A (by a different translator). The original question is then compared with the "back translation". To further assure the questions' similarity of meaning, a split-half survey is conducted; one-half receiving the original question and the other half the back-translated question, thereby investigating the questions' comparability. The second problem concerns the survey methodology, particularly the sampling and data collection methods used by survey companies. If strict random sampling is employed, few serious problems are likely to arise, apart from possible problems relating to the data collection method and procedure, non-response and response errors. In many countries, however, survey companies use only quota sampling, which gives rise to many technical problems in sampling design and data collection. Survey companies have their own particular skills and in many cases their techniques and know-how are considered proprietary and therefore kept confidential. But data quality depends upon these techniques and know-how. Some examples of these phenomena and problems are described in the following pages, focusing in particular on the differences one can observe in the results. For purposes of this study, the one-on-one, face-to-face

interview data collection procedure is adopted. In this study I will elucidate the fundamental idea of data science (Hayashi 1998), showing that design, data collection and analysis are all indispensable to data mining if we intend to obtain practically useful information.

2 Bias through question translation

In our comparative study, we use questions from several sources, based on the idea of Cultural Link Analysis (CLA) [Hayashi et. al. (1992), Hayashi (1998a), Hayashi (1998b)]. These sources include questions from continuing surveys of the Japanese National Character Study, GSS, ISR, CREDOC, EC and ALLBUS.¹ Pilot surveys were conducted by first translating each question into each nations' language. The questions were then translated back to the original language for comparison and contrast of the results. For example, an original English question would be translated into Japanese and then retranslated back into English for evaluation of any changes resulting from this process. Opinions and ideas were also exchanged between each nations' researchers regarding the results of the analysis of the pilot surveys as well as general aspects of comparative attitude studies. Through this process, the questionnaires were fixed in the relevant languages. We confirmed that some questions have no problems in translation, but that other questions seem to show degrees of congruity or dissonance. In the latter case, we adopted the following procedure. We prepared two kinds of questionnaires, Type A and Type B in Japanese. The Type A questionnaire included the problem-free questions and literal English-to-Japanese translations of the questions found to be incongruous. The Type B questionnaire included the same problem-free questions, the original Japanese questions, and literal translations into Japanese of the incongruous questions from all the languages, except English, of the questions showing incongruence with the English questions. Nationwide sample surveys by three-stage random sampling were performed using the split-half method for the two types of questionnaire A or B. In the following discussion, A denotes results from the Type A questionnaire, and B denotes results from the Type B questionnaire. The comparison of supported percentages of response categories in questions is shown in Fig.1 and Fig.2. In Figure 1, A and B are quite similar. The scatter may be due to the sampling fluctuations including sampling variance. In Figure 2, A and B are fairly similar, but certainly beyond expected sampling fluctuations, indicating that there is some survey bias. Generally speaking, these differences are at most

¹ The General Social Surveys of the National Opinion Research Center. The Institute for Social Research at the University of Michigan. Centre de Recherche pour L'Etude et L'Observation des Conditions de Vie, The "Eurobarometre" of the Commission of European Communities. Allegemeine Bevolkerungsumfrage der Sozialwissenschaften Mannheim; Zentrum fur Umfragen, Methoden und Analysen e.V.

15 %. What effect do these differences have on subsequent data analysis?

Here we shall describe the results of our seven-nation comparative survey,

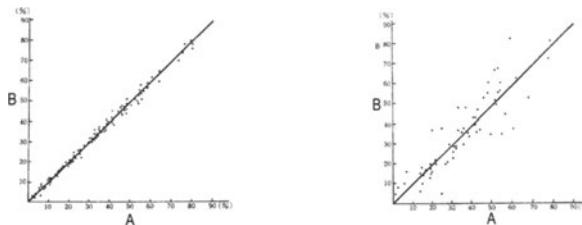


Fig. 1. Comparison of A and B: **Fig. 2.** Comparison of A and B:

Same Questions;

Incongruent Questions;

presenting, then, an example of data analysis of comparative surveys [Hayashi (1997), Research Committee(1998)]. The groups taken up in this instance include the Japanese, Americans, English, French, Germans, Dutch and Italians including Japanese-Americans and Japanese-Brazilians. It is important to add the groups of Japanese origin to seven nations surveys based on the idea of Cultural Link Analysis [Hayashi(1986), (1998b)]. Let the opinion distribution be given in each group in all questions. For the present purposes, we will focus mainly on only one key response category for each question (with some exceptions, e.g., those with more than two essential response categories or ones with key categories representing "scale" scores constructed for some questions). The number of categories is R. All questions, except those about personal characteristics (sex, age, education, etc.) are used. We can calculate the dissimilarity measure d_{ij} between the i th and j th nations as follows:

$$d_{ij} = (1/R) \sum_r^R |P_{ir} - P_{jr}|$$

where P_{ir} is the percentage of the j th nation on the key response to the r th categories. d_{ij} is a fuzzy measure of the difference between i and j . Thus we have a dissimilarity matrix between i and j . Based on this fuzzy dissimilarity matrix, we can apply a method of multidimensional analysis to create a graphic representation of groups. Specifically, the method is multidimensional data analysis (MDA-OR, for Minimum Dimension Analysis Ordered Class Belonging [(Hayashi, (1974)], a type of so-called multidimensional scaling (MDS). The idea is briefly explained as below. Without using d's as they are, d's are classified to several groups in order to minimize the variance within groups, while the distances between nations are considered in S dimensional Euclidean Space. The maximization of the measure of correspondences between these two informations (class belongings and distances) is carried out, under the practical conditions of the least S and the maximum correspondence measure.

The similar configurations of groups are obtained by the well-known idea of quantification method III [Hayashi (1956)] or correspondence analysis [Benzécri (1973), Lebart et al. (1984)], using the matrix of P's directly. Figure 3 and Figure 4 depict the results of the data analyses of only 'the Type A questionnaire', and 'the Type A and the Type B questionnaires', respectively. These are simple graphic summaries of the similarity relations. The degree of similarity is revealed as the distance in Euclidean space. Roughly speaking, consider that the distance corresponds to the similarity and that the configuration presents a reasonable summary of linked similarities. In Figure 3 and Figure 4, triangular relationships emerge. The location of A and B is very close in Fig.4, and the total configurations are quite similar in Fig.3 and Fig.4. The difference between A and B does not indicate any remarkable change in this data analysis.

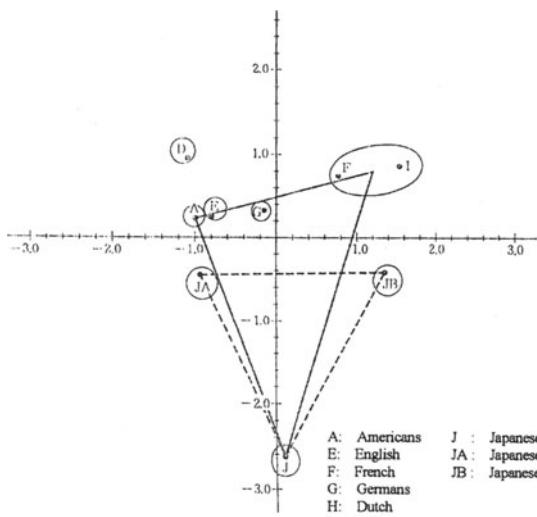


Fig. 3. Configuration of Seven Nations
with A in Japan

note) Solid lines are drawn to emboss three poles—Latin areas, English native language areas and Japan—while dashed lines are done to elucidate the relations of groups of Japanese-origin with three poles.

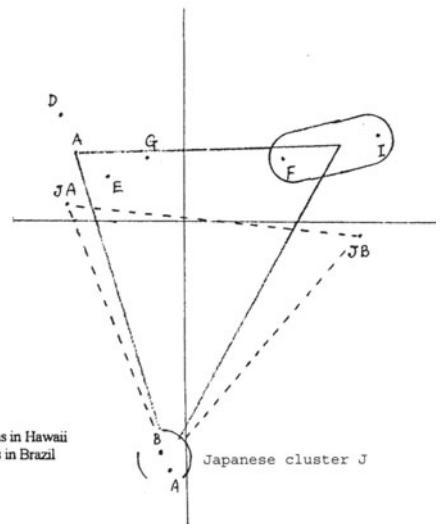


Fig. 4. Configuration of Seven Nations
with A and B in Japan

3 Bias by survey organizations

Here, only questions of interpersonal relations which seem to be consistent in time are analyzed.

3.1 Cases of random sample and quota sampling surveys in the United States

Three questions on trustfulness were used in ISR, Michigan. The ISR results (nation wide random sampling) are compared to those of Gallup (nation wide quota sampling). As can be seen in Figure 5, the results for these three questions were quite satisfactory in total [Hayashi et.al (1992)].

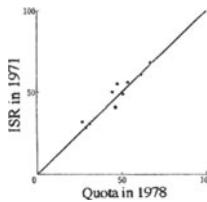


Fig. 5. Random Sampling and quota Sampling

3.2 Comparison of results of quota sampling comparative surveys

Here we take up the international surveys by quota sampling. Two surveys were conducted at different times in Germany, France, Italy, the Netherlands, the U.K. and the U.S. The same survey company in the Netherlands was used. The results are as shown in Figure 6, 7 and 8.

Some similarities are found. Generally speaking, the difference are at most

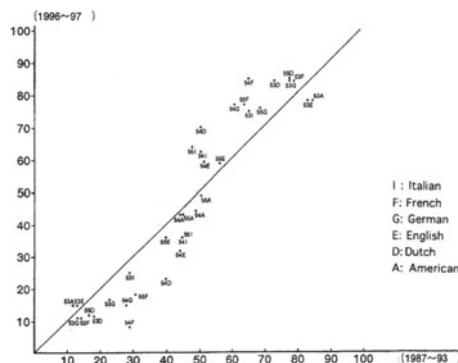


Fig. 6. Comparison between the Responses—Italy, France, the Netherlands, Germany, UK, USA—The symbols in the Figure show nations and question numbers, which have no meaning in the present discussion.

15%. Apart from detailed or formal treatments of the data, these results are satisfactory. The data analysis by quantification method III performed to determine the configuration of nations was applied to each time period, with

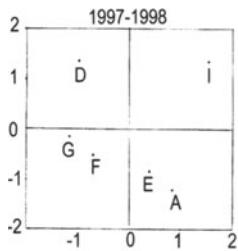


Fig. 7. Configuration of Six Nations in 1997-98

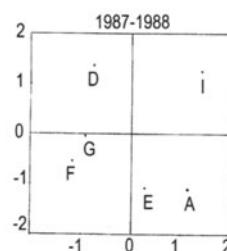


Fig. 8. Configuration of Six Nations in 1987-88

the results yielding consistent information as shown in Figure 7 and 8. This analysis indicates some practical applicability of the quota sampling method.

Concluding Remark

In our experience with comparative (cross-societal) surveys, data are satisfactory, generally speaking, if the survey procedure is carefully considered and carried out in a practical sense, given that we overlook some percentage differences in responses (e.g., less than 10%), even though these apparent differences would be significant under formal statistical testing, which does not take into consideration any survey bias.

References

- BENZÉCRI, J. P. (1973): *L'Analyse des Donnees*, Dunod.
- HAYASHI, C. (1956): Theory and Example of Quantification(II). *Proc. Inst. Statist. Math.*, 3, 69-98.
- HAYASHI, C. (1974): Minimum Dimensional Analysis MDA. *Behaviormetrika*, 1, 1-24.
- HAYASHI, C. and SUZUKI, T. (1986, 1997 Enlarged Edition), : *Data Analysis in Social Surveys*, Iwanami Shoten. The English Version by Hayashi, C. Suzuki, T. and Sasaki, M., "Data Analysis for Comparative Social Research: International Perspectives." was published by Elsevier, North-Holland in 1992.
- HAYASHI, C. (1998a): What is Data Science?, Fundamental Concepts and a Heuristic Example, In: C. Hayashi, et al. (Eds.): *Data Science, Classification and Related Methods*, Springer-Verlag Tokyo, 40-51.
- HAYASHI, C. (1998b): The Quantitative Study of National Character, Interchronological and International Perspectives, In: M. Sasaki(Ed.), *Values and Attitudes Across Nations and Time*, Brill, 91-114.
- LEBART, L. and MORINEAU, A. and WARWICK, K. M. (1984): *Multivariate Descriptive Statistical Analysis*, John Wiley.
- Research Committee on the Comparative Study of Native Character, the Institute of Statistical Mathematics (1998): *Comparative Study on Seven Nations' National Characters*, Idemitsu Shoten.

Collapsibility and Collapsing Multidimensional Contingency Tables—Perspectives and Implications*

Stefano De Cantis and Antonino M. Oliveri

Institute of Social Statistics, Demography and Biometrics, University of Palermo,
viale delle Scienze, 90128 Palermo, Italy **
(e-mail: decantis@iss.economia.unipa.it, oliveri@iss.economia.unipa.it)

Abstract. Collapsing multidimensional contingency tables is a necessary procedure in all kinds of research. Since collapsibility is subject to severe conditions, collapsing is often not admissible without incurring severe interpretative errors. After having discussed the main contributions to the statistical specification of the concept, we shall point out the logical conditions for collapsing multidimensional contingency tables.

1 Introduction

The discussion of the logical foundations of collapsing multidimensional contingency tables started in the first decades of this century when different definitions and meanings of interactions among variables were discussed (Yule (1903); Simpson (1951)), and when the conditions to verify the hypothesis of no second order interaction in three-way tables were investigated (Bartlett (1935); Darroch, (1962)).

Problems connected with collapsing multidimensional contingency tables and with collapsibility (the conditions to collapse) consist of "*determining when the size of the table may be reduced without distorting structural relationship between variables of interest*" (Bishop, et al., 1975, p. 31).

In very general terms, it will be possible to say that a contingency multi-way table is "*collapsible*" if all "*relevant information*" about its "*structure*" (information contents and, in particular, the relations among variables) is preserved in a lower-dimension table obtained by marginalizing the former table with respect to one or more variables or by condensing two or more categories.

The previous definition does not yet clarify what we mean by "*relevant information*" and "*information structure*". Indeed, the conditions for collapsibility proposed by researchers are different because of the different definition

* Thanks to G. Lovison whose seminar on *Introduction to Categorical Data Analysis*, made in February and March 1999 at the University of Palermo, represented the starting point of this paper.

** Although this paper is due to the common work of both authors, sections 1 and 3 are attributable to Stefano De Cantis, sections 2 and 4 to Antonino M. Oliveri.

of the two terms and because of different measures of association used to evaluate relations among variables or categories.

The advantages of collapsing multidimensional contingency tables consist of the possibility of fitting parsimonious models to data and consequently in the possibility of increasing the ratio between sample size and number of parameters to estimate.

2 Perspectives on collapsing and the applicability of the concept

Collapsing can be studied from at least two relevant perspectives:

1. from the methodological point of view, in particular, as regards the analytical formalization of the notion and the necessary and sufficient conditions to collapse (collapsibility);
2. from the applied point of view, in particular, regarding the interpretation of results of statistical data processing, the complexity and paradoxes of multivariate analysis and the logical difficulties of reducing dimension if related to the limits of the statistical operation of knowledge of reality (methodology of research).

In the last section, we shall discuss the second perspective. When the severe conditions under which collapsing is admissible are not satisfied, multivariate data analysis can incur paradoxical interpretations of results (Simpson (1951); Shapiro (1982), Good and Mittal (1987), Lovison and Roverato (1994)). Without control of all theoretically relevant dimensions (variables), every statistical analysis can be biased perhaps roughly but principally in a way that it is not possible to predict. This represents the inevitable logical limit of collapsing.

In the next section we shall briefly illustrate, from a historical point of view, different notions of collapsibility. We shall refer to two possible kinds of collapsing (Bishop (1971), p.546):

1. collapsing categories of a variable refers to conditions under which it is possible to aggregate modalities without affecting the structure of association within contingency tables. A historical example that clearly demonstrates the bias caused by condensing categories in the analysis of the structure of association is given by Bishop et al. (1975)

$$A = \begin{bmatrix} 4 & 2 & 6 \\ 6 & 3 & 9 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & 8 \\ 6 & 12 \end{bmatrix}, \quad C = \begin{bmatrix} 4 & 1 & 7 \\ 6 & 6 & 6 \end{bmatrix}.$$

Table B is obtained by collapsing (condensing) into one the last two columns of tables A and C. Tables A and B satisfy the condition of independence, but table C does not.

2. Collapsing variables refers to conditions requested to marginalize the conjunct distribution of a k-vector of variables over a subset of them.

The most famous result of collapsing the dimensions (variables) of a statistical analysis is probably the well-known "Yule-Simpson Paradox" (Yule (1903); Simpson (1951)), the most general version of which ("Amalgamation Paradox") is given by Good and Mittal (1987). It simply explains how information deriving from conditional associations cannot in general be reconstructed in a marginal table.

$$T_k = \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix}, \quad T = \begin{bmatrix} \sum_k a_k & \sum_k b_k \\ \sum_k c_k & \sum_k d_k \end{bmatrix}.$$

Let T_k be a 2×2 contingency table formed by two variables conditionally on the k -th category of a third variable, with $k = 1, \dots, K$.

Let a_k, b_k, c_k, d_k be the cell frequencies of T_k ; let T be a table obtained by the marginalization of (collapsing) T_k with respect to the categories of a variable, whose elements are $\sum_k a_k, \sum_k b_k, \sum_k c_k, \sum_k d_k$. Finally, let α be a measure of association; paradoxically it can alternatively be

$$\max_k \alpha(T_k) < \alpha(T), \quad \min_k \alpha(T_k) > \alpha(T).$$

3. It is also possible to think of collapsing over units with respect to sampling design¹: the analysis of this kind of collapsing seems promising and not yet deeply explored. We shall yet later refer only to the other kinds of collapsing.

3 Different notions of collapsibility

In the statistical literature the concept of collapsing (and its implications) has not always developed in a linear way, but rather by means of the confutation of some previous results. We shall later describe some notions of collapsing, by analyzing the conditions required for collapsing; in addition, we shall try to reveal their mutual differences by outlining the interpretative problems arising when the conditions for collapsing are not satisfied.

One of the first researchers trying to explain the misunderstanding between conditions for collapsing and no second order interaction in a three-way contingency table was Simpson (1951).

However the first who defined and used the terms "*collapsing*" and "*collapsibility*", was probably Bishop (1971). In 1975 (p. 47), Bishop, Fienberg and Holland enunciated two theorems, the most general of which asserted: "*Suppose the variables in a s-dimensional array are divided into three mutually exclusive groups. One group is collapsible with respect to the u-terms involving a second group, but not with respect to the u-terms involving only*

¹ The ideas of collapsing sampling design come from G. Lovison and from discussions that he solicited for a next paper.

the third group, if and only if the first two groups are independent of each other (i.e., the u-terms linking them are 0)". This theorem will be refuted because it defines false necessary and sufficient conditions. In fact, Whittemore (1978) demonstrates that conditions for collapsing multidimensional contingency tables are less restrictive. She also proposes the definition of "strict collapsibility".

Let m_{ijk} be the expected frequencies in the ijk cell of a three-way contingency table; and m_{ij} be the expected frequencies in the ij cell of a table collapsed on the third variable. The log-linear model for three and two-way tables are respectively:

$$\log(m_{ijk}) = \theta + \theta_i^1 + \theta_j^2 + \theta_k^3 + \theta_{ij}^{12} + \theta_{ik}^{13} + \theta_{jk}^{23} + \theta_{ijk}^{123},$$

$$\log(m_{ij}) = \lambda + \lambda_i^1 + \lambda_j^2 + \lambda_{ij}^{12},$$

θ and λ being parameters that are constrained to identify models. Bishop et al. define a three-way contingency table to be collapsible over the third variable with respect to the interaction of first and second variable if: $\lambda_{ij}^{12} = \theta_{ij}^{12} \forall ij$.

Since θ_{ij}^{12} is not a proper measure of association between the first and the second variable when $\theta_{ijk}^{123} \neq 0$, Whittemore defines conditions for *strict collapsibility*² adding up to $\lambda_{ij}^{12} = \theta_{ij}^{12} \forall ij$, $\theta_{ijk}^{123} = 0$.

The logical errors identified by Whittemore are the same as Simpson's and consist of mistaking "no interaction" for "collapsibility".

In 1986 (p. 198), Ducharme and LePage give the definition of "strong collapsibility": "a table that remains strictly collapsible no matter how the categories of the third variable are pooled together, that is no matter how it is partially collapsed, will be called strongly collapsible ... The important fact is that ... the structure association between the first two variables is totally independent of the levels of the third variable". Strong collapsibility implies strict collapsibility and then requires more restrictive conditions. Some authors (Aickin, (1983)) propose the notion of "pseudo-collapsibility" when conditions for strict collapsibility do not hold. The marginal measure of association is a mean of measures of conditional association and the loss of information is negligible because the variance of measures of conditional association is small. Ducharme and LePage (1986) present tests for the three conditions of collapsing (strong, strict, and pseudo-collapsibility) on the basis of odds ratios as a measure of association.

Aiming at the conditions to collapse (condensing) categories, Davis (1987, p.129) specifies the definition of *partial collapsibility* as a generalization of Whittemore's definition: "... from combining all levels of a factor to partially collapsing, i.e. combining only some levels of a factor"; she gives, in addition, necessary and sufficient conditions and analyzes the relation between

² Greenland et al. (1999) generalize Whittemore's definition to every measure of association.

partial collapsibility and conditional independence. Other contributions on properties of collapsibility regard log-linear models, measures of association, graphical models, metrical data analysis (linear regression models and generalized linear models)³.

4 Collapsing and the paradigm of "the statistical reduction of reality"

The necessity of considering only a (finite) set of variables by marginalizing over the infinite possible dimensions of reality, cannot be ignored in any data matrix. It is however impossible to verify collapsibility for variables excluded from the data matrix: the selection of variables for statistical analysis can be determined only on the basis of theoretical considerations on the nature of the studied phenomenon. The only real guarantee against biases caused by improper collapsing will be represented by the validity of theoretical models.

In extreme terms, it would even be possible to say that the researcher's main task is to evaluate in a qualitative way the hypothesis of collapsibility with regard to the (supposed) negligible dimensions.

What has so far been discussed was formalized by Dawid (1979, p.6) with his definition of a "*sufficient subset of covariates*": *"Let us label the individual units of the population by a variable I. We can consider the family of distributions for the response Y on unit I when treatment X is applied, as I and X vary... The hypothesis of no treatment effect, at the level of individual units, can be written as $Y \perp X | I$... Suppose that we have a set of covariates Z (a function of I). We say that Z is a sufficient set of covariates if $Y \perp I | (X, Z)$..."*. If this is valid, every further information about units is insignificant as regards the distribution of the response as to X when Z is known. Under the same condition, $Y \perp X | I$ equals $Y \perp X | Z$, and so it is reasonable to study the relation between Y and X without incurring biased interpretations. The knowledge of Z permits inference on the relation between Y and X conditional on Z, completely ignoring all variables not included in Z.

But, how can we be sure that Z is a sufficient set? Or that the infinite dimensions representing units in I can be reduced to the Z subset without any loss of information? Or, in the formerly used terms, that the theoretical model has been validly specified?

As we already said, this hypothesis is not verifiable. However, Dawid demonstrates that it is falsifiable. In fact, if W is a further set of covariates, we can falsify with the usual statistical procedures the hypothesis that $Y \perp W | (X, Z)$. Such falsification implies the falsification of the hypothesis that Z is a sufficient set. On the contrary, an eventual corroboration cannot say anything on the sufficiency of Z.

³ See Shapiro (1982), Greenland and Maldonado (1994), Frydemberg (1990), Geng and Asano (1993).

Someone could then think that continual corroboration of the hypothesis of the sufficiency of Z through the introduction of new W_1, W_2, \dots covariate groups is a criterion proper to accept it. Since such process cannot be endlessly extended, we must however agree with Dawid (1979, p. 7) who concludes: "At some stage it is necessary to make a subjective judgement that a set of covariates Z is sufficient."

References

- AICKIN, M. (1983): Linear statistical analysis of discrete data. *Wiley, New York*.
- BARTLETT, M.S. (1935): Contingency table interactions. *Suppl. Journal of the Royal Statistical Society, Series B*, 2 248-252.
- BISHOP, Y. M. M. (1971): Effects of collapsing multidimensional contingency tables. *Biometrics*, 27 545-562.
- BISHOP, Y. M. M., FIEMBERG, S. E., and HOLLAND P. W. (1975): Discrete Multivariate Analysis: Theory and Practice. *MIT Press, Cambridge, Mass.*
- DARROCH, J. N. (1962): Interactions in multi-factor contingency tables. *Journal of the Royal Statistical Society, Series B* 24 251-263.
- DAVIS, L. J. (1987): Partial collapsibility in multidimensional tables. *Statistics & Probability Letters*, 5 129-134.
- DAWID, A. P. (1979): Conditional independence in statistical theory (with discussion). *Journal of the Royal Statistical Society, Series B* 41 1-31.
- DUCHARME, G. R. and LEPAGE, Y. (1986): Testing collapsibility in contingency tables. *Journal of the Royal Statistical Society, Series B* 48 197-205.
- FRYDENBERG, M. (1990): Marginalization and collapsibility in graphical interaction models. *The Annals of Statistics*, 18 790-805.
- GENG, Z. and ASANO, C. (1993): Strong collapsibility of association measures in linear models. *Journal of the Royal Statistical Society, Series B* 55 741-747.
- GOOD, I. J. and MITTAL, Y. (1987): The amalgamation and geometry of two-by-two contingency tables. *The Annals of Statistics*, 15 694-711.
- GREENLAND, S. and MALDONADO, G. (1994): Inference on collapsibility in generalized linear models. *Biometrical Journal*, 36 771-782.
- GREENLAND, S., ROBINS, J.L. and PEARL, J. (1999): Confounding and collapsibility in causal inference. *Statistical Science*, 14 n.1 29-46.
- LOVISON, G. and ROVERATO, A. (1994): Grafo di indipendenza condizionata e suoi marginali: informatività, paradossi, e applicazioni. *Atti XXXVII SIS, CISU vol II* 237-244.
- SHAPIRO, S.H. (1982): Collapsing contingency tables - A geometric approach. *The American Statistician*, Vol. 36 n.1 43-46.
- SIMPSON, E. H. (1951): The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* 13 238-241.
- WHITTEMORE, A. S. (1978): Collapsibility of multidimensional contingency tables. *Journal of the Royal Statistical Society, Series B* 40 328-340.
- YULE, G. U. (1903): Notes on the theory of association of attributes in statistics. *Biometrika*, 2 121-134.

