

Studies in Classification, Data Analysis, and Knowledge Organization

Managing Editors

H.-H. Bock, Aachen
W. Gaul, Karlsruhe
M. Vichi, Rome

Editorial Board

Ph. Arabie, Newark
D. Baier, Cottbus
F. Critchley, Milton Keynes
R. Decker, Bielefeld
E. Diday, Paris
M. Greenacre, Barcelona
C. Lauro, Naples
J. Meulman, Leiden
P. Monari, Bologna
S. Nishisato, Toronto
N. Ohsumi, Tokyo
O. Optiz, Augsburg
G. Ritter, Passau
M. Schader, Mannheim
C. Weihs, Dortmund

Titles in the Series

- 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 (Eds.)
Analysis of Symbolic Data. 2000
- H. A. L. Kiers, J.-P. Rasson, P. J. F.
Groenen, and M. Schader (Eds.)
Data Analysis, Classification, and
Related Methods. 2000
- W. Gaul, O. Opitz, M. Schader (Eds.)
Data Analysis. 2000
- R. Decker and W. Gaul (Eds.)
Classification and Information
Processing at the Turn of the
Millennium. 2000
- S. Borra, R. Rocci, M. Vichi,
and M. Schader (Eds.)
Advances in Classification and Data
Analysis. 2000
- W. Gaul and G. Ritter (Eds.)
Classification, Automation, and New
Media. 2002
- K. Jajuga, A. Sokołowski, and
H.-H. Bock (Eds.)
Classification, Clustering and Data
Analysis. 2002
- M. Schwaiger and O. Opitz (Eds.)
Exploratory Data Analysis in
Empirical Research. 2003
- M. Schader, W. Gaul, and M. Vichi
(Eds.)
Between Data Science and Applied
Data Analysis. 2003
- H.-H. Bock, M. Chiodi, and
A. Mineo (Eds.)
Advances in Multivariate Data
Analysis. 2004
- D. Banks, L. House, F.R. McMorris,
P. Arabie, and W. Gaul (Eds.)
Classification, Clustering, and Data
Mining Applications. 2004
- D. Baier and K.-D. Wernecke (Eds.)
Innovations in Classification, Data
Science, and Information Systems.
2005
- M. Vichi, P. Monari, S. Mignani, and
A. Montanari (Eds.)
New Developments in Classification
and Data Analysis. 2005
- D. Baier, R. Decker, and L. Schmidt-
Thieme (Eds.)
Data Analysis and Decision Support.
2005
- C. Weihs and W. Gaul (Eds.)
Classification - the Ubiquitous
Challenge. 2005
- M. Spiliopoulou, R. Kruse, C.
Borgelt, A. Nürnberger, and W. Gaul
(Eds.)
From Data and Information Analysis
to Knowledge Engineering. 2006
- V. Batagelj, H.-H. Bock, A. Ferligoj,
and A. Žiberna (Eds.)
Data Science and Classification. 2006
- S. Zani, A. Cerioli, M. Riani, M. Vichi
(Eds.)
Data Analysis, Classification and the
Forward Search. 2006

Reinhold Decker
Hans-J. Lenz
Editors

Advances in Data Analysis

Proceedings of the 30th Annual Conference
of the Gesellschaft für Klassifikation e.V.,
Freie Universität Berlin, March 8-10, 2006

With 202 Figures and 92 Tables

 Springer

Professor Dr. Reinhold Decker
Department of Business Administration and Economics
Bielefeld University
Universitätsstr. 25
33501 Bielefeld
Germany
rdecker@wiwi.uni-bielefeld.de

Professor Dr. Hans - J. Lenz
Department of Economics
Freie Universität Berlin
Garystraße 21
14195 Berlin
Germany
hjlenz@wiwiss.fu-berlin.de

Library of Congress Control Number: 2007920573

ISSN 1431-8814

ISBN 978-3-540-70980-0 Springer Berlin Heidelberg New York

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. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2007

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.