Data Collected on the Web

Vasja Vehovar¹, Katja Lozar Manfreda² and Zenel Batagelj³

¹ Faculty of Social Sciences, University of Ljubljana
Kardeljeva ploscad 5, 1000 Ljubljana, Slovenia
(e-mail vasja.vehovar@uni-lj.si)

² Faculty of Social Sciences, University of Ljubljana
Kardeljeva ploscad 5, 1000 Ljubljana, Slovenia
(e-mail katja.lozar@uni-lj.si)

³ CATI Center
Trzaska 2, 1000 Ljubljana, Slovenia
(e-mail zby@cati.si)

Abstract. Despite its relatively short existence Web-assisted data collection has already been widely applied and, increasingly, data analysts have to work with data collected on the Web. However, we are faced with relatively contradictory views on this data collection mode. In particular, with respect to Web surveys, opinions vary from a belief that the Web will revolutionise the survey industry to the opinion that this does not represent a valid mode of data collection. This paper provides an overview of the methodology of professional Web surveys. Three essential components (self-administration, HTML basis, and automatic transmission) of the Web survey mode are defined and separated from related solicitation and selection procedures. As an illustration, the segmentation/clustering of Internet users with respect to the survey mode (telephone and Web) is presented. In addition, new possibilities for data collection on the Web are discussed, particularly as concerns the data used in network and clustering analysis.

1 Introduction

Today, technology is rapidly broadening the possibilities for Web assisted data collection. However, despite the enthusiasm arising from the low cost and simplicity of Web survey procedures, we face severe criticism, usually due to the prevailing majority of today's 'non-professional' Web surveys (Onyshkevych and McIndoe 1999).

This paper addresses the problems of Web assisted data collection; however it primarily discusses Web surveys. We use the expression *Web surveys* for the specific data collection mode that is characterised by the following features: (1) the survey is based on computer-assisted self-administered interviews without the presence of an interviewer (CSAQ component), (2) computer questionnaires are based on HTML forms, usually presented in a standard Web browser (Web specific component), and (3) responses are transferred to the server through the electronic network, usually the Internet (network component). Web surveys belong to the same computer assisted family of survey modes as CAPI (computer assisted personal interviewing),

CATI (computer assisted telephone interviewing), CASI (computer assisted self-interviewing introduced by an interviewer) or CSAQ (computerized self-administered questionnaires, i.e. self-interviewing without the presence of an interviewer; deLeeuw and Nicholls 1996). In particular, Web surveys are a specific type of CSAQ, together with DBM (Disk by Mail), EMS (Electronic Mail Surveys), CAPAR (Computer Assisted Panel Research), TDE (Touch-tone Data Entry), VR (Voice Recognition), ASR (Automatic Speech Recognition) and Audio CSAQ.

A critical issue which often interferes with the perception of these surveys, is the technology. The progress in browsers, email software, the characteristics of transmission procedure, etc., is thus extremely important. Related limitations often determine the status of this survey mode. However, we understand that technology does not present an obstacle for a well designed Web survey in developed countries.

Another issue that interferes with the perception of this survey mode is the process of selecting the units. We believe that it is critical for an understanding of the Web survey mode to separate it from the related selection procedures, which are, therefore, discussed in greater detail. As an illustration, the segmentation/clustering of Internet users with respect to the interviewing mode in RIS surveys is also presented. The RIS surveys were conducted as part of the national project RIS - Research on Internet in Slovenia (<http://www.ris.org>), at the Faculty of Social Sciences, University of Ljubljana. At the end of the paper, the specific aspects of Web-assisted data collection used in network analysis are discussed.

2 Sample selection and solicitation

The target population can be invited to participate in a questionnaire on the Web in a general manner on the Internet (through banner ads, pop-up windows, general announcements in online forums and newsgroups, etc.) or through traditional media. We label these surveys as *unsolicited* Web surveys. With *solicited* Web surveys, a specific individual invitation - though not necessarily personalised - for participation is used, often by email, but also by telephone, mail or even in person. In addition, the individually tailored invitations - such as a pop-up window - on Web pages can also be treated as solicited Web surveys.

However, solicitation itself cannot provide scientific sampling; for this we need a probability sample selection applied to all units within a population. An example would be a survey of members of a professional organisation where a list of email addresses is available. However, even the randomised sample selection cannot provide probability samples when imperfect frames are used. Typical examples of imperfect frames are lists of email addresses available in public listings or obtained with 'spider-ing' (a special software

called "spider" searches for addresses that are published on WWW pages) and previously mentioned pop-up surveys on Web pages.

One solution to the problem of the sampling frame involves 'pre-recruited samples' (Watt 1997) obtained with other modes of interviewing, particularly by telephone (Farmer 1998; Flemming and Sonner 1999). However, not all Internet users use email and not all Internet users are willing to reveal their addresses. These percentages are usually unacceptably low for the standards of probability samples.

A similar approach is used in panel surveys: users with adequate characteristics are selected from panel databases of more than a million users and post-survey adjustments are made in order to match the target population. Several studies have been performed, showing mixed results as regards the success of these methods. For example, panel Web surveys with the thoughtful use of sophisticated weighting algorithms, enabled the extrapolation of findings not only to the general population of Internet users, but also to the general U.S. population (Terhanian and Black 1999). Similarly, Nadilo (1999) claims that Web surveys, despite some differences in absolute numbers, can lead to the same marketing decisions as offline studies. However, when evaluating the quality of data collected through a Web survey, not only the structure of the respondents should be taken into account, but also other sources of discrepancies. This was clearly shown in a comparative study of results from a mall and a Web marketing study (Wydra 1999), where respondents were very similar in their demographic characteristics, but differences in responses were dramatic.

3 Segmentation of internet users in the RIS 98 web survey

The design of the RIS 98 Web survey allowed us to study more profoundly the characteristics of Web respondents. Participants ($n=6500$) were selected and solicited in different ways (announcements in classical media, banners on the Web, login messages, emails to public email addresses, etc.). At the same time, a telephone survey was conducted that included the same questions among a representative sample of 1200 Internet users. The Web respondents were compared to the telephone ones and their social-demographic characteristics, knowledge of English, computer orientation and Internet usage were included in the cluster analysis. Clustering (k-means cluster method) was first performed on the telephone respondents and two distinct clusters were obtained (similar effects were also observed with clustering to more clusters). Centroids of these clusters were used as a base for clustering the Web respondents (without iteration).

Internet users as measured in the telephone survey can be classified into two clusters of almost identical size (45% vs. 55%): *intensive users* and *less intensive users*. However, in the Web survey the intensive users represent

65% of all respondents. Obviously, differences in the characteristics (Table 1) of intensive and less intensive users are extremely large. We can expect that they will also produce large differences in target variables. The selection procedures that were used in the Web survey, therefore, did not guarantee a representative sample of Internet users.

Table 1. Cluster's centroids from the telephone survey

	Intensive users	Less intensive users	Intensive users	Less intensive users
	<i>Standardized values</i>		<i>Non-standardized values</i>	
% Have own e-mail address	0.7442	-0.4959	95%	33%
Years of Internet use	0.4974	-0.3403	3.4	2.2
Frequency of use (1-used only once, 6-few times a day)	0.6327	-0.2642	4.04	2.92
% with access from home	0.5297	-0.4291	78%	30%
Knowledge of English (1-none, 5-fluent)	0.3652	-0.2445	4.01	3.37
Computer magazine reading (1-never, 5-regularly)	0.5914	-0.4080	3.41	2.15
Age	0.0424	-0.0902	32	30
Education	0.2508	-0.1740	4.76	4.1
% of men	0.4608	-0.3397	82%	44%
% of units in each cluster				
Telephone respondents	45%	55%		
Web respondents	65%	35%		

The respondents in the Web survey were weighted in order to get the same proportion of the two clusters as in the telephone sample. The idea is that, after weighting, the attitudes of Web and telephone respondents should become closer and more similar. However, results on a set of key variables measuring attitudes towards Internet shopping show that weighting does not greatly improve the values of the key variables. In most cases, attitudes are closer to those from the telephone sample; however they still differ significantly. For example, even after weighting (a variety of weighting methods were applied), the average on agreement ("I am extremely interested in on-line shopping") on a five-point scale demonstrates a difference between 2.5 (telephone respondents) versus 3.0 (Web respondents). Similar results were found for many social-demographic variables, page visitations and general attitudes.

Of course, with increased Internet penetration (in 1998 Internet penetration was 15% in Slovenia), results would improve and discrepancies would be reduced. However, the low response (much below 20%) in Web surveys will still present a dramatic difference compared to more aggressive inter-

viewing modes that can achieve over 60% response rates. The above example of the RIS survey thus provides a warning about any data collected with non-probability Web surveys.

4 New possibilities for collection of networked data

The above discussion has been concentrated on Web surveys of unrelated individuals. However, the Web can also be used for gathering other data, such as the characteristics (amount, type, distribution) of the communication between hosts (e.g. computers linked to the Internet). In addition, a variety of computer-to-computer techniques enable us to analyse data from the 'log' files where the information of every step during the Web survey data collection process is recorded. We thus know where (from which domain) the respondent is coming from, the exact timing of each question, request for help buttons, trials and errors, etc. This is valuable in the testing of the design usability of the Web questionnaires and also in defining the origin of the respondent. The ongoing international Web study (www.rine.org) within the RIS project analyses the origin (hosts) from which respondents come to the Web questionnaire. This enables evaluation of the usefulness of placing banner ads on different pages as well as determines the country of origin of the respondents.

A typical source of data that can be analysed with the methods of network analysis can also be provided with 'spiders' and similar tools that track data on Web sites or pages. Another study within the RIS project thus currently analyses the network of links between (Slovenian) Web sites. For this purpose, a spider automatically collects all the data about the hyper-links between Web sites and thus provides a large matrix for tools capable of analysing extensive networks.

The Web also enhances a specific method of sampling - snow-ball sampling (Erickson and Nosanchuk 1983). This can be quickly and cheaply performed in the Internet network environment. In this case, a small set of Internet users is invited to answer the Web questionnaire, but they also provide contact information (email address) for some of their colleagues. In the next step, these colleagues are invited to answer the questionnaire and again provide contact information for their colleagues. These steps can be repeated several times, until the desired sample size or sample characteristics are obtained.

5 Conclusion

We have demonstrated that the quality of data from Web surveys is strongly related to the selection procedures. These procedures sharply distinguish professional and non-probability or non-professional Web surveys. Proper statistical inference to the target population is not possible in the latter two cases. In addition, the mode effect itself may also influence data quality; however,

the majority of studies showed that such an effect was negligible. The Web survey mode itself thus presents a valid tool for data collection as long as probability sampling procedures are provided.

However, Web surveys are only one specific type of Web assisted data collection. The analysis of log files and different computer-to-computer techniques can also be extremely useful in providing research data.

In future we can expect a dramatic increase in self-administered survey modes, particularly Web surveys, and also an increase in automated data collection similar to log analysis. This can become additionally important when an increasing number of products, such as home appliances, will have an IP number attached, enabling us to permanently access and observe them over the network.

References

- DE LEEUW, E. D. and NICHOLLS, W. II (1996): Technological Innovations in Data Collection: Acceptance, Data Quality and Costs. *Sociological Research Online*, 4. <http://www.socresonline.org.uk/socresonline/1/4/leeuw.html>
- ERICKSON, B. H. and NOSANCHUK, T. A. (1983): Applied Network Sampling. *Social Networks*, 5, 367-382.
- FARMER, T. (1998): Using the Internet for Primary Research Data Collection. *Market Research Library*.
<http://www.researchinfo.com/library/infotek/index.shtml>
- FLEMMING, G. and SONNER, M. (1999): Can Internet Polling Work? Strategies for Conducting Public Opinion Surveys Online. Paper presented at the 1999 AAPOR Conference, St. Petersburg, Florida, USA, May 13-16, 1999.
- NADILLO, R. (1999): Online Research: The Methodology for the Next Millennium. In: *ARF's Online Research Day - Towards Validation*. Advertising Research Foundation, New York, 50-51.
- ONYSHKOVYCH, V. and McINDOE, D. (1999): Internet Technology: Gaining Commercial Advantage. Paper presented at the 1999 AAPOR Conference, St. Petersburg, Florida, USA, May 13-16, 1999.
http://www.ronincorp.com/GAININ_1/index.htm
- TERHANIAN, G. and BLACK, G. S. (1999): Understanding the Online Population: Lessons from the Harris Poll and the Harris Poll Online. In: *ARF's Online Research Day - Towards Validation*. Advertising Research Foundation, New York, 28-33.
- WATT, J. (1997): Using the Internet for Quantitative Survey Research. *Quirk's Marketing Research Review*, June/July. <http://www.quirks.com/CGI-BIN/SM40i.exe?docid=3000:58911&%70assArticleId=248>
- WYDRA, D. (1999): Online Tracking: A New Frontier. In: *ARF's Online Research Day - Towards Validation*. Advertising Research Foundation, New York, 34-36.
- ZUKERBERG, A., NICHOLS, E., and TEDESCO, H. (1999): Designing Surveys for the Next Millennium: Internet Questionnaire Design Issues. Paper presented at the 1999 AAPOR Conference, St. Petersburg, Florida, USA, May 13-16, 1999. <http://surveys.over.net/method/zukerberg.ZIP>

Some Experimental Surveys on the WWW Environments in Japan

Osamu Yoshimura¹ and Noboru Ohsumi²

¹ Okayama University

3-1-1 Tsushima-naka Okayama-shi 700-8530, Japan
(e-mail: osamu@cc.okayama-u.ac.jp)

² The Institute of Statistical Mathematics

4-6-7 Minami-Azabu, Minato-ku Tokyo 106-8569, Japan
(e-mail: ohsumi@ism.ac.jp)

Abstract. To assess and analyze the characteristics of surveys on the World Wide Web as objectively as possible, we simultaneously conducted some experimental surveys on three Web sites and two ordinary surveys. A comparison of survey results revealed some interesting characteristics of surveys conducted on the Web. There were stable, uniform and systematically biased responses among the three Web sites surveyed, in spite of the low response rates. In addition, respondents to the Web surveys had a general tendency to participate in surveys conducted through the WWW. The findings imply that in Web surveys, it may be feasible and beneficial to conduct longitudinal surveys.

1 Background and objective of the study

In Japan, World Wide Web surveys have suddenly become popular without enough discussion about ‘what a Web survey is’ or ‘how the survey should be conducted’. As a result, surveys have been conducted not only by researchers, but also by corporations or individuals who, although familiar with the use of the Internet, are not specialized in research. This has led to the present chaotic situation where the activity of scientific research is confused with the mere collection or retrieval of information.

Taking into account this situation, in 1997, we conducted 12 trial surveys on the Web with the cooperation of a survey company, to learn about what would be observed when a survey was conducted using the Internet (Ohsumi, 1997, 1998; Yoshimura et al., 1998). The findings of the surveys led us to conclude that it would be necessary to compare Web surveys at different sites in order to inquire further into characteristics of surveys on the Web.

In this study, we conducted four successive Web surveys at three distinct web sites, using the same questions at each site. These surveys were compared with two ordinary sample surveys. The main points of the survey plan were:

1. Comparing the results of Web surveys administered almost simultaneously at three different Web sites and in which the same questionnaires were used.

2. Conducting the survey four times, with the fourth a repetition of the first survey.
3. Conducting two ordinary surveys (for example, omnibus surveys with interviewing) at two different sites at about the same time, using questionnaires as similar as possible to those used in the Web surveys.

1.1 Types of web-based surveys in Japan

To situate our experimental Web survey on the spectrum of contemporary Web survey types, we have classified existing Web surveys in Japan into three types according to their methods of securing respondents.

Type 1 – Panel style: Finds registrants by ‘want ad’ or announcement on the Web, and conducts several successive surveys targeting all registrants. The number of respondents obtained through this technique would be several thousand.

Type 2 – Resource style: Finds registrants by ‘want ad’ or announcement on the Web, and selects actual targets from among them. The number of respondents may vary from 10,000 to more than 100,000. This is the main type used in Web-based survey services and is classified into the following methods:

- a) *Intra-resource open method:* Asks the registrants for cooperation through banner ads or other means, but does not request each of the registrants to participate.
- b) *Attribute-narrowing-down method:* Narrows down the survey population by attributes including gender, age or vocation. Sends e-mail requesting cooperation. Often halts the survey when the desired number of answers is attained.
- c) *Sampling method:* Selects respondents at random from among the registrants. Sends e-mail requesting cooperation.

Type 3 – Open style: Publishes the questionnaires on the Web and asks for cooperation by banner ads or other means. Does not sample individuals. Often used in Internet user-profile surveys conducted by sites well known for their search services.

2 Method

2.1 Survey methods

The actual surveys were carried out with the collaboration of companies A, B, and C, each with Web survey environments of their own. Company D uses a survey system with some answer-only communication devices connected to telephone lines. The methods used (types of Web surveys) and the target respondents for each site are as follows:

Company A: Four Web surveys, Panel style, 4,000 registrants.

Company B: Four Web surveys, Resource style with sampling method, random sample of 5,000 selected from 21,867 registrants for each of the four surveys.

Company B: Three conventional sample interview surveys, with random samples of 1,075, 900 and 900 drawn from population of eligible voters living within 30 km of the Tokyo metropolitan area.

Company C: Four Web surveys, Resource style with sampling method, random sample of 10,000 drawn from 55,714 registrants for each of the four surveys.

Company D: Two conventional sample surveys using answer-only communication devices installed in homes, random sample of 750 drawn from population of eligible voters living within 30 km of the Tokyo metropolitan area for each of the two surveys.

2.2 Survey periods and questionnaires

The Web surveys were conducted four times, each for the duration of at least one week, and within the same time period, from February to March 1999. The themes of the four successive surveys were: 'Awareness of daily life' (*the first survey*), 'The Internet environments' (*the second survey*), 'various commercial products and services' (*the third survey*), and a repetition of the first survey (*the fourth survey*). The second survey assumed respondents use the Web daily, so the same questionnaire cannot be used in ordinary sample surveys.

2.3 Some notes on each survey

The Web surveys on Sites B and C employed the intra-resource sampling method, where respondents were randomly sampled from the database of registrants on the server machine. That is, all the registrants were assumed to be a discrete pseudo-population, from which three kinds of schedule samples were randomly extracted. Where registrants were included in more than one sample, we referred to these as 'overlapped samples'. A request was made to registrants in each of the three samples to participate in the first, second and third surveys, and to the registrants participating in the first survey to take part in the fourth survey, which was a repetition of the first. For the panel-style survey on Site A, we requested all the registrants to participate as respondents in every survey.

3 Survey results

3.1 Trends in response rates

We first examined the trends in response rates and re-response rates—one of the most important points for Web surveys. In each of the Web surveys,

the response rate was below 20%, and for every site, particularly Sites B and C, the response rate for the first survey was the highest; the response rates for the second and the third surveys were lower. This is partly because the questionnaire was longer in the second and the third surveys.

Re-response rate is defined as the response rate where the respondents of the first survey also become respondents in the fourth survey. In these cases, the re-response rate was high. Re-response rates for Sites A, B and C were about 64.0%, 71.4% and 69.9%, respectively. Members of an ‘overlapped sample’ were invited to participate in more than two different surveys, located on Sites B and C. The rate of the virtual respondents within an overlapped sample, calculated from the results of four surveys, is shown below (in parentheses). As a reference, the rate of the virtual respondents for the surveys on Site A are also shown, where all the registrants were asked to participate in all four surveys:

Site B: Requested twice (25.2%), three times (29.7%, 29.5%), four times (34.3%).

Site C: Requested twice (13.9%), three times (17.9%, 17.3%), four times (21.5%).

Site A: Requested four times (30.7%).

This result shows that over 70% of registrants did not respond to any of the four survey invitations. It must be noted that tens of thousands of registrants will not necessarily yield the same number of opinions.

3.2 Some other characteristics of the surveys

We also encountered phenomena that should be considered, although are difficult to deal with.

(1) *Undelivered mail*: Throughout the surveys on Site B about 15% of mail messages were undelivered.

(2) *Multiple responses*: Multiple response means that the same respondent gives a response several times in one survey. The survey results for Sites A and B show that there were about 5% multiple responses.

(3) *Non-registrant responses*: In the surveys on Site B, although a few responses from non-registrants were found, the rate was not large overall. In the surveys on Sites A and C, in which respondents are cross-checked with the registration information on the databases and identified after they have accessed the Web pages, there were no such responses.

(4) *Systematic bias between schedule and collected samples*: For each of Sites A, B, and C, the response rate of the 30–40 year age cohort was greater than that in the schedule samples.

(5) *Differences among demographic items*: Comparing the registered and collected samples for the demographic items on each site, we could not determine whether variations occurred by mistake or on purpose. However, for every site, a few respondents had altered some of their registered demographic details.

3.3 Typical personality characteristics of the respondents

Specific tendencies and features found in the answers to questionnaires quoted from other surveys led us to imagine the typical respondent's personality as follows:

- not satisfied in his or her present state (about life style, life stage, and so on);
- has high regard for his or her own hobbies or tastes;
- prefers simple or casual human relations to intimate ones; and
- has high confidence in, or expectations about technology.

Generally, respondents seemed to be more self-centered, or concerned with their own actions, than self-helpful, or wishing to achieve some benefit for themselves. Even though they are likely to pursue their own advantage, they do not appear to be fundamentally self-helpful people.

3.4 Survey over-participation in surveys

Respondents were asked how frequently they participated in research or questionnaire surveys. Most respondents answered 'Once a month or more': 63.6% for Site B, 77.4% for Site C, and 79.7% for Site A. As for the question about their registration, more than 10 respondents who participated in the Site A surveys were also registrants of the Site C survey. Taking this into consideration, as well as the fact that the rate of participation by virtual respondents is about 30%, we can see that an unexpectedly limited number of people participated in various surveys and made repeated responses.

4 Conclusion and future directions of Web survey

Our results clearly show that Web surveys have problems of identifying respondents and establishing representativeness of survey samples. The respondents to Web surveys seem to be neither representative of the general population, nor of registrants of a Web survey service site. However, if we accept that it is possible to discuss the effective and practical use of Web surveys in spite of such problems, we must at least consider the following.

(1) *Incentives and the size of questionnaires*: Too many questionnaires with poor incentives produce negative reactions among registrants. If they feel that sending their answers costs them too much, they may try to recoup their losses. However, that does not mean that excessive incentives are preferable, as this could endanger the reliability of the survey results.

(2) *Allaying distrust*: The respondents seem to have much greater distrust of the Internet than might be expected. In response to the question 'About the information distribution on the Net' in the second survey, many expressed hope for some limitation to anonymity, or recognition of their input, and some regulation of the uses of the Internet. Further, responses to the survey question about the conditions necessary for agreeing to participate in the Web

surveys were: 'The researchers are reliable' (60%) and 'The aim and objective of the survey is understandable' (70%). To obtain reliable results through Internet surveys, there must be mutual trust between survey researchers and respondents.

(3) *Disclosure of survey results*: More than 40% of the respondents from the second survey indicated that to be informed of the results was one of the necessary conditions for participating in surveys. The rate was as high as that to the option 'Not so many questions'.

(4) *Identification of respondents*: Many Web surveys use e-mail addresses for identifying respondents. However, our survey results showed that an e-mail address cannot necessarily identify a particular person, because:

- less than 20% of respondents had only one e-mail address; and
- about 20% of respondents shared an e-mail address with others.

Therefore, we must seek some means of tracing back and identifying respondents, such as sending requests for participation by mail.

(5) *Problems caused by conflicts among surveys by different sites*: Our results show that several sites are sharing comparatively few groups of respondents. For respondents, the sites that can promise great benefits at low cost are preferable. At present, the sites seem to be competing for registrants, but when it comes to the quality of survey results, they will be competing for a higher response rate. We are afraid that a competition to provide incentives may cause a serious deterioration in the environment. It may become necessary for incentives to be regulated in some way.

In conclusion, we propose that, to appropriately interpret and use survey results, it is necessary to understand the characteristics of respondents and how typical they are of the Internet user population on occasions when surveys are taken. In this sense, we need 'longitudinal surveys' to clarify the characteristics of the respondents on the Web, rather than a single-shot survey seeking ad hoc responses.

References

- OHSUMI, N. (1997, 1998): A Study on New Survey Methods for the Changes in Survey Environments. A research report of Micro Statistic Data Research of Priority Field of Scientific Research Expenditure of the Ministry of Education.
- OHSUMI, N. and YOSHIMURA, O. (1999): The Online Survey in Japan: An Evaluation of Emerging Methodologies. *Bulletin of the International Statistical Institute 52nd Session, Book-2*, 171–174.
- YOSHIMURA, O., OHSUMI, N., KAWAURA, Y. et al. (1998): Some Experimental Trials of Electronic Surveys on Internet Environments. In: A. Rizzi, M. Vichi, and H.-H. Bock (Eds.): *Advances in Data Science and Classification*. Springer-Verlag, Heidelberg, 663–668.

Bootstrap Goodness-of-fit Tests for Complex Survey Samples

Andrea Scagni

Dipartimento di Statistica e Matematica applicata alle Scienze Umane
Universita' di Torino, Piazza Arbarello 8 10121 Torino - Italy
(e-mail: scagni@cisi.unito.it)

Abstract. A method to implement goodness of fit tests in survey sampling, where the common independence assumption fails, is proposed. A bootstrap test following an approach similar to those of Bickel and Freedman (1984), Rao and Wu (1988), Sitter (1992) is defined and applied to stratified and two stage sampling. Extensive MonteCarlo simulations show the good behavior of the test and the conditions to achieve satisfactory power levels against reasonable alternatives.

1 Introduction

The use of the X^2 goodness of fit test on non-i.i.d. data is quite commonplace in human sciences, where finite populations are often sampled according to survey sampling (SS) schemes usually very different from bernoullian sampling. In such cases the independence condition among observations fails, even though it is necessary for the distribution $G_0(X^2)$ of X^2 under H_0 to converge to χ^2 . Use of the test then implies an actual significance level which is different from the nominal chosen level. To lessen the degree of such inaccuracy several authors have proposed various corrections to X^2 so that its distribution is more similar to χ^2 (in terms of moments). However these approaches require the potentially difficult estimation of the frequency estimates covariance matrix (Holt, Scott and Ewings, 1980; Rao and Scott, 1981, 1984). In this paper a bootstrap approach is used to avoid these difficulties and to obtain a direct estimate of $G_0(X^2)$. However, standard bootstrap techniques are inappropriate for SS, since the bernoullian resampling procedure would be different from the sample survey methods used to obtain the original data. Several authors have proposed modified bootstrap procedures (Bickel and Freedman, 1984; Rao and Wu, 1988; Sitter, 1992). The bootstrap method developed here is in part similar to these, but a different approach suitable for goodness-of-fit testing is pursued. Specifically, stratified and two stages sampling schemes are considered, appropriate bootstrap procedures are defined for them and empirically evaluated. To this end null hypotheses of Uniform and Normal distribution are considered, and the accuracy of the bootstrap tests is appraised with repeated simulation of the bootstrap. The power of the bootstrap tests is also investigated, comparing the Uniform to a Beta and the Normal to a generalized Gamma distribution.

2 The bootstrap X^2 test on SS

Let X be a random variable with unknown distribution $F(X)$. A goodness of fit test is used for $H_0: F(X) = F_0(X)$ where $F_0(X)$ can be totally specified (H_0 simple) or can be a distribution family (H_0 composite). To test this hypothesis with X^2 :

1. a partition of the domain of X in k intervals C_1, C_2, \dots, C_k is created and the number f_j of sample observations $x_i \in C_j$ is defined;
2. if H_0 is simple, the expected number of observations $x_i \in C_j$ ($j=1, 2, \dots, k$) $f_j^o = np_j^o$ ($p_j^o = \Pr\{X \in C_j | H_0\}$) according to $F_0(X)$ is computed; if H_0 is composite, the expected number of observations $x_i \in C_j$ ($j=1, 2, \dots, k$) is estimated based on $\hat{F}_0(X)$, containing appropriate estimates of the parameters not specified in H_0 .

Under i.i.d. sampling $G_0(X^2)$, the distribution of $X^2 = \sum_{j=1}^k \left[(f_j - f_j^o)^2 / f_j^o \right]$ (under H_0 , with $n \rightarrow \infty$, fixed k and $\min_{1 \leq j \leq k} np_j^o \xrightarrow[n \rightarrow \infty]{} \infty$) converges to the

χ^2 distribution.

When observations are not i.i.d. $G_0(X^2)$ behaves less simply. Let:

- $\mathbf{p}_0 = \mathbf{f}_j^o/n$ the expected relative frequency vector based on H_0 ;
- $\mathbf{P}_0 = \text{diag}(\mathbf{p}_0) - \mathbf{p}_0 \mathbf{p}_0'$ the covariance matrix of \mathbf{f}_j under i.i.d. sampling, and \mathbf{V} the same matrix under the chosen SS scheme.

Then the distribution of $G_0(X^2)$ converges to that of $\sum_{i=1}^r \lambda_i \chi_i^2$ (Holt, Scott and Ewings, 1980) where the λ_i are the eigenvalues of $\mathbf{P}_0^{-1} \mathbf{V}$. The bootstrap method allows the use of the X^2 test without knowledge of $\mathbf{P}_0^{-1} \mathbf{V}$. As stressed by Efron and Tibshirani (1993, p. 233), bootstrap tests are convenient when H_1 is non-parametric and vague, as is usually the case with X^2 . However the complex survey context requires a modification of classic bootstrap as suggested by Sitter (1992). A real empirical pseudo-population based on both the actual sample and H_0 must be built, and the original sampling method for resampling must be used. Figure 1 shows a diagram of the proposed procedure.

3 Stratified and two stages sampling

When *stratified sampling* is the case a partition of the units of population U is defined, and each strata is sampled separately with fixed sub-sample size. Then $G_0(X^2)$ is strongly dependent on how the N population units are assigned to the L strata. To simplify, consider only strata of the same size. Let $\text{St}_L = \{s_1, s_2, \dots, s_L\}$ be a stratification procedure, and let N/L be the strata sizes, with $\bigcup_{h=1}^L s_h = U$, $\bigcap_{h=1}^L s_h = \emptyset$; the optimal stratification St_L^* is obtained when, $x_{[i]}$ being the ordered population values of X , $s_h = \{x_{[1+(h-1)N/L]}, x_{[2+(h-1)N/L]}, \dots, x_{[hN/L]}\}$; $h = 1, 2, \dots, L$. Then $\frac{\sum_{h=1}^L \sigma_h^2(\text{St}_L^*) N_h}{N} \leq \frac{\sum_{h=1}^L \sigma_h^2(\text{St}_L) N_h}{N}$, where $\sigma_h^2(\text{St}_L)$ is the variance in strata

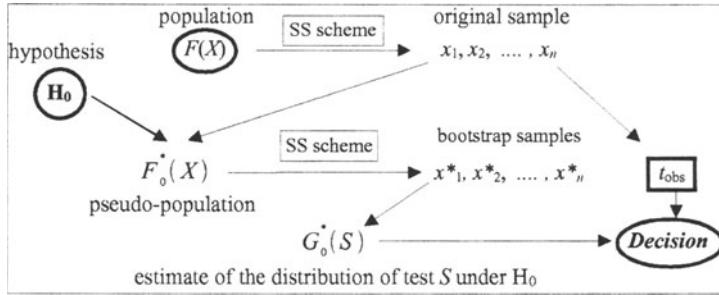


Fig. 1. Bootstrap hypothesis testing on SS - general scheme

h with stratification procedure St_L . With optimal or quasi-optimal stratification the X^2 is usually much more powerful than on i.i.d. observations. A wide range of stratification procedures have been considered here, from the optimal to the useless stratification, where all population units are randomly allocated to the strata. Sub-optimal cases are obtained as mixtures of the two extreme options.

Two stage sampling is defined as follows. Let $\text{Gr}_N = \{\gamma_1, \gamma_2, \dots, \gamma_N\}$ a partition of the population units in N blocks γ_h , each with size M_h ($h = 1, 2, \dots, N$). The total population size is then $N_T = \sum_{h=1}^N M_h$. In the first stage $n < N$ blocks $\gamma_h^{(1)}$ ($h = 1, 2, \dots, n$) are sampled without replacement, while in the second stage from each of the latter $m_h < M_h$ units are sampled without replacement. The total sample size is then $n_T = \sum_{h=1}^n m_h$. Again, X^2 power depends on the type of the actual partition in blocks. As above consider only equal-size blocks, all with size M . The optimal block partition Gr_{N*} has $\gamma_h = \{x_{[(i-1)*N+h]}, \forall i = 1, 2, \dots, N_T/N\} : h = 1, 2, \dots, N$. If $\mu_h(\text{Gr}_N)$, $\sigma_h^2(\text{Gr}_N)$ are the mean and variance of block h under Gr_N , then:

$$\text{Var}(\mu_h(\text{Gr}_{N*})) \leq \text{Var}(\mu_h(\text{Gr}_N)); \sigma_h^2(\text{Gr}_{N*}) = \sigma_h^2(\text{Gr}_N). \forall h, k$$

Sub-optimal block partitions are obtained from the optimal one reallocating a certain percentage $c\%$ of units to blocks as with optimal strata: $\gamma_h = \{x_{[1+(h-1)N_T/N]}, x_{[2+(h-1)N_T/N]}, \dots, x_{[(cN_T/N+(h-1)N_T/N)}; x_{[(i-1)*N+h]}, \forall i = cN_T/N + 1, cN_T/N + 2, \dots, N_T/N\}; h = 1, 2, \dots, N$.

4 Bootstrap procedures

Consider stratified sampling, letting f_{jh} be the number of sampled x s $\in C_j$ that come from strata h , and define $I(f_{jh}) = \begin{cases} 1 & \text{if } f_{jh} > 0 \\ 0 & \text{if } f_{jh} = 0 \end{cases}$. The pseudo-population to be built for resampling cannot in general conform to H_0 and still have the same strata sizes of the actual population (that is, unless $Np_j^o \geq \sum_{h=1}^L N_h I(f_{jh}), \forall j$). However, applying an iterative proportional fitting algorithm conformity to H_0 and pseudo-population strata sizes closest

(in MSE sense) to those of the actual population can be achieved. Let r_{jh} be the number of replications of each observation $\in C_j$ and coming from strata h , so that $N = \sum_{j=1}^k \sum_{h=1}^L r_h f_{jh}$. Also, let $r_{jh}^{(i;j)}$ be the number of replications suggested by sub-step j of step i of the algorithm¹. The procedure is then as follows, except for sub-step 1 of step 1, where starting values are set equal to $r_{jh}^{(1;1)} = f_{jh}N/n$:

- sub-step $i_1 (i \neq 1) \Rightarrow r_{jh}^{(i;1)} = r_{jh}^{(i-1;2)} N_h \left(\sum_{l=1}^k r_{lh}^{(i-1;2)} f_{lh} \right)^{-1}$;
- sub-step $i_2 \Rightarrow r_{jh}^{(i;2)} = r_{jh}^{(i;1)} N p_j^o \left(\sum_{l=1}^L r_{jl}^{(i;1)} f_{jl} \right)^{-1}$;
- sub-step i_3 (*stop check*): for a fixed $\delta > 0$, compute $D_{ij} = \left[\sum_{l=1}^k \left(r_{lh}^{(i-1;2)} - r_{lh}^{(i;2)} \right) f_{lh} \right] \left(N_h \sum_{l=1}^k r_{lh}^{(i-1;2)} f_{lh} \right)^{-1}$;
 • if $\exists j \Rightarrow D_{ij} \geq \delta$ go again to sub-step i_1 ;
 • if $D_{ij} < \delta \forall j$, the pseudo-population is created replicating each sample value $r_{jh}^{(i;2)}$ times (possibly with randomization).

Now consider two stage sampling. Here the pseudo-population must be created in two stages as well. In the first a complete pseudo-block $\gamma_h^{(1)*}$ is obtained for each first stage sampled block $\gamma_h^{(1)}$ by replicating its sample units as in the stratified case. Let f_{jh} be the second stage frequency for interval C_j on sampled block $\gamma_h^{(1)}$; if $f_{jh} > 0, \forall j, \forall h: \gamma_h = \gamma_h^{(1)}$, then $x_{ih} \in C_j$, is replicated $M_h p_j^o / f_{jh}$ times (randomization), so that each $\gamma_h^{(1)*}$ conforms to H_0 . However, if the condition above does not hold, $\gamma_h^{(1)*}$ conforming to H_0 and with sizes equal to $\gamma_h^{(1)}$ can be formed for each C_j only if $p_j^o \sum_{\gamma_h^{(1)}} M_h \geq \sum_{\gamma_h^{(1)}} M_h I(f_{jh})$. Otherwise the same iterative procedure used above for stratified sampling can be applied to create pseudo-blocks $\gamma_h^{(1)*}$ similar to $\gamma_h^{(1)}$.

In the second stage all remaining $N - n\gamma_h^{(2)}$ (totally unobserved) blocks must be recreated, depending on the kind of block partition of the population:

Optimal or quasi-optimal block partition. A partition of all observations from the n pseudo-blocks $\gamma_h^{(1)*}$ based on C_1, C_2, \dots, C_k is created; the sizes of subsets are $p_j^o \sum_{\gamma_h^{(1)}} m_h, j = 1, 2, \dots, k$ (conforming to H_0). Each pseudo-block $\gamma_h^{(2)*}$ is created by sampling without replacement $p_j^o m_h$ (randomization) units from those in the subset defined by $C_h, \forall h = 1, 2, \dots, k$.

Strongly sub-optimal block partition. Since blocks are dishomogeneous, replicating the $\gamma_h^{(2)}$ blocks from the mixture of the $\gamma_h^{(1)*}$ pseudo-blocks causes a distortion of $F^*(X)$ w.r. to $F(X)$. It is then better to separately replicate each single $\gamma_h^{(1)*}$ block $(N - n)/n$ times (randomization). Then the pseudo-population conforms to H_0 , and it mirrors the peculiarities of whole blocks.

¹ If the number of replications r is not an integer, its *randomization* will be used, i.e. an occurrence of the r. v. $\xi = \{\lceil r \rceil, \Pr(\lceil r \rceil) = p; \lfloor r \rfloor, \Pr(\lfloor r \rfloor) = 1 - p\}$ such that $E(\xi) = r$, where $\lceil r \rceil$ is the integer above r and $\lfloor r \rfloor$ is the integer below r .

After obtaining B bootstrap samples from the appropriate pseudo-population, the corresponding test values X_b^{2*} are computed and their empirical distribution $G_0^*(X^2)$ derived; the critical region is limited on the left by the $(1 - \alpha)\%$ percentile of $G_0^*(X^2)$. H_0 is rejected if the observed value X^2 lies in the critical region.

5 The simulation study

The performance of the X^2 bootstrap test on SS has been checked empirically w.r. to both accuracy (equivalence of nominal and actual significance level) and power. Two distributions were considered as H_0 , the Uniform and the Standard Normal, with appropriate Beta and Generalized Gamma alternatives for power evaluation. A finite population of 10000 units conforming to H_0 was created and simulations were conducted for a range of values of k , n , B (further details of a similar study can be found on Scagni, 1999). The main results for the X^2 bootstrap test are as follows:

- good all-round level of accuracy on stratified and two stage sampling;
- strong dependency of accuracy on two stage sampling on the type of block partition and on the way the pseudo-population is created;
- good power when compared to that of the exact X^2 test on stratified sampling based on reasonably good stratification;
- good power when compared to that of the exact X^2 test on two stage sampling based on reasonably good block partitions.

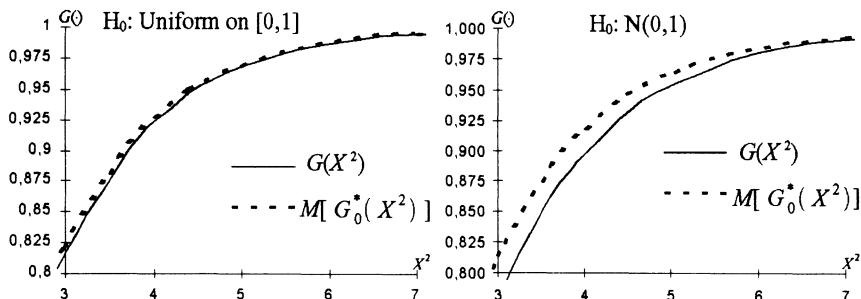


Fig.2. Accuracy in stratified sampling: H_0 : Uniform (left) and $N(0,1)$ (right)

Figure 2 shows accuracy in stratified sampling comparing the upper part ($G_0(\cdot) > 0.8$) of $G_0(X^2)$ (computed on 50000 simulations) with the mean of 100 bootstrap estimates $G_0^*(X^2)$. In the $H_0: N(0,1)$ case a bootstrap calibration procedure (Scagni, 1999) was applied and successfully eliminated the small distortion, bringing the actual significance level for $\alpha=0.05$ and 0.1 almost exactly to the chosen values.

Fig. 3 shows the behavior of mixture replication (MR) and single block replication (SBR) with two stage sampling in a quasi-optimal block partition case and in a strongly sub-optimal case (left and right side respectively). The need to use the appropriate second stage replication methods is apparent.

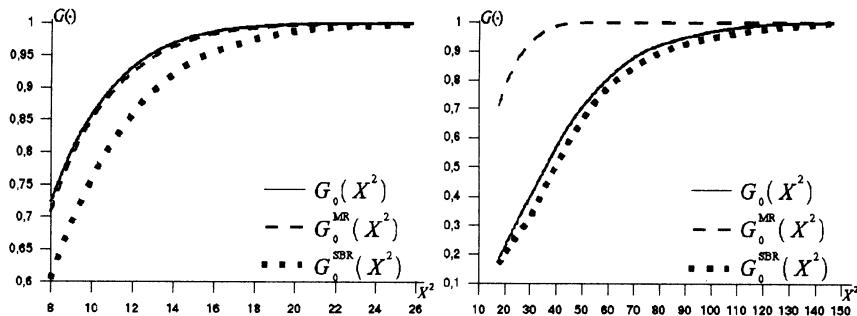


Fig.3. Accuracy, two stage sampling, optimal (left) and sub-optimal (right) block partition

Finally, Figure 4 shows how the power of the X^2 bootstrap test is only slightly lower than that of the corresponding exact test in most cases with two stage sampling.

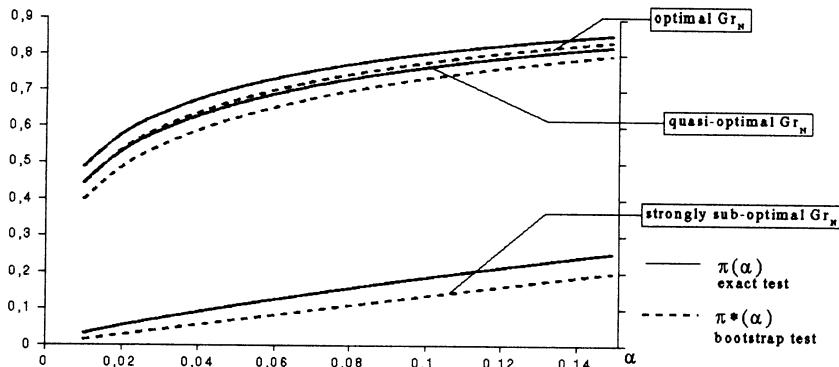


Fig.4. Power of the X^2 bootstrap test in two stage sampling, $H_0: N(0,1)$

References

- BICKEL, P.J. and FREEDMAN, D.A. (1984): Asymptotic normality and the bootstrap in stratified sampling. *Ann. Statist.* 12, 470-482.
 CHAO, M.T. and LO, S.H. (1985): A bootstrap method for finite populations. *Sankhya A*, 47, 399-405.

- EFRON, B. and TIBSHIRANI R. (1993): *An introduction to the bootstrap*. Chapman Hall, N. Y.
- HALL, P. and WILSON, S. (1991): 2 guidelines for bootstrap hypothesis testing. *Biometrics*, 47, 757-762.
- HOLT, D., SCOTT A.J. and EWINGS P.D. (1980): Chi-squared tests with survey data. *J. Royal Stat. Soc. A*, 143, 303-320.
- KOVAR, J.G., RAO J.N.K. and WU C.F.J. (1988): Bootstrap and other methods to measure errors in survey estimates. *Can. J. of Stat.*, 16 (supplement), 25-45.
- RAO, J.N.K. and SCOTT A.J. (1984): On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *Annals of Statistics*, 12, 46-60.
- RAO, J.N.K. and Wu, C.F.J. (1988): Resampling inference with complex survey data. *JASA*, 83, 231-241.
- SCAGNI, A. (1999): Test di adattamento mediante tecniche bootstrap. *Proceedings of the "SIS '99" meeting*, Udine, June 7th-9th, in print.
- SITTER, R.R. (1992a): A resampling procedure for complex survey data. *JASA*, 87, 755-765.

Part V

Symbolic Data Analysis

Regression Analysis for Interval-Valued Data

L. Billard¹ and E. Diday²

¹ Department Statistics, University of Georgia,
Athens, GA 30602 USA
(e-mail: lynne@stat.uga.edu)

² CEREMADE, Universite de Paris 9 Dauphine
75775 PARIS Cedex 16 France
(e-mail: diday@ceremade.dauphine.fr)

Abstract. When observations in large data sets are aggregated into smaller more manageable data sizes, the resulting classifications of observations invariably involve symbolic data. In this paper, covariance and correlation functions are introduced for interval-valued symbolic data. These and their associated terms are then used to fit linear regression models to such data. The methods are illustrated with an example from cardiology.

1 Introduction

Today, we are often faced with the need to make meaningful statistical analyses (such as principal components, discriminant analysis, regression, etc.) on huge data sets whose very size makes the standard form of analysis very difficult to implement. These difficulties may typically exist from a computational viewpoint only even if theoretically there may be no mathematical limitations to the implementation of the relevant statistical analysis. Therefore, before embarking upon a specific analysis, it becomes necessary to classify or to reorganize the data into summary-type classifications or classes, where the number of classes is very much smaller than the number of single individuals in the original data set.

For example, suppose an investigation involves medical records with variables such as pulse rate, blood pressure, disease, etc., as well as identifying variables such as place of residence (Paris, Lyon, London, Brussels,...), age, gender, occupation, etc., for a very large number of individuals. Then, one particular aggregation of the data is to classify individuals by residence (or, by age-gender, occupation, etc.). In another situation, in which data are available for several cities (or regions, or etc.) but already classified by occupation, it may be desired to merge and/or to compare the data for each city whilst still retaining the identifying classification of "occupation". In a different direction, it may be of interest to describe and analyse underlying concepts such as illnesses, species, unemployment, etc. It may be desired to study the issue of whether or not the data contain individuals who might be classified

according to some preassigned concept of what the analyst might be seeking. Thus, for example, the starting point could be a query relating to the presence or otherwise of certain diseases.

In these (and similarly related) examples, the resulting data set, after the classification process has been implemented, will almost invariably contain symbolic data rather than classical data values, at least on some (but more probably on all) of the variables describing each observation in the classified data set. Indeed, symbolic data methods may have been an integral part of the classification procedure itself. By symbolic data, we mean that rather than a specific x_j value, an observed value for Y_j can be multi-valued, e.g., $\xi_j = \{1, 4, 7\}$ or $\{\text{blue, green}\}$, it may be interval-valued, e.g., $\xi_j = [10, 20]$ or $\xi_j = (\geq 0)$; or it may be modal-valued, e.g., $\xi_j = \{1 \text{ with probability .1, } 0 \text{ with probability .9}\}$, and so on. In addition, there may be rules that have to hold for data integrity, such as the need to maintain underlying information or background knowledge, etc. For example, levels and frequency of cancer (say) treatments must satisfy rules governing the presence of cancer. For a detailed description of symbolic data, see, e.g., Bock and Diday (2000).

Let us suppose we have $(p + 1)$ variables Y and X_j , $j = 1, \dots, p$, with Y being a dependent variable and $\{X_j, j = 1, \dots, p\}$ being p independent predictor variables, related to Y according to the relation $Y = f(X_1, \dots, X_p)$. In particular, let us focus attention on a standard linear regression relationship

$$Y = \mathbf{X}'\boldsymbol{\beta} + e \quad (1)$$

where $\mathbf{X}' = (1, X_1, \dots, X_p)$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ and $e \sim (0, \sigma^2)$. When the variables X_j take specific values x_j , $j = 0, \dots, p$, as in classical data, then the appropriate regression analysis for (1) (and indeed for $Y = f(X_1, \dots, X_p)$ in general) has been very well studied. In this paper, we want to fit regression models of the type (1) to so-called symbolic data. We study herein interval-valued symbolic data; other forms of symbolic data can be handled with appropriate adjustments. This is presented in Section 3 with an example in Section 4. First in Section 2, we look at covariance and correlation functions for interval-valued data.

2 Covariance and correlation functions

Bertrand and Goupil (1999) developed formulae for calculating the univariate empirical frequency distribution, the relative frequency distribution (or equivalently the frequency histogram and hence the empirical distribution function), along with the symbolic empirical mean and variance for interval-valued symbolic data which must also satisfy given logical dependency rules $\{\nu\}$. We extend those ideas to obtain basic descriptive statistics for the two-dimensional variable (Y_1, Y_2) , say.

Suppose $u \in E$ is the set of m symbolic objects with observations $Y(u) = \{Y_1(u), Y_2(u)\}$, $u = 1, \dots, m$. Suppose $Y(u)$ takes specific values on the

rectangle $Z(u) = Y_1(u) \times Y_2(u) = \{\xi_1^u, \xi_2^u\} = ([a_{1u}, b_{1u}], [a_{2u}, b_{2u}])$. Analogously to the univariate case, we assume the individual description vectors $x \in vir(d_u)$ are each uniformly distributed over the rectangle $Z(u)$ where $vir(d_u)$ is the virtual description of x defined as the set of all individual descriptions vectors x which satisfy the set of rules $\{\nu\}$; see Bertrand and Goupil (1999) and Billard and Diday (2000). Then, we can define the empirical joint density function for (Y_1, Y_2) as

$$f(\xi_1, \xi_2) = \frac{1}{m} \sum_{u \in E} \frac{I_u(\xi_1, \xi_2)}{\|Z(u)\|} \quad (2)$$

where $I_u(\xi_1, \xi_2)$ is the indicator function that (ξ_1, ξ_2) is or is not in the rectangle $Z(u)$ and where $\|Z(u)\|$ is the area of this rectangle. Note also that the summation in (2) is only over values for which the logical rules hold.