Production: LE-TeX Jelonek, Schmidt & Vockler GbR, Leipzig
Cover-design: WMX Design GmbH, Heidelberg

SPIN 12022755 43/3100YL - 5 4 3 2 1 0 Printed on acid-free paper

Preface

This volume contains the revised versions of selected papers presented during the 30th Annual Conference of the German Classification Society (Gesellschaft für Klassifikation – GfKl) on “Advances in Data Analysis”. The conference was held at the Freie Universität Berlin, Germany, in March 2006. The scientific program featured 7 parallel tracks with more than 200 contributed talks in 63 sessions. Additionally, thanks to the support of the DFG (German Research Foundation), 18 plenary and semi-plenary speakers from Europe and overseas could be invited to talk about their current research in classification and data analysis. With 325 participants from 24 countries in Europe and overseas this GfKl Conference, once again, provided an international forum for discussions and mutual exchange of knowledge with colleagues from different fields of interest. From altogether 115 full papers that had been submitted for this volume 77 were finally accepted.

The scientific program included a broad range of topics from classification and data analysis. Interdisciplinary research and the interaction between theory and practice were particularly emphasized. The following sections (with chairs in alphabetical order) were established:

I. Theory and Methods

Clustering and Classification (H.-H. Bock and T. Imaizumi); Exploratory Data Analysis and Data Mining (M. Meyer and M. Schwaiger); Pattern Recognition and Discrimination (G. Ritter); Visualization and Scaling Methods (P. Groenen and A. Okada); Bayesian, Neural, and Fuzzy Clustering (R. Kruse and A. Ultsch); Graphs, Trees, and Hierarchies (E. Godehardt and J. Hansohm); Evaluation of Clustering Algorithms and Data Structures (C. Hennig); Data Analysis and Time Series Analysis (S. Lang); Data Cleaning and Pre-Processing (H.-J. Lenz); Text and Web Mining (A. Nürnberger and M. Spiliopoulou); Personalization and Intelligent Agents (A. Geyer-Schulz); Tools for Intelligent Data Analysis (M. Hahsler and K. Hornik).

II. Applications

Subject Indexing and Library Science (H.-J. Hermes and B. Lorenz); Marketing, Management Science, and OR (D. Baier and O. Opitz); E-commerce, Rec-

ommender Systems, and Business Intelligence (L. Schmidt-Thieme); Banking and Finance (K. Jajuga and H. Locarek-Junge); Economics (G. Kauermann and W. Polasek); Biostatistics and Bioinformatics (B. Lausen and U. Mansmann); Genome and DNA Analysis (A. Schliep); Medical and Health Sciences (K.-D. Wernecke and S. Willich); Archaeology (I. Herzog, T. Kerig, and A. Posluschny); Statistical Musicology (C. Weihs); Image and Signal Processing (J. Buhmann); Linguistics (H. Goebl and P. Grzybek); Psychology (S. Krolak-Schwerdt); Technology and Production (M. Feldmann).

Additionally, the following invited sessions were organized by colleagues from associated societies: Classification with Complex Data Structures (A. Cerioli); Machine Learning (D.A. Zighed); Classification and Dimensionality Reduction (M. Vichi).

The editors would like to emphatically thank the section chairs for doing such a great job regarding the organization of their sections and the associated paper reviews. The same applies to W. Esswein for organizing the Doctoral Workshop and to H.-H. Hermes and B. Lorenz for organizing the Librarians Workshop. Cordial thanks also go to the members of the scientific program committee for their conceptual and practical support (in alphabetical order): D. Baier (Cottbus), H.-H. Bock (Aachen), H.W. Brachinger (Fribourg), R. Decker (Bielefeld, Chair), D. Dubois (Toulouse), A. Gammernan (London), W. Gaul (Karlsruhe), A. Geyer-Schulz (Karlsruhe), B. Goldfarb (Paris), P. Groenen (Rotterdam), D. Hand (London), T. Imaizumi (Tokyo), K. Jajuga (Wroclaw), G. Kauermann (Bielefeld), R. Kruse (Magdeburg), S. Lang (Innsbruck), B. Lausen (Erlangen-Nürnberg), H.-J. Lenz (Berlin), F. Murtagh (London), A. Okada (Tokyo), L. Schmidt-Thieme (Hildesheim), M. Spiliopoulou (Magdeburg), W. Stütze (Washington), and C. Weihs (Dortmund). The review process was additionally supported by the following colleagues: A. Cerioli, E. Gatnar, T. Kneib, V. Köppen, M. Meißner, I. Michalarias, F. Mörchen, W. Steiner, and M. Walesiak.