The symbolic empirical covariance between Y_1 and Y_2 is derived according to

$$\begin{aligned} Cov(Y_1, Y_2) &\equiv S_{12} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\xi_1 - \bar{Y}_1)(\xi_2 - \bar{Y}_2) f(\xi_1, \xi_2) d\xi_1 d\xi_2, \end{aligned} \quad (3)$$

and, substituting from (2) and recalling that $\xi_1 = [a_{1u}, b_{1u}]$ and $\xi_2 = [a_{2u}, b_{2u}]$, we have

$$\begin{aligned} Cov(Y_1, Y_2) &= \frac{1}{m} \sum_{u \in E} \frac{1}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} \int_{a_{1u}}^{b_{1u}} \int_{a_{2u}}^{b_{2u}} \delta_1 \delta_2 d\delta_1 d\delta_2 - \bar{Y}_1 \bar{Y}_2 \\ &= \frac{1}{m} \sum_{u \in E} \frac{1}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} \int_{a_{1u}}^{b_{1u}} \delta_1 d\delta_1 \int_{a_{2u}}^{b_{2u}} \delta_2 d\delta_2 - \bar{Y}_1 \bar{Y}_2 \\ &= \frac{1}{m} \sum_{u \in E} \frac{1}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} \left[\frac{1}{2} \delta_1 \right]_{a_{1u}}^{b_{1u}} \left[\frac{1}{2} \delta_2 \right]_{a_{2u}}^{b_{2u}} - \bar{Y}_1 \bar{Y}_2 \\ &= \frac{1}{4m} \sum_{u \in E} \frac{(b_{1u}^2 - a_{1u}^2)(b_{2u}^2 - a_{2u}^2)}{(b_{1u} - a_{1u})(b_{2u} - a_{2u})} - \bar{Y}_1 \bar{Y}_2. \end{aligned}$$

Hence, the empirical symbolic covariance function is

$$Cov(Y_1, Y_2) = \frac{1}{4m} \sum_{u \in E} (b_{1u} + a_{1u})(b_{2u} + a_{2u}) - \frac{1}{4m^2} \left[\sum_{u \in E} (b_{1u} + a_{1u}) \right] \left[\sum_{u \in E} (b_{2u} + a_{2u}) \right] \quad (4)$$

where the symbolic empirical mean of Y , is, from Bertrand and Goupil (1999),

$$\bar{Y} = \frac{1}{2m} \sum_{u \in E} (b_u + a_u). \quad (5)$$

The symbolic empirical variance of Y is

$$S^2 = \frac{1}{4m} \sum_{u \in E} (b_u + a_u)^2 - \frac{1}{4m^2} \left[\sum_{u \in E} (b_u + a_u) \right]^2. \quad (6)$$

Hence, we can define the symbolic empirical correlation function between two variables Y_1 and Y_2 , denoted by $r(Y_1, Y_2)$, as

$$r(Y_1, Y_2) = S_{12}/\sqrt{S_1^2 S_2^2}. \quad (7)$$

3 Linear regression model

Let us take the multiple linear regression model (1). We know that the least squares estimator of the regression coefficient β is, for data $\mathbf{Y} = (Y_1, \dots, Y_m)$,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} (\mathbf{X}' \mathbf{Y}). \quad (8)$$

Suppose for simplicity we take $p = 1$. Then, from standard classical theory we have:

$$\hat{\beta}_1 = \frac{Cov(Y, X)}{S_X^2} = r(Y, X)(S_Y/S_X), \quad (9)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}. \quad (10)$$

When the data are symbolic data we can use the same ideas but with the sample means, variances and correlation function terms in (9) and (10) being replaced by their symbolic counterparts. Likewise, we can calculate $\hat{\boldsymbol{\beta}}$ for general p using equation (8).

Notice that since the data are interval-valued and since it is assumed that possible values are uniformly distributed across these intervals, then the formulae for the symbolic means, variances, covariance and correlation functions, and hence the regression line fits, correspond to the same values obtained by applying classical methods to the midpoints of the intervals. This would not be the case were the intervals not uniformly distributed or were the data modal or categorical values more generally.

4 Example

Suppose we have the record of the pulse rate Y and the systolic blood pressure X_1 for each of eleven patients (taken from Raju (1997)) and shown in Table 1. Let us suppose there is a logical rule that says the diastolic blood pressure (X_2) must be less than the systolic blood pressure, i.e., $X_2 \leq X_1$. Then, the $u = 7$ observation contradicts this rule, i.e., $|vir(d_u)| = 0$, for $u = 7$. Therefore, in the summation in (2), and hence also in subsequent calculations, terms corresponding to this observation are omitted. Let us find the regression equation

$$Y = \beta_0 + \beta_1 X_1. \quad (11)$$

Table 1 - Data

u	Y Pulse Rate	X_1 Systolic Pressure	X_2 Diastolic Pressure
1	44, 68	90, 100	50, 70
2	60, 72	90, 130	70, 90
3	56, 90	140, 180	90, 100
4	70, 112	110, 142	80, 108
5	54, 72	90, 100	50, 70
6	70, 100	130, 160	80, 110
7	63, 75	60, 100	140, 150
8	72, 100	130, 160	76, 90
9	76, 98	110, 190	70, 110
10	86, 96	138, 180	90, 110
11	86, 100	110, 150	78, 100

First, we calculate the symbolic statistics: $\bar{Y} = 79.1$, $\bar{X}_1 = 131.3$, $S_Y^2 = 162.29$, $S_{X_1}^2 = 495.41$, $Cov(Y, X_1) = 194.170$, and $r(Y, X_1) = 0.685$. Hence, we can calculate $\hat{\beta}_1 = 0.392$, and $\hat{\beta}_0 = 27.639$. Therefore, the regression equation (11) becomes

$$\text{Pulse Rate} = 27.639 + (.392) \text{ Systolic Pressure}. \quad (12)$$

Suppose now we wanted to predict the pulse rate when the systolic pressure is in the interval (118, 126), say. From (12), we have $Y_{118} = 27.639 + (0.392)(118) = 73.887$ and $Y_{126} = 27.639 + (0.392)(126) = 77.023$. I.e., when the systolic blood pressure is in the interval (118, 126), the predicted pulse rate would be $\xi = (73.89, 77.02)$.

In contrast, had we fitted the regression line (11) through the lower points $\{a_{ju}\}$ only, we would obtain the equation

$$\text{Pulse Rate } (a) = 29.664 + (0.330) \text{ Systolic Pressure } (a); \quad (13)$$

and likewise by fitting the model (11) through the upper points $\{b_{ju}\}$ only, we obtain the relation

$$\text{Pulse rate } (b) = 45.070 + (0.308) \text{ Systolic Pressure } (b). \quad (14)$$

Were we to use these relationships to obtain the predicted pulse rate when the systolic pressure is (118, 126), we would have the predictions, respectively, $\text{Pulse Rate } (a) = (68.60, 71.24)$, and $\text{Pulse Rate } (b) = (81.41, 83.88)$. The two regression equations (13) and (14) are standard regression fits as are the corresponding predictions for the systolic blood pressure in the interval (118, 126). If we only had these regression fits, it is not unreasonable that we might place greater confidence in the lower level prediction calculation of 68.60 [from pulse

rate (a)], and in the higher level prediction calculation of 83.88 [from pulse rate (b)], giving a prediction interval of (68.60, 83.88). When this is compared with the symbolic prediction interval of (73.89, 77.02), obtained from the symbolic regression equation (12), it is clear that the symbolic analysis gives a tighter fit.

Finally, let us fit both systolic blood pressure (X_1) and diastolic blood pressure (X_2) as predictor variables for pulse rate (Y) i.e., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. Then, we can show that, for the data of Table 1,

$$X'X = \begin{pmatrix} 10 & 1313 & 846 \\ 1313 & 177351 & 113659 \\ 846 & 113659 & 73396 \end{pmatrix}, \quad X'Y = \begin{pmatrix} 791 \\ 105800 \\ 68329 \end{pmatrix}$$

where the entries in each matrix are the symbolic counterparts of the symbolic sums and crossproducts as given in Section 2. Hence, we have $\hat{\beta}_0 = 20.703$, $\hat{\beta}_1 = -0.040$, and $\hat{\beta}_2 = 0.805$. Therefore, the regression equation becomes

$$\text{Pulse Rate} = 20.703 - (0.040) \text{ Syst. Press.} + (0.805) \text{ Diast. Press.} \quad (15)$$

Prediction and other uses of the regression equation (15) then follow in the usual ways.

5 Conclusion

Regression models for other types of symbolic data can be fitted analogously with appropriate adjustments. Thus, for example, if the data were modal, then each rectangular region $Z(u)$ (or, more generally, a hypercube) would have weights corresponding to the probabilities of the modal data lying in that rectangle. Then, the empirical frequency joint density function $f(\xi_1, \xi_2)$ in (3) would take the necessary form reflecting these weights or probabilities (instead of the uniform rectangular form used above). The subsequent results therefore lead to weighted regression analogues. Logistic regression and generalised linear regression models more broadly defined can likewise be handled.

References

- BERTRAND, P. and GOUPIL, F. (1999). Descriptive Statistics for Symbolic Data. In: H.-H. Bock and E. Diday (Eds). *Symbolic Official Data Analysis*. Springer, 103-124.
- BILLARD, L. and DIDAY, E. (2000). From the Statistics of Data to a Statistics of Knowledge: Symbolic Data Analysis. In preparation.
- BOCK, H. -H. and DIDAY, E. (eds.) (2000). *Symbolic Official Data Analysis*. Springer.
- RAJU, S. R. K. (1997). Symbolic Data Analysis in Cardiology. In: E. Diday and K. C. Gowda (Eds). *Symbolic Data Analysis and Its Applications*. CEREMADE, Paris, 245-249.

Symbolic Approach to Classify Large Data Sets

Francisco de A.T. de Carvalho, Cezar A. de F. Anselmo, and
Renata M.C.R. de Souza

Centro de Informatica - CIN / UFPE,
Av. Prof. Luiz Freire, s/n - Cidade Universitaria,
CEP: 50.740-540 Recife-PE BRASIL
(e-mail: {fatec,rmc,cafa}@di.ufpe.br)

Abstract. The aim of this work is to present an approach to classify large data sets based on a Boolean symbolic classifier as described in Yaguchi et al (1996) and Ichino et al (1996). Compared with this last classifier, our system keeps the concept of *mutual neighbours* between examples (Ichino et al (1996)) but introduces some modifications in both learning step (generalisation tools) and allocation step (matching functions). As an example of large data set processing, a particular kind of simulated images will be classified according to this approach.

1 Introduction

With the recent advances in information technologies in all domains of human activities huge sets of data are now stored in large databases. Nowadays, some approaches have been proposed in order to discovery regularities, to extract knowledge and to summarise the informations stored on these large data sets. Our approach is the so-called *Symbolic Data Analysis (SDA)* (see Diday (1988)). Its first step concerns knowledge extraction from large data bases as in "Data Mining". These knowledge are described by more complex data called "symbolic data" as they contain internal variation and are structured.

The second step in SDA is to apply new tools in such extracted knowledge in order to extend "Data Mining" in "Knowledge Mining" (Diday (1998)). Therefore, an extension of exploratory data analysis and statistical methods to the symbolic data is needed.

This work concerns the classification of large data set by means of the symbolic approach. Our reference here is the work from Ichino et al (1996) where they introduce a Boolean symbolic classifier as a region oriented approach. The main idea of this approach is the adaptation of the concept of *mutual neighbours*, introduced by Gowda et al (1978), to define the concepts of mutual neighbours between Boolean symbolic objects (hereafter denoted SO's) and *Mutual Neighbourhood Graph (MNG)* between groups.

In this approach, the training set may be usual data or Boolean symbolic data. In the learning step, each group is described by a Boolean SO or by a disjunction of Boolean SO's which are obtained using as tools of generalisation an approximation of the *MNG* and a Boolean symbolic operator (join).

At the end of this step, a complete and discriminant description of the examples of each group is furnished by the classifier. The classification rule is based on a matching function between an usual description of an observation and a Boolean SO which describes a group.

In this work, we modify this approach at several levels (generalisation tools, *MNG* approximation, matching functions) and we use it to classify a special kind of simulated images (*Synthetic Aperture Radar (SAR)* images) showing areas with different degrees of roughness, assuming multiplicative model (Frery et al (1977)).

2 Basics on Symbolic Objects

Extracting knowledge means getting new concepts, that is why SO are introduced. They form a mathematical modelling of concepts. They are used as input and as explanatory output of a SDA (Diday (1998)).

Let Ω be a set of individuals and $\{D_1, \dots, D_p\}$ the domains of the variables y_j ($j = 1, \dots, p$) selected to describe the individuals as well as the Boolean SO's. These variables are of different types: nominal, ordinal, discrete or continuous, taking multi-values or intervals for a Boolean SO.

A Boolean SO s is expressed by the triple (a, R, d) where i) $d = (d_1, \dots, d_p)$, $d_j \subseteq D_j$ is the set of values assumed by the variables in its description; ii) $R = (R_1, \dots, R_p)$ is the set of relations (e.g.: \subseteq, \in, \dots); iii) a is a Boolean function which allows to compare $\omega \in \Omega$ with s in such a way that if $a(\omega) = \wedge_j [y_j(\omega) \in R_j \text{ and } d_j] \in \{\text{true}\}$ this means that ω respects all the properties in d by means of the relations in R . The extent of a SO s in Ω is defined as $\text{Ext}(s/\Omega) = \{\omega \in \Omega / a(\omega) = \text{true}\}$.

Example. A segment (set of pixels) described as usual data by the grey level mean (y_1) and standard deviation (y_2) may be represented by a Boolean SO as $seg = [y_1 \in [44.50, 44.50]] \wedge [y_2 \in [7.50, 7.50]]$.

Moreover, a SO may be described also by *modal* variables, when it is associated to the categories or intervals, a *mode*, i.e., a measure of probability, possibility, or belief. The SO is then termed *modal*.

3 Symbolic classifier

We start this section by reminding some definitions which will be necessary to describe the modifications which we have introduced in the learning step (generalisation tools and Mutual neighbourhood Graph approximations) and in the allocation step (matching functions) of the classifier.

3.1 Boolean symbolic operations

Let $a = [y_1 \in \bar{d}_1] \wedge \dots \wedge [y_p \in \bar{d}_p]$ and $b = [y_1 \in d_1] \wedge \dots \wedge [y_p \in d_p]$ be two Boolean SO's. The following operations may be defined (Diday (1998), Ichino et al (1996)):

- a) The join between a and b is defined as $a \oplus b = [y_1 \in \bar{d}_1 \oplus \underline{d}_1] \wedge \dots \wedge [y_p \in \bar{d}_p \oplus \underline{d}_p]$, where i) if y_i is quantitative or ordinal qualitative, $\bar{d}_i \oplus \underline{d}_i = [\min(\bar{l}_i, \underline{l}_i), \max(\bar{u}_i, \underline{u}_i)]$, and $(\bar{l}_i, \underline{l}_i)$ and $(\bar{u}_i, \underline{u}_i)$ are the lower bound and upper bound of the intervals \bar{d}_i and \underline{d}_i , respectively; min and max furnishes, respectively, the minimum and the maximum between two values; ii) if y_i is nominal qualitative, $\bar{d}_i \oplus \underline{d}_i = \bar{d}_i \cup \underline{d}_i$.
- b) The disjunction between a and b is defined as $a \vee b = \{[y_1 \in \bar{d}_1] \wedge \dots \wedge [y_p \in \bar{d}_p]\} \vee \{[y_1 \in \underline{d}_1] \wedge \dots \wedge [y_p \in \underline{d}_p]\}$.
- c) The conjunction between a and b is defined as $a \wedge b = [y_1 \in \bar{d}_1 \cap \underline{d}_1] \wedge \dots \wedge [y_p \in \bar{d}_p \cap \underline{d}_p]$.

3.2 Mutual neighbourhood relation

Let $C = \{C_i\}$ a set of m groups of SO's, where $C_i = \{a_{i1}, \dots, a_{in_i}\}$, $a_{ij} = \wedge_{k=1}^p [y_k \in d_{ijk}]$, $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n_i\}$. Assume $a_{ip}, a_{iq} \in C_i$. The *Mutual Neighbourhood Relation (MNR)* R_v (Ichino et al (1996)) is defined saying that $a_{ip} R_v a_{iq}$ is true iff $\forall a_{jl} \in \bar{C}_i = \bigcup_{j \neq i}^m C_j, \exists k \in \{1, \dots, p\}$ such that $d_{jlk} \cap (d_{ipk} \oplus d_{iqk}) = \emptyset$.

3.3 Mutual neighbourhood graph

The *MNG* (Ichino et al (1996)) from C_i versus \bar{C}_i , denoted as $MNG(C_i/\bar{C}_i)$, is a graph where the nodes are SO's from C_i and the edges are distinct couples of SO's from C_i satisfying the *MNR*, i. e., $MNG(C_i/\bar{C}_i) = (N, E)$, where $N = C_i$ and $E = \{(a_{ip}, a_{iq}) \in C_i \times C_i / p \neq q \text{ and } a_{ip} R_v a_{iq} \text{ is true}\}$.

3.4 Learning step

In theory, the approach proposed by Yaguchi et al (1996) works in the following way: i) if the *MNG* is a complete graph, the group will be described by a Boolean SO obtained by the join of its members; ii) otherwise the group will be described by a disjunction of Boolean SO, each Boolean SO being obtained by the join of the members of this group which forms a clique.

Unfortunately, finding all the cliques on a graph is a NP-complete problem and in fact in Yaguchi et al (1996) is used the following approximation of the *MNG*: i) choose the element of the group which has the maximum number of mutual neighbours as the seed; ii) find the element of the group which has the maximum number of common mutual neighbours with the seed. If this element and the seed are also mutual neighbour, then to update the seed by the join between it and this element; iii) repeat step ii) until there is no more elements on the group which are mutual neighbours of the actual seed. At this point, a clique has been constructed with the actual elements of the group; iv) delete all elements of the group which have been joined in the steps i) to iii); v) repeat steps i) to iv) until there is no more elements in the group.

The complexity of this step is on $O(n^3)$, n being the number of elements in the group, it is why, in order to improve the run time of our system in processing large data sets, we have implemented the following modifications in the last algorithm: a) at the step i) the seed is now searched according to the lexicographic order presented by the elements of the group; and b) at the step ii) the search of the elements of the group which are mutual neighbours of the seed is also doing considering theirs lexicographic order. With these last modifications, the complexity is now on $O(n^2)$.

Moreover, in the original algorithm, both the seed and the element of the group which may be joined with it to form a new seed are selected only if they have never been selected before. In our system, we maintain this constrain concerning the seed but we leave it concerning the other elements of the group in two different manner: a) either an element which is in the border of the hypercube (Boolean SO in the Cartesian space) which forms a clique may always be joined with the actual seed; b) either all the elements of the group may always be joined with the actual seed.

Another important modification on this algorithm concerns the generalisation tools. In the original algorithm, a clique is described by a hypercube where the edges are obtained by the join of its elements. In our system, a clique is always described by an hypercube but its edges are obtained by the mean (or median) and standard deviation confidence interval at a level α of its elements. Of course, this last modification is possible only in the case where the input data are usual data.

3.5 Allocation step

The allocation of a new observation to a group is based on matching functions between a standard description of an individual and a Boolean SO which describes a group. Remember that an individual, which is described in a standard way, may also be described as a Boolean SO.

Let $a = \wedge_{j=1}^p [y_j \in d_j^a]$ a Boolean SO which describes a new observation to be classified and let $r_i = \vee_v r_{iv}$, $r_{iv} = \wedge_{j=1}^p [y_j \in d_{ivj}^r]$, be a disjunction of Boolean SO's which describes the group C_i , $i \in \{1, \dots, m\}$.

In the algorithm of Ichino et al (1996), it is used the following affectation rule: a new observation a is affected to group C_i , described by r_i , if $f(a, r_i) \geq f(a, r_h), \forall h \in \{1, \dots, m\}$, where $f(a, r_h) = \max_v f(a, r_{hv})$ and

$$f(a, r_{hv}) = \frac{1}{p} \sum_{j=1}^p \frac{\mu(d_{hvj}^r)}{\mu(d_{hvj}^r \oplus d_j^a)} \text{ with } \mu(d) = \begin{cases} \text{cardinal}(d), & \text{if } d \text{ is a finite set} \\ \text{range}(d), & \text{if } d \text{ is an interval} \end{cases}$$

Concerning this kind of affectation rule, we have used, as alternative, this other similar criterion:

$$t(a, r_{hv}) = \frac{\pi(r_{hv})}{\pi(r_{hv} \oplus a)}, \text{ where } \pi(s) = \prod_{j=1}^p \mu(d_j^s) \text{ and } s \text{ is a SO}$$

Notice that f and t are non symmetric similarities functions which measures the matching between a usual description of an observation and the Boolean symbolic description of a group.

We have used also, as alternative, dissimilarity matching functions. In that case, the affectation rule becomes: an observation represented by a is affected to group C_i , described by r_i , if $d_\lambda(a, r_i) \leq d_\lambda(a, r_h), \forall h \in \{1, \dots, m\}$, where $d_\lambda(a, r_h) = \min_v d_\lambda(a, r_{hv})$ and (De Carvalho et al (1999))

$$d_\lambda(a, r_{hv}) = \frac{1}{p} \left[\sum_{j=1}^p \{\psi(a_j, r_{hvj})\}^\lambda \right]^{\frac{1}{\lambda}}, k, r \in \{1, 2, \dots\}, \text{ with}$$

$$\psi(a_j, r_{hvj}) = \left\{ \frac{1}{2} \left[\left(\frac{\phi_{1\gamma}(a_j, r_{hvj})}{\mu(d_j^a \oplus d_{hvj}^r)} \right)^k + \left(\frac{\phi_{2\gamma}(a_j, r_{hvj})}{\mu(d_j^a \oplus d_{hvj}^r)} \right)^k \right] \right\}^{\frac{1}{k}}, \text{ and}$$

$$\phi_{1\gamma}(a_j, r_{hvj}) = (1 - 2\gamma)\mu(d_j^a \cap \overline{d_{hvj}^r}) + \mu(\overline{d_j^a} \cap d_{hvj}^r) + \mu(\overline{d_j^a} \cap \overline{d_{hvj}^r} \cap (d_j^a \oplus d_{hvj}^r))$$

$$\phi_{2\gamma}(a_j, r_{hvj}) = \mu(d_j^a \cap \overline{d_{hvj}^r}) + (1 - 2\gamma)\mu(\overline{d_j^a} \cap d_{hvj}^r) + \mu(\overline{d_j^a} \cap \overline{d_{hvj}^r} \cap (d_j^a \oplus d_{hvj}^r))$$

Considering this kind of affectation rule, we have used also this other similar criterion:

$$u(a, r_{hv}) = \sum_{j=1}^p \frac{\mu(a \oplus r_{hvj}) - \mu(r_{hvj})}{\mu(r_{hvj})}$$

Notice that d_r is a symmetric and u is a non symmetric dissimilarity function which measures the matching between a usual description of an observation and the Boolean symbolic description of a group.

4 The Monte Carlo experiences

As an example of large data set processing, a special kind of simulated images (*SAR* images) will be classified according to the approach described on the previous section. The multiplicative model has been widely and successfully used in the modelling, processing and analysis of *SAR* images (Frery et al (1997)). This model assumes that the observed return Z is a random variable defined as the product between the random variables X (the terrain backscatter) and Y (the speckle noise).

Different types of region and kinds of detection (intensity or amplitude format) can be associated to different distributions for the return. Regarding to region types, the homogeneous (agricultural fields, bare soil, etc.), heterogeneous (primary forest, etc.), and the extremely heterogeneous (urban areas, etc.) are considered in this work. A common hypothesis is that the return in amplitude case follows the Square Root of Gamma, the K -Amplitude and the $G0$ -Amplitude distribution in homogeneous, heterogeneous, and extremely heterogeneous areas, respectively (Frery et al (1997)).

The process for obtaining simulated images consists of creating an idealised classes image (a phantom), and to associate each class to a particular distribution. The Monte Carlo experience was performed with images of size 256×256 and two situations ranging from *moderate* to *great difficulty* of classification. 100 replications were obtained with identical statistical properties and the error rate of the classifications were recorded. The results were, in mean, around 3 % and 23 % for each considered situation, matching function, generalisation tool and *MNG* approximation.

5 Final comments and conclusions

We have presented a modified version of the Boolean symbolic classifier proposed by Yaguchi et al (1996). The usefulness of this classifier is illustrated by the satisfactory error rate under situations ranging from *moderate* case to *great difficulty* of classification in the framework of a Monte Carlo experience of simulated images. Another interesting problem concerning this approach is to extend it to modal symbolic data.

Acknowledgements. This paper is supported by grants from CNPq-Brasil Proc. 301387/92-3

References

- DE CARVALHO, F. A. T. and SOUZA, R. M. C. (1999): New metrics for constrained Boolean symbolic objects. In: *Studies and Research: Proceedings of the Conference on Knowledge Extraction and Symbolic Data Analysis (KESDA '98)*, Office for Official Publications of the European Communities, Luxembourg, 175–187.
- DIDAY, E. (1998): Symbolic Data Analysis: a Mathematical Framework and Tool for Data Mining. In: A. Rizzi, M. Vichi and H.-H. Bock. (Eds.): *Advances in Data Science and Classification*, Springer-Verlag, heidelberg, 409–416.
- FRERY, A.C., MUELLER, H. J., YANASSE, C.C.F. and SANT'ANNA, S. J. S. (1997): A model for extremely heterogeneous clutter. *IEEE Transactions on Geoscience and Remote Sensing*, 1, 648–659.
- GOWDA, K. C. and KRISHNA, G. (1978): Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognition*, 10, 105–112.
- ICHINO, M., YAGUCHI, H. and DIDAY, E. (1996): A fuzzy symbolic pattern classifier. In: E. Diday, Y, Lechevallier and O. Opitz (Eds.): *Ordinal and Symbolic Data analysis*. Springer, Heidelberg, 92–102.
- YAGUCHI, H., ICHINO, M. and DIDAY, E. (1996): A knowledge acquisition system based on the Cartesian space model. In: E. Diday, Y, Lechevallier and O. Opitz (Eds.): *Ordinal and Symbolic Data analysis*. Springer, Heidelberg, 113–122.