The great success of this conference would not have been possible without the support of many people mainly working in the backstage. Representative for the whole team we would like to particularly thank M. Darkow (Bielefeld) and A. Wnuk (Berlin) for their exceptional efforts and great commitment with respect to the preparation, organization and post-processing of the conference. We thank very much our web masters I. Michalarias (Berlin) and A. Omelchenko (Berlin). Furthermore, we would cordially thank V. Köppen (Berlin) and M. Meißner (Bielefeld) for providing an excellent support regarding the management of the reviewing process and the final editing of the papers printed in this volume.

The GfKI Conference 2006 would not have been possible in the way it took place without the financial and/or material support of the following institutions and companies (in alphabetical order): Deutsche Forschungsgemeinschaft, Freie Universität Berlin, Gesellschaft für Klassifikation e.V., Land Software-Entwicklung, Microsoft München, SAS Deutschland, Springer-

Verlag, SPSS München, Universität Bielefeld, and Westfälisch-Lippische Universitätsgesellschaft. We express our gratitude to all of them.

Finally, we would like to thank Dr. Martina Bihn of Springer-Verlag, Heidelberg, for her support and dedication to the production of this volume.

Berlin and Bielefeld, January 2007

Hans-J. Lenz
Reinhold Decker

Contents

Part I Clustering

Mixture Models for Classification

Gilles Celeux 3

How to Choose the Number of Clusters: The Cramer Multiplicity Solution

Adriana Climescu-Haulica..... 15

Model Selection Criteria for Model-Based Clustering of Categorical Time Series Data: A Monte Carlo Study

José G. Dias 23

Cluster Quality Indexes for Symbolic Classification – An Examination

Andrzej Dudek..... 31

Semi-Supervised Clustering: Application to Image Segmentation

Mário A.T. Figueiredo..... 39

A Method for Analyzing the Asymptotic Behavior of the Walk Process in Restricted Random Walk Cluster Algorithm

Markus Franke, Andreas Geyer-Schulz 51

Cluster and Select Approach to Classifier Fusion

Eugeniusz Gatnar..... 59

Random Intersection Graphs and Classification

Erhard Godehardt, Jerzy Jaworski, Katarzyna Rybarczyk 67

Optimized Alignment and Visualization of Clustering Results

Martin Hoffmann, Dörte Radke, Ulrich Möller..... 75

Finding Cliques in Directed Weighted Graphs Using Complex Hermitian Adjacency Matrices <i>Bettina Hoser, Thomas Bierhance</i>	83
Text Clustering with String Kernels in R <i>Alexandros Karatzoglou, Ingo Feinerer</i>	91
Automatic Classification of Functional Data with Extremal Information <i>Fabrizio Laurini, Andrea Cerioli</i>	99
Typicality Degrees and Fuzzy Prototypes for Clustering <i>Marie-Jeanne Lesot, Rudolf Kruse</i>	107
On Validation of Hierarchical Clustering <i>Hans-Joachim Mucha</i>	115
<hr/>	
Part II Classification	
<hr/>	
Rearranging Classified Items in Hierarchies Using Categorization Uncertainty <i>Korinna Bade, Andreas Nürnberger</i>	125
Localized Linear Discriminant Analysis <i>Irina Czogiel, Karsten Luebke, Marc Zentgraf, Claus Weihs</i>	133
Calibrating Classifier Scores into Probabilities <i>Martin Gebel, Claus Weihs</i>	141
Nonlinear Support Vector Machines Through Iterative Majorization and I-Splines <i>Patrick J.F. Groenen, Georgi Nalbantov, J. Cor Bioch</i>	149
Deriving Consensus Rankings from Benchmarking Experiments <i>Kurt Hornik, David Meyer</i>	163
Classification of Contradiction Patterns <i>Heiko Müller, Ulf Leser, Johann-Christoph Freytag</i>	171
Selecting SVM Kernels and Input Variable Subsets in Credit Scoring Models <i>Klaus B. Schebesch, Ralf Stecking</i>	179