Factorial Methods with Cohesion Constraints on Symbolic Objects

N.C. Lauro¹, R. Verde², and F. Palumbo³

¹ Dip. di Matematica e Statistica, Università di Napoli Federico II. Monte S. Angelo, I-80126 Napoli, Italia (e-mail: claudio@unina.it)

² Facoltà di Economia, SUN - Seconda Università di Napoli, P.zza Umberto I, I-81043 Capua, Italia (e-mail: verde@dms.unina.it)

³ Dip. di Istituzioni Economiche e Finanziarie, Università di Macerata Via Crescimbeni, 14, Macerata, Italia (e-mail: palumbo@unimc.it)

Abstract. In this paper we generalize some results of Factorial Analysis to the complex data structure defined in Symbolic Data Analysis. The proposed treatments are based on a multi-steps *symbolic-numerical-symbolic* procedure and on the geometric results interpretation. The paper generalizes the constrained Factorial Approach (Lauro and Palumbo, 2000), that permits to take into account the Symbolic Data Structure in the analysis of the coded data.

1 Introduction

Real world phenomena are often characterized by complex information, expressed by multi-valued variables and their relationships. A statistical treatment of complex structured data (Diday, 1987) has been recently considered in the context of Symbolic Data Analysis (Bock and Diday, 1999). In this framework, different authors proposed some approaches extending factorial data analysis techniques to the study of the relationships between Symbolic Objects (SO's) in a reduced sub-space. Cazes et al. (1997) and Chouakria et al. (1998) developed the *Vertices Principal Component Analysis* (V-PCA) as an extension of Principal Component Analysis to SO's described by *interval valued* variables. Specifically, V-PCA, based on a geometric representation of the SO's as hypercubes, is a classical PCA on a suitable numerical coding of their vertices which are considered as statistical units. Nevertheless, in most cases, the SO's are described by different kinds of descriptors: nominal, multi-nominal, modal, *interval valued* variables. Under these assumptions, Verde (1997) proposed to homogenize the descriptors by a fuzzy and/or crisp coding system, and perform the Generalized Canonical Analysis.

The present paper offers a review of the recent proposals in Factorial Symbolic Analysis and, in addition, it points out the weak points affecting these approaches and proposes some new enhancements.

In detail, Section 2 summarizes the basic notations and definition of Symbolic Data Analysis (SDA); Section 3 treats three different factorial approaches for Symbolic Data: Principal Component Analysis, Generalized Canonical Analysis and Factorial Discriminant Analysis.

2 Symbolic objects definition and coding

A SO s is defined (Bock and Diday, 1999) as a triple (a, R, d) , where $d = (d_1, \dots, d_p)$ is the set of description values assumed by s , with respect to p descriptors (y_1, \dots, y_p) with domains in $D = (D_1, \dots, D_p)$ respectively; a is a mapping function; $R = (R_1, \dots, R_p)$ is a set of relationships (e.g. $R\{=, \in, \leq, \dots\}$). The SO descriptors can be nominal, numerical-discrete and continuous (*interval valued*). Nominal variables are called *modal variables* if a *mode* (i.e. a measure of belief, possibility or probability) can be associated to the categories.

Each SO is described by a conjunction of its descriptors y_j assuming values in d_j ($j = 1, \dots, p$). Moreover, given a set of n individuals or SO's $\omega_i \in \Omega$, the extent of a SO s can be computed through the *mapping function* $a : \Omega \rightarrow \{0, 1\}$. The latter compares the characteristics of the elements ω 's to the descriptions d of a SO i , by a relationship of $R = (R_1, \dots, R_p)$. The comparison function $a(\omega)$ takes value 1 when: $y_j(\omega) \in d_j, \forall j = 1, \dots, p$.

The Symbolic Data Analysis strategies, we refer to in the following, share the following *symbolic-numerical-symbolic* procedures steps: *i*) SO descriptors coding (or transformation); *ii*) numerical treatment of the coded descriptors; *iii*) results interpretation according to the original symbolic data.

In accordance with the descriptors nature, different transformations or codings of SO's are proposed. In the case of p *continuous interval valued descriptors*, each SO i is associated to a matrix \mathbf{Z}_i , having 2^p rows and p columns.

The matrix \mathbf{Z}_i is made of all possible combinations of the lower bounds $y_{i,j}$ and upper bounds $\bar{y}_{i,j}$ of the intervals of values assumed by the SO i for p numerical descriptors y_j ($j = 1, \dots, p$). The stack matrix \mathbf{Z} ($N \times p$), of the n coded SO's, is obtained by superposing the n matrices \mathbf{Z}_i , with $N = n2^p$. In a geometric view, the rows of each matrix \mathbf{Z}_i (for $i = 1, \dots, n$) correspond to the vertices of the hypercube associated to each i -th SO.

In the case of p nominal descriptors, each SO i is associated to a boolean matrix \mathbf{Z}_i partitioned in p blocks. The generic block \mathbf{Z}_{ij} ($j = 1, \dots, p$) has r_i rows and k_j columns. Particularly, $r_i = \prod_{j=1}^p k_{ij}$, which is obtained by the product of the number of categories k_{ij} , where $k_{ij} \subseteq D_j$.

The matrix $Z_{N \times K}$ of the n SO's, with $N = \sum_{i=1}^n r_i$ rows and $K = \sum_{j=1}^p k_j$ columns, is obtained by superposing the n matrices \mathbf{Z}_i (with $1 \leq i \leq n$). In the case of p *modal descriptors*, each nominal descriptor with associated relative frequency, belief or probability (i.e. *modes*), is coded in a single row, whose elements are the K modes. Therefore, \mathbf{Z} is a column-partitioned matrix with n rows and p blocks matrices \mathbf{Z}_j , each block having k_j columns.

In general, when dealing with different kinds of descriptors, a suitable transformation is required before treatment to homogenize their coding. In this respect, the numerical *interval valued* variables are first categorized and then transformed according to the *fuzzy* approach, using nonlinear functions

(Basic splines) of low degree. Interval bounds are coded into different rows of the coding matrices \mathbf{Z}_{ij} . The coding matrix \mathbf{Z}_i of the SO i is a partitioned matrix in p blocks \mathbf{Z}_{ij} ($j = 1, \dots, p$), whose elements assume values in $[0,1]$. The number of columns depends on the number of categorized intervals in where the numerical descriptors have been split (or the number of knots of the B-spline functions) and 2^p rows, corresponding to every possible combination of the coded bounds of the numerical *interval valued* descriptors.

In this case, the *global coding* matrix $\mathbf{Z}_{N \times K}$ has N rows, corresponding to the number of vertices of the n SO's. Let us remark that N tends to be very large in the case of nominal descriptors or numerous continuous *interval valued* descriptors. This event can seriously affect the numerical treatment of the SO's. The total number of columns is equal to $K = \sum_{j=1}^p k_j$, where k_j is the categories number in D_j of the nominal descriptors or the number of categories where numerical descriptors have been categorized.

3 Factorial methods extended to SO's

3.1 Principal Component Analysis on SO's

The V-PCA on the vertices of the SO's, proposed by Cazes et al. (1998) consists of performing a classical PCA on the standardized matrix $\mathbf{Z}_{N \times p}$ of the p numerical *interval valued* descriptors. The principal axes \mathbf{v}_m ($1 \leq m \leq p$) are obtained by maximizing the sum of the squared coordinates of the projected vertices. As known, they are obtained as solutions of the characteristics equation in \mathbb{R}^N :

$$\frac{1}{N} \mathbf{Z}' \mathbf{Z} \mathbf{v}_m = \lambda_m \mathbf{v}_m \quad 1 \leq m \leq p, \quad (1)$$

under the usual orthonormality constrains: $\mathbf{v}_m' \mathbf{v}_m = 1$ and $\mathbf{v}_m' \mathbf{v}_{m'} = 0$ for $m \neq m'$, being: \mathbf{v}_m and λ_m the m -th eigenvector and the m -th eigenvalue respectively, associated to the matrix $\frac{1}{N} \mathbf{Z}' \mathbf{Z}$. The vertices coordinates of the i -th SO on the principal axes are given by the vector: $\psi_{i,m} = Z_i \mathbf{v}_m$. Then, the SO's are visualized on a factorial plane by the *maximum covering area rectangles* (MCAR), that have their sides parallel to the axes and cover all projected vertices belonging to the same SO (Cazes et al., 1997). Over-sized rectangles frequently occur, because in V-PCA the vertices are treated as independent without considering any relationship among vertices belonging to the same SO. So doing, V-PCA maximizes the sum of squared distances among vertices, rather than a sum of distances between SO's.

The SO-PCA proposed by Lauro and Palumbo (1998, 2000) is based on the search of the factorial axes which maximize the *SO's between variance* rather than the *total variance* of the vertices, like in V-PCA. Actually, the SO-PCA aims at representing the SO's on a factorial plane as best separated from each other. In this sense, it seems the most coherent extension of the

classical PCA which, as known, aims at maximizing the difference between statistical units. Therefore, to relate the vertices of the same hypercube, a *vertices cohesion constraint* is defined in the analysis. The latter is introduced by the indicator matrix \mathbf{A} ($N \times n$), which describes the belonging of the N vertices to the n SO's. The principal axes are obtained (Lauro and Palumbo, 1998), in the space \mathbb{R}^N , as solutions of the following characteristic equation:

$$\frac{1}{N} [\mathbf{Z}' \mathbf{A} (\mathbf{A}' \mathbf{A})^{-1} \mathbf{A}' \mathbf{Z}] \tilde{\mathbf{v}}_m = \tilde{\lambda}_m \tilde{\mathbf{v}}_m, \quad (2)$$

where $\tilde{\mathbf{v}}_m$ is defined under the usual orthonormality constraints, already considered in V-PCA. Because $\mathbf{P}_A = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ is an orthogonal projector, this approach is a particular case of the so called PCA on a reference subspace (D'Ambra, Lauro, 1982), here spanned by the columns of \mathbf{A} . The coordinates of the i -th SO vertices on the axis m , are given by the vector $\tilde{\psi}_{i,m} = Z_i \tilde{\mathbf{v}}_m$. Notice that the coordinates can be directly obtained by the analysis in \mathbb{R}^p based on the following eigenequation: $(\mathbf{A}'\mathbf{A})^{-\frac{1}{2}}(\mathbf{A}'\mathbf{Z}\mathbf{Z}'\mathbf{A})(\mathbf{A}'\mathbf{A})^{-\frac{1}{2}}\tilde{\mathbf{w}}_m = \tilde{\lambda}_m \tilde{\mathbf{w}}_m$. Then, analogously to V-PCA, the SO's are represented by the MCAR's.

The symbolic interpretation of principal axes is accomplished with reference to the variables z_j 's having maximal contributions. The measures of contributions of descriptors to the axes are expressed, as in the classical PCA, as the squared correlations between variables and factors (Lebart et al. 1995).

3.2 Generalized Canonical Analysis on SO's

An extension of the classical Generalised Canonical Analysis (GCA) has been proposed (Verde, 1997) to analyze SO's described by any kind of descriptors. The method is performed on the global coding matrix $\mathbf{Z}_{N \times K}$ that can also be considered as a juxtaposition of the p fuzzy and/or crisp coding matrices: $\mathbf{Z} = [\mathbf{Z}_1 | \dots | \mathbf{Z}_j | \dots | \mathbf{Z}_p]$. The GCA searches for the synthesis axes ν_j in each sub-space E_j spanned by the column vectors of the coding matrices \mathbf{Z}_j , and in addition, the orthogonal vectors ξ_m , as a global synthesis of every ν_j . The criterion optimized is the average multiple correlation ratio: $\sum_{j=1}^p (\xi_m | \nu_j)^2$ where ν_j and ξ_m are normalized to 1. The maximum inertia axes ξ_m are given by the following characteristic equation solution, under the usual orthonormality constraints $\xi'_m \xi_m = 1$ and $\xi'_m \xi_{m'} = 0$ (with $m \neq m'$):

$$\frac{1}{Np} \mathbf{Z} \Sigma^{-1} \mathbf{Z}' \xi_m = \mu_m \xi_m \quad (3)$$

(with $m = 1, \dots, M$; $M = \min(N - 1, K - p + 1)$), where ξ_m and μ_m are the m -th eigenvector and eigenvalue respectively of the decomposed matrix; Σ^{-1} is a block diagonal matrix of elements $(\mathbf{Z}'_j \mathbf{Z}_j)^{-1}$ ($j = 1, \dots, p$).

It can be easily demonstrated that the vector ξ_m is a linear combination of \mathbf{Z} : $\xi_m = \mathbf{Z} \beta_m$, where β_m ($m = 1, \dots, M$) are the solutions of the dual analysis in \mathbb{R}^K (i.e. the eigenvector of $\Sigma^{-1} \mathbf{Z}' \mathbf{Z}$). The coordinates of the vertices of the SO's on the factorial m -th axis are given by the row elements of ξ_m (alternatively $\mathbf{Z} \beta_m$).

Representing the coding categories by $(\mu_m)^{-1/2}\beta_m$, factors can be interpreted according to their correlation with the coding categories. The SO's are represented on a factorial plane by the MCAR's. As before, in SO-PCA, we introduce a cohesion constraint in this most general factorial approach (GCA) by the already defined matrix $\mathbf{A}_{N \times n}$.

The constrained GCA can also be interpreted as the Non-Symmetrical Multiple Correspondence Analysis approach proposed by Lauro and D'Ambra (1988) and further generalized by Verde and Lauro (1993) to the analysis of fuzzy coded data. The maximized criterion expresses the explanatory power of the coded descriptors with respect to the set of *hypercube vertices* which are a geometric representation of each SO. According to this approach the factorial axes are obtained as solution of the characteristics equation:

$$\frac{1}{N} \mathbf{A}' \mathbf{Z} \Sigma^{-1} \mathbf{Z}' \mathbf{A} \xi_m = \tilde{\mu}_m \tilde{\xi}_m \quad (4)$$

under the orthonormality constraints : $\tilde{\xi}_m' \tilde{\xi}_m = 1$ and $\tilde{\xi}_m' \tilde{\xi}_{m'} = 0$ ($m \neq m'$).

The maximum number of non-trivial eigenvalues is $(K - p + 1)$, if $K < N$ (where $K = \sum_{j=1}^p k_j$). It is worth noticing that, assuming that the matrix \mathbf{A} is centered, the *trace* of the matrix in 4 is equal to the sum of the Goodman-Kruskal predictive τ indices, except for a constant term, and respect to each binary or fuzzy coded descriptor. The coordinates of the vertices of the i -th SO on the m -th factorial axis are $\tilde{\xi}_{i,m}$, and can also be expressed as $\mathbf{Z}_i \tilde{\beta}_m$, where $\tilde{\beta}_m$ is a solution of the dual analysis in \mathbb{R}^K .

3.3 SO-Factorial Discriminant Analysis

Given a set of n SO's described by p independent generic descriptors (numerical *interval valued*, nominal and modal) and one dependent categorical variable, we assume that the latter defines a partition into r different groups of the n SO's. The aim of the Factorial Discriminant Analysis on SO's (SO-FDA) (Lauro et al. 1999) is to define a geometric classification rule in order to assign new SO's to one of the r classes on the basis of the same set of p descriptors. The SO-FDA aims at defining the best discriminant subspace which is defined first via a descriptors quantification and selection step, then performing a classical FDA on the new descriptors. According to Lauro et al. (1999), the SO-GCA has been proposed as a quantification procedure of coded descriptors in the SO-FDA. In addition, the additivity of the optimized criterion in the SO-GCA provides a way to select the best discriminant predictors. The SO's to classify and the classes are represented by the MCAR's making us to define several different decision rules on different cases: *i*) if a SO image is included in the representation of C_i (of the training set) it is assigned to the same class of C_i ; *ii*) if a SO is partially or completely outside any C_i representation or it is in an overlapping area between two or more representations, we shall consider a suitable measure of similarities to assign

the element. The new element will be assigned to the class with which it presents the highest similarity on the basis of a geometric rule.

The *Description Potential* quantity $\pi(\cdot)$, defined by De Carvalho (1992) as the volume obtained by the Cartesian product of the domains of the variables, was used to define a proximity measure by Lauro's et al. (1999), but other measures can also be considered (see Bock and Diday, 1999).

4 Concluding remarks

In this paper, we show how factorial methods can be improved in order to get suitable visualization/discrimination of SO's preserving their wholeness by cohesion constraints.

Further prominent aspects, strictly related to the factorial SDA, have not been treated in the present paper because of the lack of space. Here follows a short list of open problems that could be topics for new contributions: quality and interpretation of the representation; the computational cost aspects that could be very heavy in the case of many mixed SO descriptors.

References

- BOCK, H. H. and DIDAY, E. (eds): 1999, *Analysis of Symbolic Data*, Springer-Verlag, Heidelberg.
- CAZES, P., CHOUAKRIA, A., DIDAY, E. and SCHEKTMAN, Y.: 1997, Extension de l'analyse en composantes principales à des données de type intervalle, *Revue de Statistique Appliquée XIV*(3), 5–24.
- CHOUAKRIA, A., DIDAY, E. and CAZES, P.: 1998, An improved factorial representation of symbolic objects, *KESDA '98 27-28 April*, Luxembourg.
- CHOUAKRIA, A., DIDAY, E. and CAZES, P.: 1998, Vertices Principal Components with an Improved Factorial Representation, In: A.Rizzi, M.Vichi, and H.H. Bock Eds.: *Advances in Data Science and Classification*. Springer-Verlag.
- D'AMBRA, L. and LAURO, C. N.: 1982, Analisi in componenti principali in rapporto a un sottospazio di riferimento, *Rivista di Statistica Appl.* **15**(1), 51–67.
- DE CARVALHO, F. A. T.: 1992, *Méthodes Descriptives en Analyse de Données Symboliques*, Thèse de doctorat, Université Paris Dauphine, Paris.
- DIDAY, E.: 1987, Introduction à l'approche symbolique en analyse des données, *Journées Symbolique-Numerique*, Université Paris Dauphine.
- LAURO, C., VERDE, R. and PALUMBO, F.: 1999, Factorial Discriminant Analysis on Symbolic Objects, in Bock, H. H. and Diday, E. (eds): 1999, *Analysis of Symbolic Data*, Springer Verlag, Heidelberg.
- LAURO, C. and PALUMBO, F.: 2000, Principal Component Analysis of Interval Data: a Symbolic Data Analysis Approach, Comp.Stat. (forthcoming).
- LEBART, L., MORINEAU, A. and PIIRON, M.: 1995, *Statistique exploratoire multidimensionnelle*, Dunod, Paris.
- VERDE, R. and LAURO, C.: 1993, Non symmetrical data analysis of multiway fuzzy coded matrices, *ISI*, Firenze.
- VERDE, R.: 1997, Symbolic object decomposition by factorial techniques. *Franco-Indian Meeting*, LISE-CEREMADE, Université Paris IX Dauphine.

A Dynamical Clustering Algorithm for Multi-nominal Data

Rosanna Verde¹, Francisco de A. T. de Carvalho², and Yves Lechevallier³

¹ Facolta de Economia, SUN - Seconda Universita di Napoli, P.za Umberto I, 81043 Capua, Italia (e-mail: verde@dms.unina.it)

² Centro de Informatica - UFPE, Av. Prof. Luiz Freire s/n, Cidade Universitaria, 50740-540 Recife - PE, Brasil (e-mail: fatc@di.ufpe.br)

³ INRIA - Rocquancourt, Domaine de Voluceau - Rocquencourt B. P. 105, 78153 Le Chesnay Cedex, France (e-mail: Yves.Lechevallier@inria.fr)

Abstract. In this paper we present a dynamical clustering algorithm in order to partition a set of multi-nominal data in k classes. This kind of data can be considered as a particular description of *symbolic objects*. In this algorithm, the representation of the classes are given by prototypes that generalize the characteristics of the elements belonging to each class. A suitable allocation function (*context dependent*) is considered in this context to assign an object to a class. The final classes are described by the distributions associated to the multi-nominal variables of the elements belonging to each class. That representation corresponds to the usual description of the so called *modal symbolic objects*.

1 Introduction and symbolic object notation

This work aims at adapting the general dynamic clustering algorithm (Diday et al (1976), Celeux et al (1989)) for partitioning multi-nominal symbolic data (Bock and Diday, 1999). The criterion optimized in the classical algorithm is based on the best agreement between the classes and their representations. In order to represent the classes, we use "*prototypes*" such as means, axes, probability laws, groups of elements, etc. In the classical dynamical algorithm it is necessary to specify: *i*. the set E of objects to be clustered in k classes; *ii*. the set of variables describing objects; *iii*. the prototypes representing the classes; *iv*. an allocation function based on a proximity measure between objects and prototypes; *v*. a structure of object partitions; *vi*. the criterion of the best fitting between the structure of the partition in k classes and the k prototypes.

In this work we consider a set E of Symbolic Objects (SO's), described by multi-nominal variables y_j ($j = 1, \dots, p$) with domain in the set $D = (D_1, \dots, D_p)$. This data corresponds to the description of symbolic data according to standard definition (Diday (1998)). Each SO $x_i \in E$ is expressed by the triple (a, R, d) . The $d = (d_1, \dots, d_p)$, with $d_j \subseteq D_j$, is the set of values assumed by SO descriptors y_j which in the classical definition of SO's can be also ordinal or numerical *interval valued* variables: usually, such SO's

are called *Boolean* SO's, while *modal* SO's are characterized by distributions associated to the descriptors. Furthermore $R = (R_1, \dots, R_p)$ is the set of relations and a is a *mapping function* in order to compute the extent of SO.

According to the nature of the data, the first phase of the proposed algorithm consists of defining a suitable way to represent the classes of objects. Similarly to the dynamical clustering method on individual data, we refer to suitable prototypes G_h in order to represent the clusters. In this context, we propose to construct prototypes by summarizing the whole information of the SO's belonging to the different classes. Each prototype associated to a class can be modeled as a *modal SO* (Diday, 1998). The description of a modal SO is given by frequency (or probability) distributions associated to the p variables. The other main aspect of this algorithm concerns the choice of the proximity function in order to assign SO's to the classes. In particular we consider a suitable *context dependent proximity function* δ (De Carvalho et al., 1998) in order to compare the objects to prototypes.

The clustering algorithm proposed here consists of two steps: the first is the description step of the classes by prototypes. The peculiarity of the proposed approach consists of getting a class descriptions in terms of modal SO's; the second one is an allocation step in order to assign the objects to the classes, according to their proximity to the prototypes. The convergence of the algorithm to a stationary value of the criterion function Δ is guaranteed by the best fitting between the type of representation of the classes and the allocation function.

An example on *Microorganism data* is proposed to corroborate the procedure.

2 Dynamic clustering objects algorithm

According to the standard dynamic clustering algorithm (Celeux et al., 1989) we look for the partition $P \in P_k$ of E in k classes, among all the possible partitions P_k , and the vector $L \in L_k$ of k prototypes representing the classes in P , such that a criterion Δ of fitting between L and P is minimized:

$$\Delta(P^*, L^*) = \text{Min}\{\Delta(P, L) / P \in P_k, L \in L_k\}$$

The dynamic algorithm is performed by the following steps:

- a) *Initialization*; a partition $P = (C_1, \dots, C_k)$ of E is randomly chosen.
- b) *representation step*:
 - for $h = 1$ to k , find the prototype G_h associated with C_h , such that $\sum_{x_i \in C_h} \delta(x_i, G_h)$ is minimized
- c) *allocation step*:
 - test $\leftarrow 0$
 - for all x_i do
 - find m such that C_m is the class of x_i

```

find  $l$  such that:  $l = \arg \min_{h=1, \dots, k} \delta(x_i, G_h)$ 
if  $l \neq m$ 
    test  $\leftarrow 1$ 
     $C_l \leftarrow C_l \cup \{x_i\}$  and  $C_m \leftarrow C_m - \{x_i\}$ 
d) if  $test = 0$  then stop, else go to b)

```

Then, the first choice concerns the definition of the representation structure of the classes $\{C_1, \dots, C_k\} \in P$ by prototypes (G_1, \dots, G_k) . The criterion $\Delta(P, L)$, optimized in the dynamic clustering algorithm, is usually defined as the sum of the measures $\delta(x_i, G_h)$ of fitting between each object x_i belonging to a class $C_h \in P$ and the class representation $G_h \in L$:

$$\Delta(P, L) = \sum_{i=1}^k \sum_{x_i \in C_h} \delta(x_i, G_h)$$

where δ is a suitable dissimilarity or distance function, see Section 2.1.

2.1 Representation and allocation functions

The representation space of the SO's $x_i \in E$ to be clustered and the representation space L associated to each partition P of E , are not homogeneous, as the x_i are Boolean SO's, and the prototypes, representing the classes, are *modal* SO's. In order to retrieve the two configurations of data in the same representation space, we associate suitable distributions to the variables, in the description of each SO.

Let d_{j,x_i} be the set of categories of a multi-nominal variable y_j in the description of the SO x_i . The distribution value q_{j,x_i} , associated to the category c_s ($s = 1, \dots, \text{card}(d_{j,x_i})$) of y_j , is given by:

$$q_{j,x_i}(c_s) = \begin{cases} \frac{1}{|d_{j,x_i}|} & \text{if } \{c_s\} \in d_{j,x_i} \\ 0 & \text{otherwise} \end{cases}$$

The h -th prototype description d_{j,G_h} is defined by a linear combination of the distributions q_{j,x_i} ($j = 1, \dots, p$) of the $x_i \in C_h$: $g_{j,G_h} = f(q_{j,x_i})$.