Part III Data and Time Series Analysis

Simultaneous Selection of Variables and Smoothing Parameters in Geoadditive Regression Models
Christiane Belitz, Stefan Lang 189

Modelling and Analysing Interval Data
Paula Brito 197

Testing for Genuine Multimodality in Finite Mixture Models: Application to Linear Regression Models
Bettina Grün, Friedrich Leisch 209

Happy Birthday to You, Mr. Wilcoxon! Invariance, Semiparametric Efficiency, and Ranks
Marc Hallin 217

Equivalent Number of Degrees of Freedom for Neural Networks
Salvatore Ingrassia, Isabella Morlini 229

Model Choice for Panel Spatial Models: Crime Modeling in Japan
Kazuhiko Kakamu, Wolfgang Polasek, Hajime Wago..... 237

A Boosting Approach to Generalized Monotonic Regression
Florian Leitenstorfer, Gerhard Tutz 245

From Eigenspots to Fisherspots – Latent Spaces in the Nonlinear Detection of Spot Patterns in a Highly Varying Background
Bjoern H. Menze, B. Michael Kelm, Fred A. Hamprecht 255

Identifying and Exploiting Ultrametricity
Fionn Murtagh 263

Factor Analysis for Extraction of Structural Components and Prediction in Time Series
Carsten Schneider, Gerhard Arminger 273

Classification of the U.S. Business Cycle by Dynamic Linear Discriminant Analysis
Roland Schuhr 281

Examination of Several Results of Different Cluster Analyses with a Separate View to Balancing the Economic and Ecological Performance Potential of Towns and Cities
Nguyen Xuan Thinh, Martin Behnisch, Alfred Ultsch 289

Part IV Visualization and Scaling Methods

VOS: A New Method for Visualizing Similarities Between Objects
Nees Jan van Eck, Ludo Waltman 299

Multidimensional Scaling of Asymmetric Proximities with a Dominance Point
Akinori Okada, Tadashi Imaizumi..... 307

Single Cluster Visualization to Optimize Air Traffic Management
Frank Rehm, Frank Klawonn, Rudolf Kruse 319

Rescaling Proximity Matrix Using Entropy Analyzed by INDSCAL
Satoru Yokoyama, Akinori Okada 327

Part V Information Retrieval, Data and Web Mining

Canonical Forms for Frequent Graph Mining
Christian Borgelt 337

Applying Clickstream Data Mining to Real-Time Web Crawler Detection and Containment Using ClickTips Platform
Anália Lourenço, Orlando Belo 351

Plagiarism Detection Without Reference Collections
Sven Meyer zu Eissen, Benno Stein, Marion Kulig..... 359

Putting Successor Variety Stemming to Work
Benno Stein, Martin Potthast 367

Collaborative Filtering Based on User Trends
Panagiotis Symeonidis, Alexandros Nanopoulos, Apostolos Papadopoulos, Yannis Manolopoulos 375

Investigating Unstructured Texts with Latent Semantic Analysis
Fridolin Wild, Christina Stahl 383

Part VI Marketing, Management Science and Economics

Heterogeneity in Preferences for Odd Prices <i>Bernhard Baumgartner, Winfried J. Steiner</i>	393
Classification of Reference Models <i>Robert Braun, Werner Esswein</i>	401
Adaptive Conjoint Analysis for Pricing Music Downloads <i>Christoph Breidert, Michael Hahsler</i>	409
Improving the Probabilistic Modeling of Market Basket Data <i>Christian Buchta</i>	417
Classification in Marketing Research by Means of LEM2-generated Rules <i>Reinhold Decker, Frank Kroll</i>	425
Pricing Energy in a Multi-Utility Market <i>Markus Franke, Andreas Kamper, Anke Eßer</i>	433
Disproportionate Samples in Hierarchical Bayes CBC Analysis <i>Sebastian Fuchs, Manfred Schwaiger</i>	441
Building on the Arules Infrastructure for Analyzing Transaction Data with R <i>Michael Hahsler, Kurt Hornik</i>	449
Balanced Scorecard Simulator – A Tool for Stochastic Business Figures <i>Veit Köppen, Marina Allgeier, Hans-J. Lenz</i>	457
Integration of Customer Value into Revenue Management <i>Tobias von Martens, Andreas Hilbert</i>	465
Women’s Occupational Mobility and Segregation in the Labour Market: Asymmetric Multidimensional Scaling <i>Miki Nakai</i>	473
Multilevel Dimensions of Consumer Relationships in the Healthcare Service Market M-L IRT vs. M-L SEM Approach <i>Iga Rudawska, Adam Sagan</i>	481