The allocation function δ can be considered as a suitable comparison function which measures the matching degree between the SO x_i and the prototype G_h ($\forall x_i, G_h$), according to their description d_{G_h} and d_{x_i} . The particular allocation function that we consider is based on the following two additive components:

$$\delta(x_i, G_h) = \sum_{j=1}^p \sum_{s: c(s) \in D_j} (\gamma_s \cdot g_{j,G_h}(c_s) + \gamma'_s \cdot q_{j,x_i}(c_s))$$

where γ_s and γ'_s can take values $\{0, 1\}$, under the following conditions:

$$\gamma_s = \begin{cases} 1 & \text{if } c_s \in d_{j,G_h} \& c_s \notin d_{j,x_i} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \gamma'_s = \begin{cases} 1 & \text{if } c_s \in d_{j,x_i} \& c_s \notin d_{j,G_h} \\ 0 & \text{otherwise} \end{cases}$$

3 Application to microorganism virtual data

The proposed clustering approach is exemplified on a remaking of the data set used by Diday and Michalski (1981) in conceptual classification context. The objects to be clustered are "Microorganisms", as shown in Figure 1.

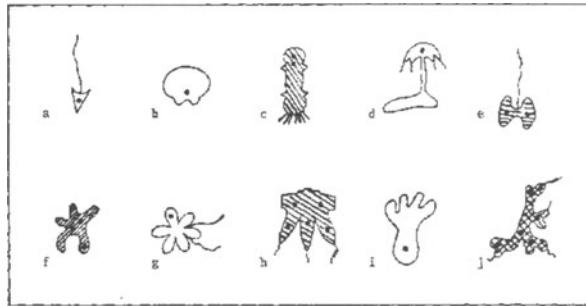


Fig. 1. Microorganisms picture.

The symbolic descriptions of the 10 Microorganisms are given in Table 1.

Micro-organism	body spots	body part	tail type	texture
a	bottom	triangle	top	white
b	bottom	circle	nobody	white
c	top	rectangle, circle	bottom	black, white
d	top	rectangle, circle	top, bottom	white
e	bottom, right, left	circle	top	black, white
f	top, right, left	nobody	striped	black, white
g	top	circle	right	black
h	top, bottom, right, left	triangle, rectangle	bottom	black, white
i	bottom	rectangle, circle	bottom, right	white
j	top, bottom, right, left	rectangle, circle	bottom, right	black, white

Table 1. Symbolic Microorganisms descriptions

According to our approach, the 10 objects will be clustered in 3 classes, by using the *referred* context dependent proximity measure as allocation function.

In this application, the algorithm is initialized by considering the following partition: $C_1 = \{a, b, c, e, j\}$, $C_2 = \{d, g, h\}$ and $C_3 = \{f, i\}$.

The symbolic description of the prototypes are shown in Table 2. The matching values between the objects and the classes are presented in Table 3,

Prototypes	body spots	body part	tail type	texture
G_1	$\text{top}_{(1/4)}$, $\text{bottom}_{(31/60)}$ $\text{right}_{(7/30)}$, $\text{left}_{(7/30)}$	$\text{triangle}_{(1/5)}$, $\text{rectangle}_{(1/5)}$ $\text{circle}_{(3/5)}$	$\text{top}_{(2/5)}$, $\text{bottom}_{(3/10)}$ $\text{right}_{(1/10)}$, $\text{nobody}_{(1/5)}$	$\text{black}_{(3/10)}$ $\text{white}_{(7/10)}$
G_2	$\text{top}_{(3/4)}$, $\text{bottom}_{(1/12)}$ $\text{right}_{(1/12)}$, $\text{left}_{(1/12)}$	$\text{triangle}_{(1/6)}$, $\text{rectangle}_{(1/3)}$ $\text{circle}_{(1/2)}$	$\text{top}_{(1/6)}$, $\text{bottom}_{(1/2)}$ $\text{right}_{(1/3)}$	$\text{black}_{(1/6)}$ $\text{white}_{(5/6)}$
G_3	$\text{top}_{(1/6)}$, $\text{bottom}_{(1/2)}$ $\text{right}_{(1/6)}$, $\text{left}_{(1/6)}$	$\text{rectangle}_{(1/4)}$ $\text{circle}_{(3/4)}$	$\text{nobody}_{(1)}$	$\text{black}_{(1/4)}$ $\text{white}_{(3/4)}$

Table 2. Symbolic descriptions of prototypes - at Step 1

	$\delta(., G_1)$	$\delta(., G_2)$	$\delta(., G_3)$	partition
a	2.18	2.75	4.75	C_1
b	1.98	3.58	1.0	C_3
c	1.65	0.92	2.83	C_2
d	1.55	0.92	3.08	C_2
e	1.25	2.08	2.42	C_1
f	1.72	2.58	0.75	C_3
g	2.35	1.58	3.33	C_2
h	1.3	1.0	3.25	C_2
i	1.78	3.25	0.75	C_3
j	0.8	0.33	2.0	C_2

Table 3. Dissimilarity between objects and prototypes - at Step 1

where the last column contains the assignment class of each object.

The achieved new partition $P^{(1)}$ of the Microorganism, is: $C_1 = \{a, e\}$, $C_2 = \{c, d, g, h, j\}$ and $C_3 = \{b, f, i\}$, differing from the previous one for the elements b, c, j that move from the class C_1 to the classes C_2 and C_3 .

The partial criterion, evaluated with respect to each cluster, is: $\Delta_1^{(1)} = 3.43$; $\Delta_2^{(1)} = 4.75$; $\Delta_3^{(1)} = 2.5$, and the global criterion is: $\Delta^{(1)} = 10.68$.

The partition $P^{(2)}$, obtained at Step 2, is: $C_1 = \{a, e\}$, $C_2 = \{c, d, g, h, j\}$ and $C_3 = \{b, f, i\}$. It is coincident with the one achieved at the previous step. The partial criterion values are: $\Delta_1^{(2)} = 2.0$; $\Delta_2^{(2)} = 4.7$; $\Delta_3^{(2)} = 1.58$ and the global criterion value is equal to 8.28, which is less than $\Delta^{(1)}$.

The algorithm is stopped at this best partition $P^{(2)}$, whose objects in the clusters are represented in Figure 2.

It is worth noticing that the partition we have achieved with the proposed clustering approach is equivalent to the best partition given by *PAF* algorithm in Diday and Michalski (1981). The main difference between the two approaches concerns the clustering context and the description of the data. In fact, they proposed a conceptual criterion to cluster the same Microorganism set, but described by classical nominal variables, in order to obtain a partition, corresponding to best human solution (Diday, Michalski. 1981).

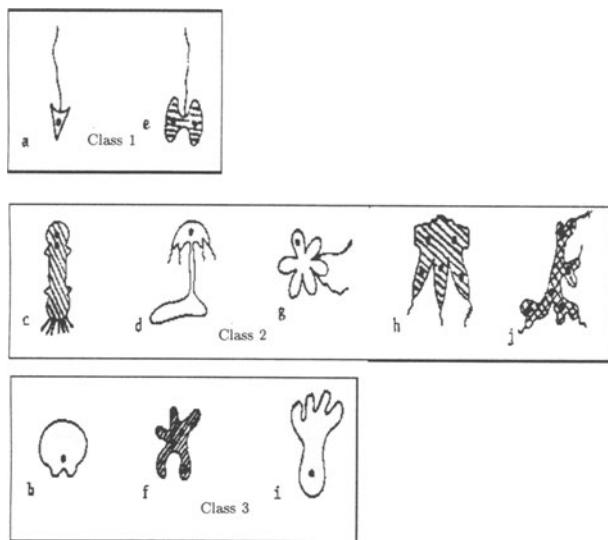


Fig. 2. Final partition

4 Final comments

The results we have obtain on the multi-nominal data are coherent with the partition given by a conceptual clustering algorithm, as we have shown in the application. However, our approach is based on the optimization of a numerical function and it gives an efficient solution. Therefore, it allows to overcome the combinatorial complexity of the conceptual clustering algorithm.

Acknowledgements This paper is supported by grants: CNPq proc. 301387/92-3 and Cofin. MURST 1999 "Modelli statistici di classificazione e di segmentazione"

References

- BOCK, H. H., DIDAY, E. (eds.), (1999): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.
- CELEUX, G., DIDAY, E., GOVAERT, G., LECHEVALLIER, Y., RALAMBONDRAINY, H. (1989): *Classification Automatique des Données*. Bordas, Paris.
- DE CARVALHO, F.A.T. and SOUZA, R. M. C. (1998): Statistical proximity functions of Boolean symbolic objects based on histograms. In: Rizzi, A., Vichi, M., Bock, H.-H. (Eds.): *Advances in Data Science and Classification*, Springer-Verlag, Heidelberg, 391–396.
- DE CARVALHO, F.A.T., VERDE, R. and LECHEVALLIER, Y. (1999): A dynamical clustering of symbolic objects based on a context dependent proximity

- measure. In: Bacelar-Nicolau, H., Nicolau, F.C. and Janssen, J. (Eds.): *Proc. IX International Symposium - ASMDA'99*. LEAD, Univ. de Lisboa, 237–242.
- DIDAY, E. and SIMON, J. C. (1976): Clustering Analysis. In: Fu, K. S. (Eds.): *Digital Pattern Recognition*. Springer-Verlag, Heidelberg, 47–94.
- MICHALSKI, R. S., DIDAY, E. and STEPP, R. E. (1981): A recent advance in data analysis : Clustering Objects into classes characterized by conjunctive concepts. In : Kanal L. N. and Rosenfeld A. (Eds.): *Progress in pattern recognition*. North-Holland, 33–56.

DB2SO : A Software for Building Symbolic Objects from Databases

Georges Hébrail¹ and Yves Lechevallier²

¹ EDF- Research and Development Division

1, Av. du Général de Gaulle
92141 Clamart Cedex, France
(e-mail: Georges.Hebrail@edf.fr)

² INRIA-Rocquencourt,
78153 Le Chesnay CEDEX, France
(e-mail: Yves.Lechevallier@inria.fr)

Abstract. The SODAS project, funded by EC, has developed a software for extending statistical data analysis methods to more complex objects. Objects processed by these methods are complex in the sense that they represent groups of individuals, featuring variation among each group of individuals. Within the context of the SODAS project, the complex objects are called symbolic objects. In this paper, we present a part of the SODAS software, which enables the user to acquire datasets of symbolic objects, by extracting information from relational databases.

1 Introduction

During the last decade, Diday (1988) introduced the concept of symbolic objects and symbolic data analysis. The basic idea was to extend standard data analysis methods to analyze objects described by a more complex structure than a simple table row. The first part of his work was to define more complex data structures associated with some semantics. The second part was to extend standard statistical data analysis methods to treat these new complex data structures. The main goal of the SODAS project, funded by the EC, was to build a software which demonstrates the approach to be valid and operational. This project ended at the end of 1999 (Bock and Diday (1999)).

In this paper, we present a part of the SODAS software: the DB2SO module, which enables the user to create a set of symbolic objects from data stored in a relational database (Stéphan et al. (1999)).

In Section 2, we present the principles of symbolic object construction from the contents of a relational database. Two main steps are performed: (1) building of symbolic objects corresponding to descriptions of groups of individuals (by a generalization process), and (2) refining the description to make them simpler by removing some untypical individuals. The user defines queries to the database: these queries must fit specified expected structures.

In Section 3, we show that symbolic objects built by the basic process can be enriched by metadata, still picking up information from the database:

DB2SO supports addition of taxonomies in variable domains, and definition of mother/daughter variables.

Section 4 finally presents some additional features which facilitate the process of constructing of a set of symbolic objects from a database.

This paper is an introduction to construction of symbolic objects from the contents of relational databases. Readers wishing to have complete theoretical information about this process can refer to Stéphan (1998), Stéphan et al. (1997, 1999). Hereafter, symbolic object will be referred as *SO*.

2 Symbolic object construction by generalization

DB2SO generates SO's called *boolean* and *modal* symbolic objects in the framework of E. Diday. As described in (Bock and Diday (1999)), a SO is defined by a triple (a, R, d) where d is a *description*, R is a *comparison operator* between descriptions, and a is a *mapping* which defines the extension of the SO. We focus here on the construction of the description part of SO's, as illustrated in Fig.1. Each description of SO generated by DB2SO is the representation of a group of individuals by some symbolic variables. Symbolic variables describe variations among each group of individuals by figuring:

- The interval of observed values on individuals in the group for numerical variables,
- The list of observed values on individuals in the group for nominal variables,
- The probability distribution of observed values on individuals in the groups for nominal variables, when the user asks for a modal SO.

2.1 Basic process

Input of the basic process is a table describing individuals from a population. Individuals are described either by numerical or nominal variables. The table below is an example of such data input. The user has to write an SQL query which returns such a table with the following expected structure: the first column describes the individual ID, the second one the group ID the individual belongs to, and others columns represent characteristics of individuals.

In our example, individuals are private households : groups are regions of United Kingdom, household characteristics are the number of bedrooms and dining/living rooms, the presence/absence of a garage and its size when there is one, and the socio-economic group of the household.

CASENO	Region	Bedroom	Dining Living	Garage	Garage size	Socio-Economic Group
114051	Northern metropolitan	2	1	No		Managers
114111	Northern metropolitan	2	1	Yes	2	Managers
114131	Northern metropolitan	1	1	No		Intermediate non-manual
...						
201081	North non-metropolitan	2	1	No		Semi-skilled manual
201091	North non-metropolitan	4	3	Yes	1	Skilled manual
...	...					

In DB2SO, we assume that the second column of the returned query is nominal. It is used for defining groups of individuals : each group will be described by one SO. The variables associated with generated SO's are thus all query columns except the two first ones.



```

dbgip2 WordPad
Fichier Edition Affichage Insertion Format ?
variable Bedroom
real [1:4]
interval;
variable Dining\living
real [1:3]
interval;
variable Garage
nominal ("No", "Yes")
multiple,mode=probabilist;
variable Garage_size
real [1:2]
interval;
variable Socio_Economic_Group
nominal ("Intermediate non_manual ancill", "Junior non_manual", "Managers : large establishment", "Own account nor
multiple,mode=probabilist;
|
os "Northern metropolitan"(9) =
[Bedroom = [1:3]]
^(Dining\living = [1:2])
^(Garage = ("Yes"(0.222222), "No"(0.777778)))
^[Garage_size = [1:2]]
^Socio_Economic_Group = ("Semi-skilled manual"(0.111111), "Unskilled manual"(0.111111), "Skilled manual"(0.111111))
|
os "North non-metropolitan"(7) =
[Bedroom = {2:4}]
^(Dining\living = [1:3])
^(Garage = ("Yes"(0.285714), "No"(0.714286)))
^[Garage_size = [1:2]]
^Socio_Economic_Group = ("Semi-skilled manual"(0.142857), "Unskilled manual"(0.142857), "Junior non_manual"(0.14
|
Pour défaire, appuyez sur F1

```

Fig. 1. Symbolic Objects generated by DB2SO

The description of SO is generated using a *generalization* function to aggregate individuals of each group:

- Numerical variables describing individuals lead to interval variables describing groups: the generalization function is here to generalize the set of variable values to the interval of values from the minimum to the maximum of them (see for instance the *bedroom* variable in Fig.1),
- Nominal variables describing individuals lead either to *boolean* or *modal* multi-valued variables describing groups. If the user chooses to generate a boolean multi-valued variable, the aggregation function is simply building the list of observed values within the group. If the user chooses to generate a modal multi-valued variable, the aggregation function builds the probability distribution of the nominal variable among individuals of the group. In Fig.1, a modal multi-valued variable has been selected for the *socio-economic-group* variable.

2.2 Symbolic object refinement

The basic generalization functions described in the previous section may lead to group descriptions of poor quality. Since the generalization process is applied separately on each variable, one may obtain similar descriptions for different groups, even if groups have very different contents. In particular, if groups contain untypical individuals, intervals are large.

The idea of the refinement process is to remove from each group some untypical individuals. Each SO, i.e. each group of individuals is processed separately. With each SO an extension mapping is associated which says if an individual is recognized by a SO or not. In the simple case of interval variables or boolean multi-valued ones, this function says 'yes' if individual values for every variable are in the interval or list of values of the corresponding symbolic variables. Untypical individuals are removed and simpler description of SO is built again from the remaining individuals. The following optimization constraints guide the removal of individuals :

- A minimum threshold is defined on the number of individuals of the group still recognized by the refined description of SO (this threshold is given by the user, typically 80%),
- A volume measure is associated with each SO, representing the amount of subspace its variable values occupy in the Cartesian product of variable domains. The choice of untypical individuals to be removed is done to maximize the decrease of volume associated with the refinement process.

Details of this refinement process and a description of the algorithm can be found in (Stephan (1998)).

3 Adding metadata

The model of symbolic objects allows to associate metadata with a set of SO's. In DB2SO, two ways are available to specify metadata associated with generated SO's: the definition of mother-daughter variables, and the definition of taxonomies linking together possible values of nominal variables.

3.1 Mother-daughter variables

Mother-daughter variables are variables which are linked semantically by the fact that daughter variables have a value (we say: *are applicable*) only for some values of a mother variable. In our example, the *garage_size* variable is applicable only if the *garage* variable is set to 'yes'. The *garage* variable is said to be the mother variable of the *garage_size* variable.

DB2SO enables the user to define mother-daughter variables in a very simple way, including a checking of the corresponding integrity constraints on available data.

3.2 Taxonomies on variable domains

The second way to introduce metadata associated with SO's is to define taxonomies within nominal variable possible values. DB2SO assumes that the taxonomy is stored in the database in a child/parent representation as shown in the table below. The user writes a query to the database which returns such a table. In the example below, the taxonomy is associated with the *socio-economic-group* variable.

Socio-Economic Group	Parent Group
Employers	Self-employed
Employers: large establishment	Employers
Employers: small establishment	Employers
Farmers	Self-employed
Farmers:emp&mgrs	Farmers
Farmers:own account	Farmers
Managers	Employed
Manual	Employed
Own account non-professional	Self-employed
Professional - self employed	Self-employed

Once a taxonomy is acquired by DB2SO, it is available in a usual graphical representation as a tree. Taxonomies associated with SO's can then be used by the SODAS methods as metadata.

4 DB2SO useful additional features

The last sections presented the main operations that can be performed with DB2SO. This module also includes some other less important features, but which really help the user in his(her) task.

A first facility is the possibility of *sampling* the result of the query returning the initial set of individuals from the database. As a matter of fact, this result can be very large and not fit in main memory. A sampling facility is available in this case. It is important to note that the algorithm does not have to know in advance the number of returned records by the query: a 'reservoir' algorithm has been chosen (Vitter (1985)).

In the basic process described in Section 2, the only variables that can describe SO's are variables initially describing individuals. In many applications, the database may contain data describing directly the groups of individuals. For instance, in our example, we may have variables describing regions independently of households, like the surface area of the region. DB2SO enables the user to add such variables to SO's already built from a set of individuals.

An interesting and typical use of SO's is to describe some real-world objects by several variables, possibly describing very different aspects of the objects. For instance, we may describe regions by household characteristics, as well as by school characteristics. Beyond the construction of region SO's based on households, we can build SO's which still represent regions, but

described by characteristics of the schools of the region: in this case, the individuals used by DB2SO are schools in the country. If the user wants to have in the same SO's both information describing households and schools, DB2SO provides a *merging* facility, which takes as input two sets of SO's and produces a new set of SO's described by the union of variables of the two input sets.

Finally, DB2SO provides several basic facilities for the user, like browsing of individuals and generated SO's, exportation of SO's symbolic data matrix to the SODAS software.

5 Conclusion and further work

In this paper, we have presented very briefly a module of the SODAS software which helps the user to build SO's from the contents of a relational database. This module is the result of many interactions between end-users of the software and specialists in databases and symbolic data analysis.

Some further work can be undertaken to improve this module. A very promising direction is to extend this module to be able to run again all queries submitted to the database, in a temporal analysis perspective. For instance, if it is possible to rerun the query returning individuals distributed into groups, new SO's can be built a year later, and differences with old SO's can be very interesting for analyzing evolutions during the period.

References

- BOCK, H. H., DIDAY, E. (eds.), (1999): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*. Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.
- DIDAY, E., (1988): The symbolic approach in clustering and related methods of data analysis: The basic choices. In IFCS-87, H. H. Bock (ed.), 673-684.
- STEPHAN, V., HEBRAIL, G., LECHEVALLIER, Y., (1999): Generation of Symbolic Objects from Relational Databases. In *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Springer-Verlag, 78-105.
- STEPHAN, V., (1998): *Construction d'objets symboliques par synthèse des résultats de requêtes SQL*. PhD thesis, Université Paris IX Dauphine.
- STEPHAN, V., HEBRAIL, G., LECHEVALLIER, Y., (1997): Improving symbolic descriptions of sets of individuals : the reduction of assertions. In *8th international symposium on Applied Stochastic Models and Data Analysis*, 407-412, Anacapri, Italy.
- VITTER, J. S., (1985): Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11, 37-57.

Symbolic Data Analysis and the SODAS Software in Official Statistics

Raymond Bisdorff¹ and Edwin Diday²

¹ Dpt Gestion et Informatique, Centre Universitaire
162a, avenue de la Faïencerie, L-1330-LUXEMBOURG
(e-mail: bisdorff@cu.lu)

² CEREMADE, Université Paris-Dauphine
F-75775 Paris Cédex 16, France
(e-mail: diday@ceremade.dauphine.fr)

Abstract. The need to extract new knowledge from complex data contained in relational databases is increasing. Therefore, it becomes a task of first importance to summarise huge data sets by their underlying concepts in order to extract useful knowledge. These concepts can only be described by more complex data type called "symbolic data". We define "Symbolic Data Analysis" (SDA) as the extension of standard Data Analysis to symbolic data tables. The "Symbolic Data Analysis" theory is now enhanced by a new software tool called "SODAS" which results from the effort of 17 European teams (sponsored by EUROSTAT). This is shown by several applications in Official Statistics.

1 Introduction

The data descriptions of the units are called "symbolic" when they are more complex than the standard ones due to the fact that they contain internal variation and are structured. Symbolic data happen from many sources, for instance in order to summarise huge sets of data and to describe underlying concepts of a Data Base (as regions , socio-demographic groups, unemployed types, scenario of traffic accidents,etc). They need more complex data tables called "symbolic data tables" because a cell of such data table does not necessarily contain as usual, a single quantitative or categorical value. For instance, a cell can contain, a distribution (Schweitzer (1985) says that "distributions are the number of the future"!), or several values linked by a taxonomy, or intervals with logical rules, etc. SODAS is a new way for DATA WAREHOUSING in order to extract symbolic data files from a Relational Data Base, and for DATA MINING in order to extract knowledge from this file.

"Extracting knowledge" means getting new explanatory concepts, which is why "symbolic objects" are introduced. Symbolic Objects as defined in Diday (1999) or in the Chapter 1 of Bock and Diday (2000) model concepts and constitute an explanatory output of a Symbolic Data Analysis. Moreover they can be used to define queries of a Data Base, in order for instance to propagate concepts discovered from a country, to another country.

Hereunder, a first example illustrates SODAS facilities to manipulate complex statistical survey results such as given by the European labour force survey (LFS). By the way this example shows how statistical data from different European regions may be joined, visually explored and compared. A second example, again concerned by the LFS results, illustrates some symbolic clustering techniques of the SODAS software. Finally a third example shows symbolic discrimination techniques for time series from Social Security data¹

2 Manipulating European labour force survey results on a regional level

The example data consists of statistical results from the quarterly Portuguese (INE) and Basque (EUSTAT) Labour Force Survey (LFS), a total of 56,049 records describing individual persons. 17 important variables have been selected for illustration: gender, marital status, age, level of education, relation to labour force, territorial unit, search of employment, search of employment (first or another), profession, branch of economic activity, professional situation, time seeking employment, type of inactivity, length of working day, type of contract, working hours. To these variables has been added a sampling weight. All first sixteen variables are of qualitative nominal type whereas the

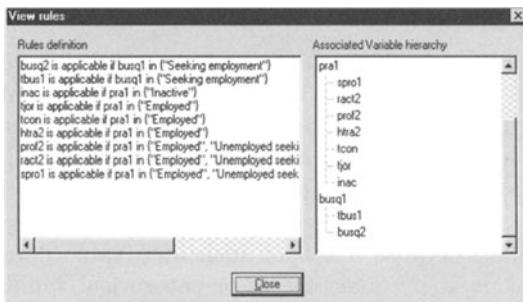


Fig. 1. Rules in SODAS



Fig. 2. Taxonomies in SODAS

sampling weight variable is of a quantitative continuous type. Due to the complex structure of the LFS questionnaire, some of our variables are logically dependent one on the others as for instance variable: 'search of employment (first or another)' which is only applicable if the variable 'search of employment' takes the value 'seeking employment'. Figure 1 shows the use of rules for handling such hierarchical mother-daughter variable structures within the

¹ More details on these examples are provided in chapter 13 of Bock and Diday (2000)

SODAS software tool. By the way, we may notice on Figure 2 that a variable marital status is constructed as a taxonomy first distinguishing between "single" and "not single" persons and then for "not single" persons whether they are "divorced or separate", "widow (er)" or "married". Both these features may be formulated in the SODAS software tool so as to be taken into account by all SODAS methods and tools.

As a result we are able, with the help of the database extraction tool (DB2SO) provided in the SODAS software package, to construct from statistical data tables through a standard database connection and corresponding SQL queries, a symbolic data table as shown in Figure 3. We obtain 10 sec-

	SEXO	eciv
Lisboa e Vale do Tejo	Man (0.47), Woman (0.53)	Single (0.26), Married (0.63), Widow/Mido (0.07), Divorced o (0.03)
Centro	Man (0.48), Woman (0.52)	Single (0.25), Married (0.65), Widow/Mido (0.08), Divorced o (0.02)
Norte	Man (0.48), Woman (0.52)	Single (0.28), Married (0.63), Widow/Mido (0.07), Divorced o (0.02)
Açores	Man (0.49), Woman (0.51)	Single (0.28), Married (0.62), Widow/Mido (0.09), Divorced o (0.01)
Alentejo	Man (0.48), Woman (0.52)	Single (0.24), Married (0.65), Widow/Mido (0.09), Divorced o (0.02)
Algarve	Man (0.48), Woman (0.52)	Single (0.23), Married (0.65), Widow/Mido (0.08), Divorced o (0.03)
Madeira	Man (0.46), Woman (0.54)	Single (0.33), Married (0.55), Widow/Mido (0.10), Divorced o (0.02)
Gipuzkoxa	Man (0.49), Woman (0.51)	Single (0.38), Married (0.54), Widow/Mido (0.07), Divorced o (0.01)
Bizkaia	Man (0.49), Woman (0.51)	Single (0.37), Married (0.54), Widow/Mido (0.08), Divorced o (0.01)
Araba	Man (0.50), Woman (0.50)	Single (0.36), Married (0.57), Widow/Mido (0.06), Divorced o (0.01)

Fig. 3. Extract of symbolic data table

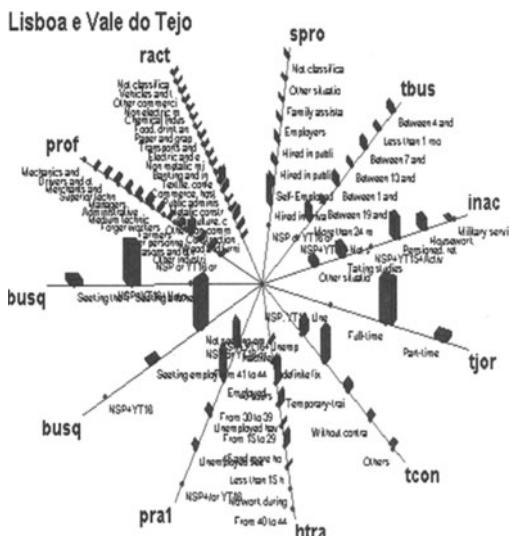


Fig. 4. Portuguese regional LFS results

ond order symbolic data units aggregating individual observations from the same territorial unit, i.e. Basque and Portuguese regions. Each region is described by 16 symbolic variables, each one associated with a finite categorical domain and a corresponding frequency distribution (see Figure 4).

One of the obvious applications of Symbolic Objects for Official Statistics concerns visual exploration and comparison of statistical data from different geographical regions out of the EU. Most EU countries do similar surveys, such as the Labour Force Survey (LFS), and the SODAS analysis software package proposes tools to easily explore and compare aggregated regional results without necessarily having to join *a priori* basic statistical data. To get an idea of similarities and/or differences among regions, we may directly compare the ten regions graphically by using the SODAS symbolic object editor (SOE). Figure 4 above shows for instance the Lisboa e Vale do Tejo region from Portugal as 3D Zoom Stars.

Let us now in the next example briefly illustrate symbolic data clustering results.

3 Clustering symbolic objects built from statistical surveys

The second illustrative analysis comes from the National Statistical Institute of Portugal (INE). The statistical data used here are again from the European Labour Force Survey providing quarterly results. It is regionally stratified by level NUTS II: Norte, Centro, Lisboa e Vale do Tejo, Alentejo, Algarve, Açores and Madeira. In each quarter, the national sample comprises around 22,000 households, representing about 45,000 individuals. This example uses the data concerning all individuals classified as workers on the LFS from the 2nd quarter of 1998, i.e. 22,660 individuals. Beside the classical personal information, i.e. gender, age group, marital status and education level, some activity variables of nominal or Boolean type were added such as: economic activity, profession, professional status, type of enterprise, searching employment, part/full time, social security contributor, over qualifications for the work, etc. The analyses were made to compare the activity by gender and age group. Therefore the number of build symbolic objects was that of $2 \times 6 = 12$. In order to uncover discriminating variables for these gender \times age groups, the SODAS factorial discriminant analysis (FDA) was used. To do so, a class variable was added in which the 12 symbolic objects were grouped based upon their age: the youngest, the oldest, and all the other workers in a third group. The graphical outcome of the FDA result is shown in Figure 5, where a nice discrimination of these three groups is illustrated. On the first axis we find the opposition between retired persons and all the others, and on the second axis we find the youngest opposed to all the others. A second analysis of the same data was conducted on the basis of interval data associated with probabilistic survey outcomes. The SODAS-PCM (Princi-

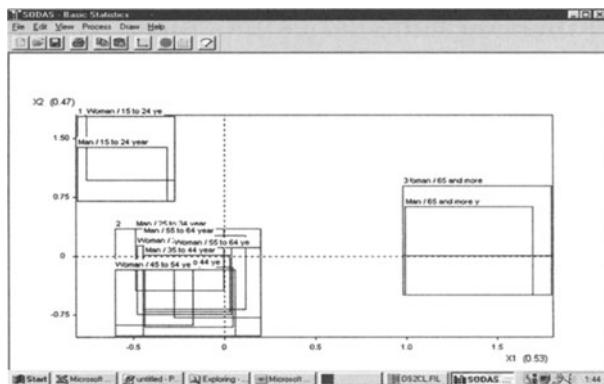


Fig. 5. SODAS-Factorial Discriminant Analysis

pal Components) method reveals the position of each symbolic object in the factorial plane (see Figure 6). The first factor shows again an opposition

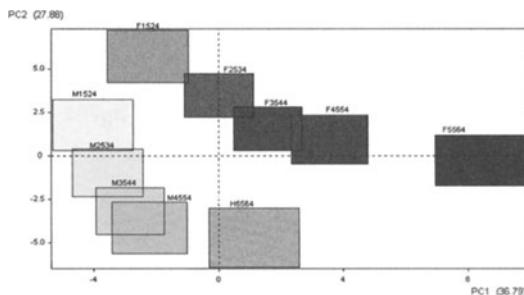


Fig. 6. SODAS-Principal Component Analysis

between the oldest workers (related with agriculture as part-time job) and the majority of the other symbolic objects (manufacturing, craft and related trades workers as employees in full-time jobs). In addition, there appears a growing on the age group variable from left to right. In the second factor, we observe an opposition in the gender variable: on the positive semi-axis the male group is more characterised in the legislators, senior officials and managers group as self-employed with employees. On the negative side, females are mainly defined in the clerks and in other services, especially in hotels and restaurants.

Let us now turn to the last example.

4 Discriminating time series modelled as symbolic objects

The statistical data in this last illustrative analysis describes exhaustively all complete (40 years) professional careers of persons having worked in Luxembourg and retiring within the year 1991, 1223 professional careers extracted from the administrative records of the Social Security Office in Luxembourg. Each individual career is described by the following 85 variables: birth year, gender, pension fund (workers, employees, independents, farmers), monthly pension, activity sector (17 main NACE rev. 1 sectors), yearly salaries from 1991 to 1952 (40 observations) and finally, yearly health care allowances from 1991 to 1952 (40 observations). Gender, pension fund and NACE rev. 1 (NACE) are qualitative variables of nominal type. The birth year is a qualitative variable of ordinal type. Monthly pension, yearly salaries as well as yearly health care allowances from 1952 to 1991 are all quantitative variables of a continuous type. The monthly pension and all salaries and health care allowances are represented in constant Luxembourgish francs expressed in base 100 in 1948. All these data were gathered as 86 symbolic objects of a Cartesian type by taking the product of birth year, gender, pension fund and activity sector. The purpose of the study now is to choose an appropriate small subset of years capturing a sufficient amount of the total variance over all these 40 years. To detect these most relevant variables, we used the divisive clustering method DIV provided in the SODAS software package to obtain the clustering tree shown in Figure 7. Following the advice of the segmentation, we could restrict our attention to the salaries from the following seven years: 1989, 1985, 1982, 1975, 1973, 1971 and 1958. Furthermore, only

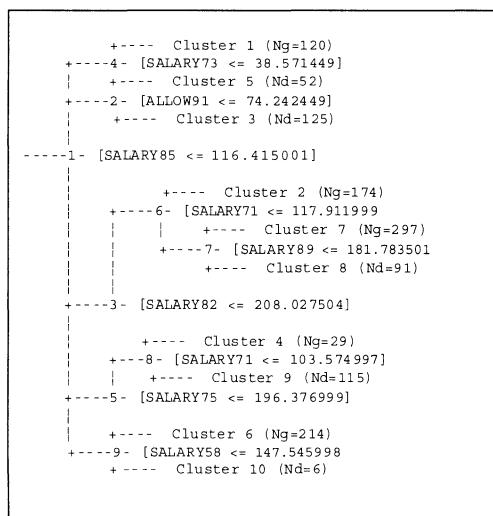


Fig. 7. Cluster tree from SODAS-DIV method

the health care allowances from the last working year 1991 seem to be of importance in order to discriminate the professional careers under review. Non-trivial as it is, this result is of great practical usefulness since a classical approach for selecting a subset of variables would have suggested a uniform distribution of the years to consider as for instance every fifth year.

5 Conclusion

The SODAS software facilitates the use of analysis techniques for numeric or symbolic data, and can be used in particular for data having a complex structure, to provide better explanations of statistical results, and to represent, manipulate or analyse better concepts and metadata. It provides as well a concept-oriented system for official statistics and databases recorded in companies.

References

- BOCK H. and DIDAY E. (2000): *Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data*, Springer Verlag, Heidelberg, 2000, 425 pages, ISBN 3-540-66619-2.
- DIDAY E. (1999): An Introduction to Symbolic Data Analysis and its Application to the SODAS European Project. *Cahiers du CEREMADE (UMR 7534)*, N. 9914.
- SCHWEITZER B. (1985): Distributions are the numbers of the future. In: *Proceedings of the Napoli meeting on "The mathematics of fuzzy systems"*. Instituto di Mathematica delle Faculta di Mathematica, Universita degli Studi di Napoli, 137–149.

Strata Decision Tree SDA Software

M. Carmen Bravo

Universidad Complutense de Madrid, Centro de Proceso de Datos, Av. Paraninfo s/n, 28040 Madrid, Spain. e-mail:carmen@sim.ucm.es

Abstract. The SDT and SDTEDITOR software are presented. The SDT (Strata Decision Tree) implements a generalised recursive tree-building algorithm for populations partitioned into strata and described by symbolic data, that is, more complex data structures than classical data. Symbolic objects describe decisional nodes and strata. The SDTEDITOR is a graph editor for strata decision trees. The SDT and SDTEDITOR are modules integrated into the SODAS Software (Symbolic Official Data Analysis System), partially supported by ESPRIT-20821 SODAS.

1 Introduction

The main aim of symbolic data analysis (SDA) is to extend classical methods in data analysis to more complex data structures. For example, data values may be given by a probability distribution: for *Employment*, the distribution given by $(yes(0.8), no(0.2))$ can represent either a subpopulation 20% of them unemployed, or an individual with 20% of his working life unemployed. Such a variable will be called a probabilistic-modal variable; Hierarchical dependence can be represented by *NA* rules: *If* (*smoke = no*), *then cigarette_mark = Non_applicable*; Groups of individuals are called strata.

The Strata Decision Tree (SDT) program implements a generalised recursive tree-building algorithm (Breiman et al. (1989)) for populations partitioned into strata, such as individuals of a country are divided into regions. Common predictor (including probabilistic variables and hierarchical dependencies) and criterion variables describe population in all strata. The algorithm considers strata structure. Symbolic objects describe decisional nodes and strata.

Some advantages of symbolic data analysis are: treatment of complex data structures and aggregated data, symbolic data can be derived from data bases or given by an expert, confidentiality of individuals is guaranteed, input and output language are the same understandable language to the user.

2 Algorithm

Input data. Let be Ω a set of individuals, $E = \{S_1, \dots, S_m\} \subset \mathcal{P}(\Omega)$ a partition of Ω . Thus, each element of E , $S_i \subset \Omega$ is a group of individuals, called a stratum (for $\omega \in \Omega$, $M(\omega) = i \iff \omega \in S_i$). Let individuals $\omega \in \Omega$ be described by the predictors $Y_j, j = 1, \dots, q$ and the criterion variable Z .

Two different input variable types are considered: (1) Ω a set of classical data: Variables Y_j are categorical single-valued mappings from Ω to \mathcal{Y}_j ; (2) Ω a set of probabilistic-modal data: Variables Y_j are probabilistic-modal variables with finite domain \mathcal{Y}_j . In both cases, Z is a categorical single-valued mapping from Ω to $\mathcal{Z} = \{1, \dots, s\}$.

Output data. A decision tree can be represented by an *organised* set of *assertions*, Ciampi et al. (1996). Each decisional node, described by the assertion $t_k = \beta_k \wedge \alpha_k \wedge \mu_k$, (see (1)) represents a set of strata for which the same rule for prediction of the criterion variable can be applied. An assertion a , Diday (1993), is a mapping from Ω into $[0, 1]$ and represents a logical predicate to be applied on the individuals $\omega \in \Omega$. The *extension* of a , $Ext_{\Omega}(a)$, contains the individuals that fulfil the logical predicate to a certain extent. $a(\omega)$ measures this fulfilment (See below). Tree is represented by:

$$T = \{\beta_k \wedge \alpha_k \wedge \mu_k\}_{k=1, \dots, K} \quad (1)$$

where K is the number of decisional nodes; β_k is a conjunction of objects (of $B = \{b = [Y_j \in D_j], b^c = [Y_j \notin D_j] | D_j \subset \mathcal{Y}_j\}$; D_j is a subset of the space of categories \mathcal{Y}_j) defined in the predictors Y_j ; α_k is a probabilistic symbolic object describing the prediction for Z ; and, $\mu_k = [M \in S^k]$ with $S^k \subseteq \{1, \dots, m\}$ is a Boolean object in the multi-valued categorical variable M . The μ_k is true for all individuals $\omega \in \Omega$ that belong to a stratum in S^k . Stratum indicators in S^k are identified in steps 3, 4 of the algorithm (see below). Assertion $\beta_k \wedge \mu_k$ (its extension) describes the population for which the prediction of Z is described by α_k . For example, a decisional node described by $[sex = f] \wedge [salh25 = yes] \wedge [clerk \sim (no(0.10), yes(0.90))] \wedge [NACE \in \{services, electric\}]$ gives for individuals in services and electricity $NACE$ sectors the rule *if sex is woman and mean gross hourly earnings is in the first quartile then estimated probability to be clerk is 0.9*.

Each stratum is also described by an *organised* set of *weighted assertions*, the decisional tree node descriptions where the stratum belongs to, Bravo and García-Santosmases (2000b).

Algorithm. The aim is to build recursively an *organised* set of assertions $T = \{t_k\}_{k=1 \dots K}$ (see (1)), by binary partitioning the population and combining at each step maximisation of an extended information content (EIC) measure of the tree with respect to E and selection of new decisional nodes. The EIC criterion measures quality of prediction for the criterion variable in a new partition, taking into account stratum membership in the cut. Quality of prediction is tested for subsets of strata in order to build decisional nodes. A decisional node is a leaf for some strata, while the other strata follow the recursive method. For the latest strata, a stopping criterion is also checked. In each step of the algorithm, T is composed of exploitable (obtained from the recursive partition, they can be binary split further on) and decisional nodes (split from an exploitable node, they are terminal). Quality of prediction for the criterion variable by the predictors and strata is given by

the information content measure (IC) of the tree with respect to Ω , Bravo and García-Santosmases (1997,2000a). This paper shows the measure choices implemented in the SDT program (see also section 4). Let X be the set of exploitable nodes in a step.

The IC and EIC measures are defined as:

$$IC\{T, \Omega\} = -\sum_{k=1}^K P(\beta_k \wedge \mu_k) Ent(Z|\beta_k \wedge \mu_k) \quad (2)$$

$$\begin{aligned} EIC\{T, r, b, E\} &= IC\{T(r), \Omega\} \\ &- P(\beta_r \wedge b \wedge \mu_r) \sum_{i \in S^r} P(S_i|\beta_r \wedge b \wedge \mu_r) Ent(Z|\beta_r \wedge b \wedge [M = i]) \\ &- P(\beta_r \wedge b^c \wedge \mu_r) \sum_{i \in S^r} P(S_i|\beta_r \wedge b^c \wedge \mu_r) Ent(Z|\beta_r \wedge b^c \wedge [M = i]) \end{aligned} \quad (3)$$

where $T(r) = T - \{\beta_r \wedge \alpha_r \wedge \mu_r\}$ is the tree that results from T when the node r is removed; $P(\cdot)$ is the probability (classical data) or weight (probabilistic data) of a node and $Ent(Z|\cdot)$ is the entropy (classical) or fuzzy entropy (probabilistic) (Quinlan (1990)) for Z in the corresponding node.

For classical data and a Boolean assertion (all in (2) and (3), in this case) $a : \Omega \rightarrow \{0, 1\}$, $P(a) := Card(Ext_\Omega(a))/Card(\Omega)$ and $P(S_i|a) := P(S_i|Ext_\Omega(a))$ are estimated in a frequentist way.

For probabilistic data and an assertion $a : \Omega \rightarrow [0, 1]$, $P(a) := \sum_{\omega \in \Omega} a(\omega)$, with $a(\omega)$ the probability of $\omega \in \Omega$ being in the node described by a , and $P(S_i|a) := P([M = i]|a)$ is the relative weight of stratum S_i in the node described by a .

Very briefly, the algorithm's main steps are:

Step 0: Initialisation and evaluation of IC at initial step, $IC\{T, \Omega\}$, that is minus the entropy of Z in the whole population. The first exploitable node contains the whole population (and all strata).

Step 1: Check admissibility condition. For each $r \in X$ (if any), build $B_r \subseteq B$ set of admissible splitting statements to be explored from node r considering information given by NA rules and *number of levels* permitted.

Step 2: Obtain the best split. For each $r \in X$, maximise in $b \in B_r$, the EIC measure, $EIC\{T, r, b, E\}$ of T expanded from node r by split b with respect to E . Maximise in $r \in X$ these measures and select the best node r' and split. This node r' is removed from X (given that we explore it now). Make the split and add to parent node descriptions (in Y_j , M) the description of the new split.

Step 3: Decisional node criterion. For the new children nodes, i.e., the new exploitable nodes, check the set of strata for which the decisional node condition is satisfied. Split from these nodes these strata.

Step 4: Strata terminal node condition. For the new exploitable nodes (if not empty), check the set of strata for which the stopping propagation condition from the node is satisfied. Remove these strata from exploitable nodes, obtaining a terminal node.

Step 5: Update node descriptions. New decisional/terminal node descriptions are added to parent node descriptions in Y_j , descriptions in M with

the strata they contain. For exploitable nodes, update their descriptions in M , removing the strata split in *Steps 3,4*. Descriptions for Z are given by probability distributions. For classical data, probabilities are estimated as relative frequencies of each class in a node. For probabilistic data, they are the relative weight of each class in a node. In the latter case, the weight of an individual into a node is the probability of the individual in the node; and, the weight of a node is the sum of these probabilities. Compute $IC\{T, \Omega\}$, go to *Step 1*.

3 The SDT and SDTEDITOR programs

The programs SDT and SDTEDITOR are written in Visual C++ language (v5.0), Bravo (1999). The SDT identifies design and prediction samples, builds the strata tree, predicts the prediction sample, and writes report, log and graph files. The SDTEDITOR reads the graph file and visualises the strata tree. Input limits to SDT depend on the machine where SDT is executed and on memory management. Minimum SDT requirements are: PC Pentium Windows 95, 133 Mz, 32 Mg. RAM. A Limit in the number of symbolic data objects has been established to be 50,000. Error messages are produced when a lack of memory happens. The SDT has been successfully tested with minimum SDT requirements with 10,000 symbolic data objects and 10 variables and with 5,000 symbolic data objects and 32 variables.

Binary predictors and criterion variable, non-observable values and NA rules are admitted. Extension to non-binary predictors can be done by building variables with binary partitions of categories and adding NA rules.

In the SODAS software, the DB2SO (Stéphan et al.(2000)) builds symbolic data object files from ODBC Data Bases. The user has to give an identification group variable for individuals described by classical data, to be aggregated and described by symbolic data objects. Input of the SDT method requires that at least strata and criterion variable values must be unique in each group. When the identification group variable is the result of a concatenation of several categorical variables, strata and criterion variables should be in this list of variables. The DB2SO makes the identification of NA rules.

4 SDT input parameters

SODAS has a graph user interface to give the predictor, criterion and strata variables and a few parameters, provided with a default value, Morineau (2000). These parameters (or a linear transformation) passed to SDT are: (i) *Maximum percentage of null*. When the percentage of non-observable values in a predictor is lower than this value, then these objects are removed from analysis. Otherwise, the predictor is removed from analysis. Input data objects with non-observable values in strata or criterion variable are removed

for analysis. These latter objects comprise the prediction sample. (ii) *Decisional node condition*. This is the minimum probability for a stratum and a class of Z to be split in a decisional node. (iii) *Strata terminal node condition*. Strata with weight lower than this value in an exploitable node do not follow the recursive method. (iv) *Maximum number of levels in a branch*. Possible values: 1 to 15. (v) *Minimum improvement of IC* in a branch to let recursion continue. This value is compared to $ABS(\frac{IC\{T_{new}, \Omega\} - IC\{T_{last}, \Omega\}}{IC\{T_{last}, \Omega\}})$, with T_{new} the new tree and T_{last} the previous one. If this value is lower then: (1) algorithm Steps 2,3,4 are undone (we go back to T_{last}), and; (2) the explored parent node r' is split into two terminal nodes, splitting the subset of strata by their quality of prediction of Z . This latter action is also taken when the maximum number of levels is attained. (vi) *Printout of a short report*. (vii) *Minimum (relative decisional) node weight in a stratum description* to be written in the report file to describe a stratum.

5 SDTEDITOR

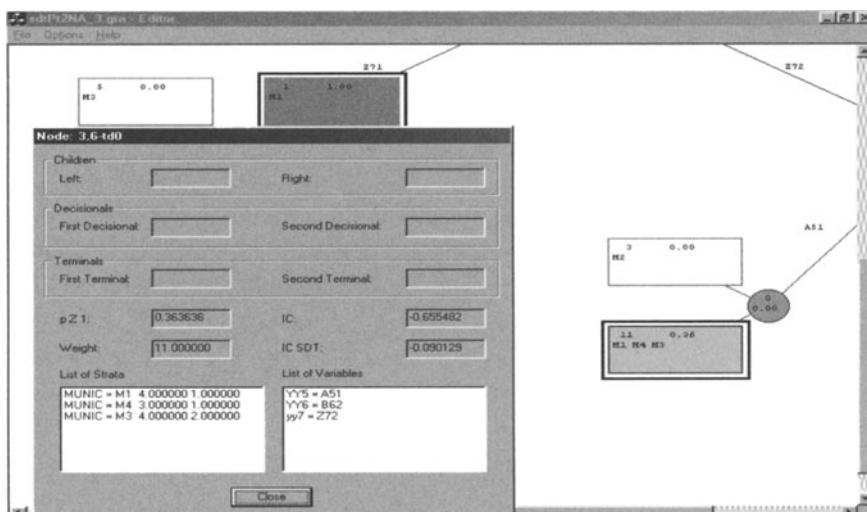


Fig. 1. A 3-level strata tree with node information window

The program SDTEDITOR is a graph editor for strata trees. Some characteristics are: different shapes and colors depending on node types; circles for exploratory nodes and squares for decisional/terminal nodes; different color intensities related to class of Z ; mark nodes with a specified stratum, Figure 1; fit of the strata tree in one window, Figure 2. Information on: split

predictors and categories; data on weight and Z class probabilities; data of stratum identifications in decisional/terminal nodes; and deeper node information in a different window, Figure 1. In the left upper corner, initial and final IC measures, and color and color intensity information are shown, Bravo (1999). The SDTEDITOR user interface, through menus, text boxes,

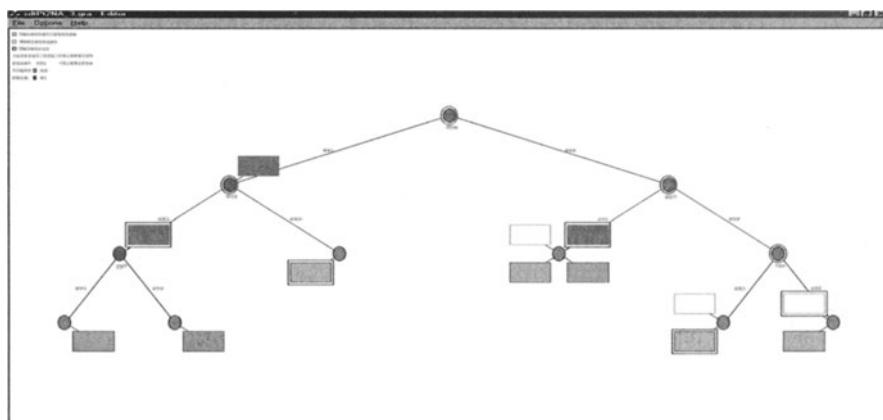


Fig. 2. A 3-level strata tree fitted in one window

toggles, and buttons, let the user change some SDTEDITOR parameters: maximum number of strata/characters per stratum and number of characters per line shown in a decisional/terminal node; minimum weight threshold to show decisional/terminal nodes; font types and sizes; printing capabilities; and possibility of showing or not showing the terminal/decisional nodes and data in all nodes.

References

- BRAVO,M.C.(1999): *Software User Manual for the Strata Decision Tree CSCI v04*. SODAS Project (20821 DG34/D-3/300536). Comm. of the EC-DgIII-Eurostat.
- BRAVO,M.C. and GARCÍA-SANTESMASES,J.M.(1997): Segmentation Trees for Stratified Data. In:Jansen,J. and Lauro,C.N.Ed: *Applied Stochastic Models and Data Analysis: The Ins/Outs of Solving Real Problems*. Curto, Napoli, 37-42.
- (2000a): Segmentation Trees for Stratified Data. In: Bock,H.H. and Diday,E. Eds.: *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heid., 266-293.
- (2000b), Symbolic Object Description of Strata by Segmentation Trees. In: *Computational Statistics*. Physica Verlag, Heidelberg, 12p. To appear.
- BREIMAN,L., FRIEDMAN,J.H., OLSHEN,R.A., STONE,C.J. *Classification and Regression Trees*. Wadsworth, Belmond, Ca.

- CIAMPI,A., DIDAY,E., LEBBE,J., PÉRINEL,E. and VIGNES,R.(1996): Recursive partition with probabilistically imprecise data. In: Diday,E. et al. Eds.: *Ordinal and symbolic data analysis*. Springer Verlag, 201–212.
- DIDAY,E.(1993): *An introduction to Symbolic Data Analysis*, Tutorial of the 4th conference of IFCS, Rapport INRIA no. 1936, Paris.
- MORINEAU, A.(2000): The SODAS Software Package. In: Bock,H.H. and Diday,E. Eds.: *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heid., 386-391.
- QUINLAN,J.R.(1990), Probabilistic Decision Trees, In: Kodratoff,Y., Michalski,R. Ed. *Machine Learning, an Artificial Intelligence A., III*. Kaufmann, 140-152.
- STÉPHAN, V., HÉBRAIL, G. and LECHEVALIER, Y. (2000): Generation of Symbolic Objects from Relational Data Bases. In: Bock,H.H. and Diday,E. Eds.: *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer Verlag, Heidelberg, 78-105.

Marking and Generalization by Symbolic Objects in the Symbolic Official Data Analysis Software

Mireille Gettler Summa

Lise Ceremade, Université Paris IX Dauphine
1, Place du M^e de Lattre de Tassigny 75016 Paris France
(e-mail:summa@ceremade.dauphine.fr)

Abstract. In this paper we propose an automatic method of generating Symbolic Objects in the following framework: description of a partition by symbolic objects that takes into account two aspects, that may be called homogeneity and discrimination criteria. This method belongs to a family of algorithms named MGS (Marking and Generalization by Symbolic Objects), which may be applied either to Factorial Analysis interpretation, to interpretation of partitions or for summarizing huge databases.

1 Introduction

We restrict attention in this paper to a theoretical presentation of the MGS algorithm implemented for SODAS (Symbolic Official Data Analysis Software of EUROSTAT) which builds Discriminant Symbolic Descriptions on a subset of observations. It will thus allow the application of symbolic methods DIDAY (1995) starting from any classical statistical data array, because Markings can be considered as data abstracts with a symbolic object format. This method belongs to exploratory contexts and is not concerned with decision phases (MICHALSKI, 1983). We consider here the restricted case of an input data array which contains only classical nominal data. Outputs are then modal and multi-valued symbolic objects.

In the proposed method, single-valued boolean objects are built at first exploring the level graph, by conjunction of some input levels, according to various chosen criteria. This step builds what are called the Cores of the cluster Markings. Modal and multi-valued symbolic objects are then computed by describing the extensions of the Marking Cores on the variables which do not appear in the results of the first step. The output of the marking process consists of a symbolic matrix which has exactly the same variables as the classical nominal input ones, but in a modal and multi-valued form. The number of rows is of course smaller than that of the initial one, as the scope consists of summarizing raw data.

2 The input classical data matrix

2.1 n observations on p nominal variables

We consider a set $\Omega = \{1, \dots, n\}$ of n objects observed for p classical nominal variables Y_j . \mathcal{Y}_j is the domain of the variable Y_j , that is a finite set of categories. If there are some quantitative variables, they should be transformed into nominal ones. If no expert bounds for the intervals exist, an automatic coding is necessary. As all Marking approaches take into account a discrimination criterion, the quality of the Markings depends on the chosen bounds for the intervals which will define the categories of the new nominal variables. The generalized Fisher algorithm (1958) provides a solution to the problem of finding optimized intervals for Y_j . The output of this process is a categorical single-value matrix.

2.2 The partition of the initial data

Partition resulting from an automatic clustering process. Let us assume that $\Omega = \{1, \dots, n\}$ is partitioned into k_c known disjoint classes C_1, \dots, C_{k_c} , resulting from an automatic clustering process. Each element of the population is described by the p categories of the pattern matrix, and C , the class variable.

Externally provided partition. Let us assume that $\Omega = \{1, \dots, n\}$ is partitioned into r known disjoint classes C_1, \dots, C_r , resulting from an external specification, without any reference to an automatic clustering procedure, such as a randomly generated partition, the partition provided by an expert, or the level-induced partition from a categorical variable. Let C be the partition variable. In this case, classes are likely to be less homogeneous or/and less isolated than those provided by a clustering process which optimize criteria based on variances.

As the quality of the discrimination of the Markings depends on the extent to which the classes are separated, it is necessary to look for optimized levels of the partition variable, as an appropriate aggregation of the initial levels. Marking end users insist upon the importance of this phase which can be managed through Datawarehouse tools.

Let $L = \{l_1, \dots, l_r\}$ be the initial set of levels ; let $\mathcal{P}'(L)$ be the set of possible partitions of L with at least 2 classes; let $P'_s(L)$ be one of those partitions

$$\begin{cases} k \equiv \text{Card}[P'_s(L), P'_s(L) \in \mathcal{P}'(L)] \\ 2 \leq k \leq l \end{cases}$$

Let C_j^s be the subset of Ω induced by the j^{th} class of $P'_s(L)$. One is looking for a partition $P'_s(L)$ such that the Within-class Variance, $\sum_{1 \leq j \leq k} \text{Var}(C_j^s)$, is minimized according to a chosen metric.

For real industrial, marketing or census data, experts may provide a taxonomy of the initial levels. That is to say that C is a tree-structured variable. The search for the best partition of the initial levels must consequently be carried out with the constraint of the knowledge of the taxonomy. The search is thus shortened because only partitions which can be derived from the taxonomy are to be examined for the Within-class Variance optimization.

3 Marking and Generalization by Symbolic descriptions (MGS) algorithm in SODAS

3.1 Marking cores

Let us assume C_1 is the class to be marked, defined from one level of C . Marking cores are generally formalized by multi-valued Boolean categorical symbolic descriptions which do not necessarily contain the same variables. For SODAS, only the restricted situation of single valued Boolean descriptions is developed. Let M_g be a generic marking core for C_1 .

Three criteria for constructing the marking cores. One is looking for conjunctions of initial levels (each conjunction will be a marking core) such that: (i) the cardinal of the union of their extensions in C_1 is maximal; (ii) the cardinal of the union of their extension in $\Omega \setminus C_1$ is minimal; (iii) each hypercube representing conjunctions of levels is statistically significant with respect to the test which has been chosen for measuring the quality of the linkage between C_1 and markings.

According to Gordon (1999), (i) and (ii) can be written as (i) Minimize false negatives, i.e. objects belonging to the extension of a marking but not to C_1 ; (ii) Minimize false positives, i.e. objects belonging to C_1 but not to the union of the extensions of markings. For criterion (iii), in SODAS the Test-Value of Morineau is computed. It is used for ranking the Markings with respect to the quality of the link they have with the class to be marked. The greater it is, the more relevant is the corresponding marking to the class.

The algorithmic approach. Various heuristics have already been proposed to construct Marking cores. Main differences are whether they are top-down, Gettler-Summa (1994), or bottom-up Pham Ti Tong et al. (1996), greedy or not, depth first or breadth first, allowing overlapping branches or not, etc. Let $L = \{l_g, 1 \leq g \leq v\}$ denote the set of the levels of all the variables (except C), S_M a set of Markings, $Cov(l_g) \equiv Card[\Omega \setminus C_1(l_g)]$ (first criterion), and $Err(l_g) = Card[ext_{\subset_{\Omega} \{C_1\}}(l_g)]$ (second criterion). Two a priori thresholds are to be chosen:

- The final degree in which C_1 is covered by the union of the markings;
- The errors made by the markings by covering elements out of C_1 .

Let r_{Cov} denote the first threshold; the ratio for covering obtained by the final markings should be such that:

$$\frac{\text{ext} \left[\bigcup_g M_g \right]}{\text{Card}(C_1)} \geq r_{Cov}.$$

Let r_{Err} denote the second threshold; the error ratio for a marking should be such that:

$$\forall M_g \in S_M \quad \frac{\text{Err}(M_g)}{\text{ext}_\Omega(M_g)} \leq r_{Err}.$$

STEP 1

- Levels are ordered by their measures in the framework of criterion (iii). Let $T(l_g)$ denote this value for a generic level l_g .
- All first levels build a first set of marking cores, which will eventually be improved at further steps; l_g can thus be denoted as M_g^1 . For each marking $Cov(M_g^1)$ and $\text{Err}(M_g^1)$ are computed.
- The two following quantities are also computed:

$$Cov(S_M^1) \equiv \text{Card} \left\{ \text{ext}_{C_1} \left(\bigcup_{S_M^1} M_g^1 \right) \right\}$$

and

$$\text{Err}(S_M^1) \equiv \text{Card} \left\{ \text{ext}_{\Omega \setminus C_1} \left(\bigcup_{S_M^1} M_g^1 \right) \right\}.$$

STEP 2

- Each element of S_M^1 will be a root for descending branches built as follows.
- The constituents of S_M^1 are ordered by their corresponding values $\{T(M_g^1), 1 \leq g \leq v_1\}$ (third criterion).
- The greatest value corresponds to the root which is processed at first and so on.
- Branches are constructed from each node by choosing the levels with the above defined order.
- For each branch, one has to check if it has not yet been constructed for avoiding redundancy.

Each branch as a whole is a new marking. A second set of marking cores S_M^2 is thus substituted to the first one. For each marking $T(M_g^2)$ (criterion (iii)) is computed; for S_M^2 , $Cov(S_M^2)$, $\text{Err}(S_M^2)$ are also computed.

FURTHER STEPS

Step 2 is iterated and stops according to the stopping rules which are described in the following paragraph.

3.2 Stopping and non-stopping rules

As the number of starting levels is limited and redundancy of branches is avoided, the algorithm naturally proceeds with a finite number of steps and gets to an end. Some stopping rules can shorten the process:

- A step f is the last one if $Cov\left(S_M^f\right) / Card_{C_1}\left(S_M^f\right) \geq r_{Cov}$, i.e. C_1 has been sufficiently marked.
- If one does not want more than h levels in a description (for example for providing efficient graphics in an application) no branches will be developed after the h^{th} step which is at the most the last one.
- If a final marking M_f is such that $Err\left(M_f\right) / Card_{C_1}\left(M_f\right) \geq r_{Err}$, it can be cancelled, as an option of the algorithm, from the results.

The markings which are the results of the above process are the so called marking cores for the class C_1 . They are Boolean descriptions, such that each mentioned level has a hundred per cent presence in the description.

3.3 From marking cores to full specified symbolic description

Each Marking core has some missing initial variables ; it does not mean that their values are unknown or that the range is the whole domain. In the framework of SODAS, Symbolic Object Language requires that marking cores were completed on the not specified variables. Results are thus formalized as modal multivalued symbolic descriptions. Actually, the choice which has been implemented in SODAS consists in substituting a missing value by the (discrete) distribution of the missing category on the extension of a marking core in class C_1 .

3.4 Performance of the algorithm

Some work has been carried out on the performance of MGS algorithm by comparing different indexes for measuring the quality of the link between markings and class C_1 . Experimentation has been done on four different data sets coming from the UCI Machine Learning Repository site (<ftp://ics.uci.edu/pub>): WINE, VOTE, WAVE, ZOO. Data have been processed with four different indexes for criterion (iii): the Test-value based on the hypergeometric statistic, the Shannon entropy, the J-measure, and the χ^2 . Each data set Ω is randomly divided into 10 mutually exclusive subsets $\Omega, \Omega_2, \dots, \Omega_{10}$ of approximatively equal size. Each measure is tested 10 times. At each time, marking cores are constructed on $\Omega \setminus \Omega_k$. The following table presents the average of the error rate of misclassifications:

	Test-value	Shannon entropy	J-measure	χ^2
WINE	0.05	0.08	0.08	0.09
VOTE	0.06	0.08	0.6	0.14
WAVE	0.26	0.18	0.20	0.35
ZOO	0.50	0.47	0.23	0.40

4 Applications for official statistical institutes

An extract of INE (National Institute of Statistics of Portugal) Labor Force Survey have been processed by SODAS implementation of MGS (named DSD): 2193 units and 34 categories such that "look for job" had three levels, "full-part" nine, "principal activity" twelve, "month" six, "sex" three, etc. An external partition has been given, in two classes: employed, unemployed.

Here is an example of a Marking core which has been obtained as a result through MGS procedure:

```
princact ∈ {pers_serv&sic}) "and"(prinprof ∈ {pers_serv&sic}) "and"
(fullpart ∈ {full}) "and"(act ∈ {yes}) "and"(inscr ∈ {no}) "and"(lookforjob ∈ {no}) "and"(bestway ∈ {nr}) "and"(months ∈ {NA}) "and"(typesec ∈ {NA}).
```

This Marking covers 9.6% of the respondents of the class "employed", with no error at all; its Test-Value is equal to 16.91.

Variable 'Status' which is missing in the Marking Core is completed as follows: (Status ∈ {married(80%), single(20%)}) .

The full associated completed markings have then be processed as inputs for a Symbolic Discriminant Factorial Analysis through SODAS.

5 Concluding remarks

Marking and Generalisation by Symbolic Objects implemented as the Discriminant Symbolic Descriptions procedure in SODAS provides good abstracts on data which do not have more than a hundred levels from the active variables and ten thousand individuals to be summarised. For bigger data sets, a new version is being developed which is faster thanks to the possibility of taking into account taxonomies and making unions and intersections on the Marking cores. Some new validation tools will also be available based on Symbolic constraints.

References

- DIDAY, E. (1995) : Probabilistic objects for a symbolic data analysis, *Series in Discrete Mathematics and Theoretical Computers*, 19.
- GETTLER-SUMMA, M., PERINEL, E. and FERRARIS, J. (1994) : New automatic aid to symbolic cluster interpretation. In *New Approaches in Classification and Data Analysis*, Springer Ed.
- GORDON, A.G. (1999) : *Classification* . (2nd Edition), Chapman&Hall/CRC, Boca Raton,FL. 256pp.
- PHAM TI TONG , H. and GETTLER-SUMMA, M., (1996) : IFCS 96, Kobe,Japan.
- MICHALSKI, R.S., (1983) : STEPP.-Machine Learning: an artificial intelligence approach, chap. Learning from observations:conceptual clustering, pp.331-363. Morgan Kaufmann.

List of Reviewers

J.P Asselin de Beauville	H. Bacelaar-Nicolau	D. Banks
M. Bardos	J.P. Barthélémy	V. Batagelj
F. Bavaud	H. Bensmail	H. Berkhof
V. Bertholet	T. Bijmolt	L. Billard
C. Bravo	I. Brito	H.H. Bock
C. Borgelt	A. Carlier	A. Ciampi
S. Contassot-Vivier	R. Coppi	R. Couturier
C. Croux	G. Cucumel	P. de Boeck
C. Dehon	M. de Rooij	E. Diday
G.B. Dijksterhuis	E. Dusseldorp	B. Efron
P.H.C. Eilers	A.S. Ferreira	B. Fichet
P. Filzmoser	P.H. Franses	A. Gordon
J.C. Gower	A. Guenoche	D.J. Hand
A. Hardy	W.J. Heiser	C. Hennig
S. Hirtle	M. Hubert	D. Jacquemin
K. Jajuga	G.G.H. Jansen	K.C. Klauer
T. Kodawaki	P.M. Kroonenberg	F.J. Lapointe
C.N. Lauro	L. Lebart	Y. Lechevallier
P. Legendre	I.C. Lerman	S. Liu
V. Makarenkov	J.A. Martín Fernández	R. Mathar
J. Meulman	R. Miglio	F. Mola
H.J. Mucha	N. Ohsumi	O. Opitz
V. Petko	W. Polasek	M. Remon
C. Robardet	G. Saporta	R. Siciliano
T.A.B. Snijders	E.W. Steyerberg	M. Summa
M.E. Timmerman	L. Torgo	M. Ueno
S. van Aelst	I. van der Lans	M. van der Meulen
M.A.J. van Duijn	F. van Eeuwijk	I. van Mechelen
K. van Montfort	B.J. van Os	W.H. van Schuur
R. Verde	N. Verhelst	J. Vermunt
M. Vichi	H.Vos	M. Wedel

Index

- Action rule 23
Additive tree 35, 149
Affectation 187
Affinity coefficient 181
Aggregation invariance 131
Alexandrov, M.A. 83
Alpha-trimmed regions 17
Ambroise, C. 161
Analyse des donn'ees 329
ANN 223
Anselmo, C.A.F. 375
Application–oil industry 95
Asymmetric data scatter 235
Atypicality index 187
Autonomous agents 23, 229
Average profile 131
- Bacelar-Nicolau, H.* 107, 181
Backfitting 205
Backward induction 101
Barceló-Vidal, C. 155
Bardos, M. 169
Batagelj, Z. 347
Bavaud, F. 131
Bayesian rule 187
Becker, C. 315
Benmiloud, B. 247
Billard, L. 369
Billaudel, P. 63
Bisdorff, R. 401
Bonnevay, S. 303
Bootstrap 59, 359
Botte-Lecocq, C. 69
Brand management 89
Bravo, M.C. 409
Brazdil, P. 119
Breakdown point 315
Breakdown value 309
Brito, I. 175
- Canonical correlation 321
Cappelli, C. 193
CART 205
Carvalho, F.A.T. 375, 387
- Castillo, W.* 297
Celeux, G. 175, 181
Censoring 223
Centroid method 95
Chamal, H. 187
Chauchat, J.H. 199
Ciampi, A. 223
Ciok, A. 41
Classification probabilities 3
Classification trees 211
Classifier combination 205
CLUSTAN 89
Cluster analysis 23, 41, 47, 125, 155
Cluster interpretation 417
Clustering 63, 113, 161, 217
Collapsibility 341
Color image segmentation 69
Combining models 181
Comparison of surveys 353
Compositional data 155
Conditional linear independence 279
Consensus 125
Context dependent proximity 387
Contingency tables 315
Conversano, C. 205
Convex geometry 241
Convex hull 235
Cooperative social networks 229
Coppi, R. 273
Core cluster detection 247
Correlation 369
Correspondence analysis 41, 329, 335
Costa, J. 119
Costa, J.P. 217
Covariance 369
Covariance matrices 321
Credit risk portfolio analysis 169
Cross-societal survey 335
Cross-validation 175
Croux, C. 321
Cucumel, G. 125
- D'Urso, P.* 273
Data acquisition 329

- Data analysis 303, 329
 Data depth 17
 Data mining 229, 395
 Data quality 335
 Data science 329
De Cantis, S. 341
 Decision tree 199, 409
Dehon, C. 321
 Dependence graph 285
deSoete, G. 59
Devillez, A. 63
Diday, E. 369, 401
 Discretization 199
 Discriminant analysis 95, 169, 175, 241
 Discrimination 181, 187
 Disparity function 267
 Dissimilarities 35, 131
 Domain-oriented dictionary 83
Doria, I.P. 107
 Dynamic programming 101
 Dynamical clustering 387
- EDDA models 175
 EM algorithm 161
 Entropy-based clustering 29
 Expected loss 169
 Experimental web surveys 353
- Factorial analysis 381
 Failure horizon 169
Ferreira, A.S. 181
Fichet, B. 137
Filzmoser, P. 321
 First-order independence model 181
 Fixed point clusters 53
 Fuzzy clustering 77
 Fuzzy logic 63
 Fuzzy morphology 69
 Fuzzy segmentation 69
 Fuzzy time array 273
 Fuzzy time dissimilarity 273
 Fuzzy time trajectory 273
- GAM 205
Gather, U. 315
 General network representation 35
 Generalized power divergence 131
 Genetic algorithms 229
- Gettler Summa, M.* 417
Gillet, A. 69
 Gini index 83
González, J. 297
 Goodness of fit 267, 359
Govaert, G. 161
Grandcolas, S. 143
 Graph partitioning 113
 Graph theory 285
Guénoche, A. 143
- Hébrail, G.* 395
Hajnal, I. 47
Hartigan, J.A. 3
Hayashi, C. 335
Henaux, F. 113
Hennig, C. 53
 Hierarchical clustering 95, 107
 Hierarchical coupling 181
Hoberg, R. 17
 Huygens' principle 131
 Hybrid approaches 217
 Hypervolume criterion 17
- Identification of respondents 353
 Image analysis 375
Imaizumi, T. 77
 Incomplete data 161
 INDSCAL 297
 Induction 3
 Inertia 131
 Influence function 309
 Information content measure 409
 Information matrix 261
 Inhomogeneity 17
 Initial values 47
 Interaction 341
 Internet surveys 347
- Jardino, M.* 29
- K-means 29, 77, 89
Kikuchi, T. 77
Kuhnt, S. 315
 Kullback-Leibler divergence 131
Kuntz, P. 113
- Landry, P.-A.* 149
Lapointe, F.-J. 125, 149
 Large data set 335

- Largeron-Leteno, C.* 303
 Latent class model 59
 Latent trait analysis 41
Lauro, N.C. 381
Le Calvé, G. 107
 Least-squares 35
 Least-squares method 149
Lechevallier, Y. 223, 387, 395
Lecolier, G.V. 63
Legendre, P. 35
Levasseur, C. 149
 Likelihood ratio test 261
 Linear regression 369
 Linear transformation 23
Liu, S. 261
Loosveldt, G. 47
 Lower maximal approximation 137

Macaire, L. 69
Makagonov, P.P. 83
Makarenkov, V. 35
Manfreda, K.L. 347
 Market segmentation 89
 Marking validation 417
 Markovian model 247
Martín-Fernández, J.A. 155
 Masking effect 235
 Matching functions 375
 Maximum likelihood 261
 MDS with restrictions 297
 Metric MDS 297
Miglio, R. 211
Minami, H. 255
 Minimax principle 101
 Minimum covariance determinant 309
 Missing data 149
 Mixture models 47, 53, 161
Mizuta, M. 255
 Mode detection 69
 Mokken Double Monotonicity model 267
Mola, F. 193, 205
 Monte-Carlo process 29
Moussa, A. 247
 Multidimensional scaling 297
 Multinomial model 181
 Multiple regression 217
 Multivariate categorical data 341
 Multivariate graph 285

 Multivariate normal model 315
 Multivariate regression 309

 Network data collection 347
 Neural networks 77, 211
 Non convex pattern recognition 241
 Non-hierarchical clustering 29
 Normal mixtures 53

Ohsumi, N. 329, 353
Oliveri, A.M. 341
 Online surveys 347
 Optimal partition 29
 Outlier identification 315
 Outliers 95

 Pairwise decision 175
Palumbo, F. 381
 Partial correlation 279
 Partial distances 143
Pattison, P. 285
Pawlowsky-Glahn, V. 155
 Pertinence 125
 Phylogenetic tree 149
Pillati, M. 211
 PL and VALAw coefficients 107
 Poisson point process 241
Polasek, W. 261
Porzio, G.C. 235
Postaire, J.-G. 69, 247
 Posterior probability of failure 169
 Prediction 223
 Pretopology 303
 Principal components analysis 95
 Probabilistic similarity index 107
 Projection pursuit 255
 Pseudoclosure 303

 Quantification methods 329, 335
 Quasi-hierarchy 137
 Quasi-ultrametric 137

Rémon, M. 241
 Radial basis functions 211
Ragozini, G. 235
Rakotomalala, R. 199
 Random graph 285
 Random trials 89
 Randomization test 125

- Ranking evaluation 119
 Recovery 47
 Reduction of explanatory variables 255
 Regression trees 193
 Relational databases 395
 Replacement of zeros 155
 Representativeness 353
 Response rates 353
 Reticulogram 35
Rivas Moya, T. 267
 Robust approach 95
 Robust methods 315
 Robustness 309, 321
Rousseeuw, P.J. 309
 Row/column duality 131
 Sample isotone regression 267
 Sample selection 347
 Sample surveys 359
 Sampling 199
Sato, Y. 23
Sbihi, A. 247
Sboychakov, K. 83
Scagni, A. 359
 Segmentation tree 409
 Sequential classification 101
Siciliano, R. 193, 205
 Similarity index on dichotomics 107
 Simpson's paradox 341
 Simulated annealing 297
 Simulation 47, 149
Slaoui Chah, S. 187
 Sliced inverse regression 255, 315
 Smoothing 205
Soares, C. 119
 Social sciences applications 401
 Solicitation methods 347
Souza, R.M.C.R. 375
 Spearman's rho 41
 Spectral methods 113
 Statistical reliability 193
 Structural method 303
 Sufficient covariates 341
 Supervised classification 119, 375
 Supervised classifiers 211
 Survival data 223
 Symbolic data analysis 369, 381, 395, 401, 409, 417
 Symbolic objects 375, 381, 387, 409
 Text mining 83
 Threshold loss 101
Torgo, L. 217
 Tree growing 409
 Tree reconstruction 143
Trejos, J. 297
 Ultrametric matrix 107
 Ultrametric tree 149
Ultsch, A. 229
 Unsolicited/solicited surveys 347
 Unsupervised classification 63, 247
 Unsupervised clustering 29
 Validation 125
 Validity of partition 63
Valois, J.-P. 95
Van Aelst, S. 309
Van Driesssen, K. 309
 VAR-ARCH model 261
Vehovar, V. 347
Verde, R. 381, 387
Vicari, D. 279
Villalobos, M. 297
 Voronoï tessellation 241
Vos, H.J. 101
 Wald test 261
 Ward's method 89, 95
Wasserman, S. 285
Watanabe, N. 77
 Web surveys 347
 White misspecification test 261
Winsberg, S. 59
Wishart, D. 89
Yoshimura, O. 353
 Zipf law 83
 zonoid 17