Data Mining in Higher Education

Karoline Schönbrunn, Andreas Hilbert 489

Attribute Aware Anonymous Recommender Systems

Manuel Stritt, Karen H.L. Tso, Lars Schmidt-Thieme 497

Part VII Banking and Finance

On the Notions and Properties of Risk and Risk Aversion in the Time Optimal Approach to Decision Making

Martin Bouzaima, Thomas Burkhardt 507

A Model of Rational Choice Among Distributions of Goal Reaching Times

Thomas Burkhardt 515

On Goal Reaching Time Distributions Estimated from DAX Stock Index Investments

Thomas Burkhardt, Michael Haasis 523

Credit Risk of Collaterals: Examining the Systematic Linkage between Insolvencies and Physical Assets in Germany

Marc Gürtler, Dirk Heithecker, Sven Olboeter 531

Foreign Exchange Trading with Support Vector Machines

Christian Ullrich, Detlef Seese, Stephan Chalup 539

The Influence of Specific Information on the Credit Risk Level

Miroslaw Wójciak, Aleksandra Wójcicka-Krenz 547

Part VIII Bio- and Health Sciences

Enhancing Bluejay with Scalability, Genome Comparison and Microarray Visualization

Anguo Dong, Andrei L. Turinsky, Andrew C. Ah-Seng, Morgan Taschuk, Paul M.K. Gordon, Katharina Hochauer, Sabrina Fröls, Jung Soh, Christoph W. Sensen 557

Discovering Biomarkers for Myocardial Infarction from SELDI-TOF Spectra

Christian Höner zu Siederdissen, Susanne Ragg, Sven Rahmann 569

Joint Analysis of In-situ Hybridization and Gene Expression Data

Lennart Opitz, Alexander Schliep, Stefan Posch 577

Unsupervised Decision Trees Structured by Gene Ontology (GO-UDTs) for the Interpretation of Microarray Data <i>Henning Redestig, Florian Sohler, Ralf Zimmer, Joachim Selbig</i>	585
--	-----

Part IX Linguistics and Text Analysis

Clustering of Polysemic Words <i>Laurent Cicurel, Stephan Bloehdorn, Philipp Cimiano</i>	595
--	-----

Classifying German Questions According to Ontology-Based Answer Types <i>Adriana Davidescu, Andrea Heyl, Stefan Kazalski, Irene Cramer, Dietrich Klakow</i>	603
---	-----

The Relationship of Word Length and Sentence Length: The Inter-Textual Perspective <i>Peter Grzybek, Ernst Stadlober, Emmerich Kelih</i>	611
--	-----

Comparing the Stability of Different Clustering Results of Dialect Data <i>Edgar Haimler, Hans-Joachim Mucha</i>	619
--	-----

Part-of-Speech Discovery by Clustering Contextual Features <i>Reinhard Rapp</i>	627
---	-----

Part X Statistical Musicology and Sound Classification

A Probabilistic Framework for Audio-Based Tonal Key and Chord Recognition <i>Benoit Catteau, Jean-Pierre Martens, Marc Leman</i>	637
--	-----

Using MCMC as a Stochastic Optimization Procedure for Monophonic and Polyphonic Sound <i>Katrin Sommer, Claus Weihs</i>	645
---	-----

Vowel Classification by a Neurophysiologically Parameterized Auditory Model <i>Gero Szepannek, Tamás Harczos, Frank Klefenz, András Katai, Patrick Schikowski, Claus Weihs</i>	653
--	-----

Part XI Archaeology

**Uncovering the Internal Structure of the Roman Brick and
Tile Making in Frankfurt-Nied by Cluster Validation**
Jens Dolata, Hans-Joachim Mucha, Hans-Georg Bartel 663

**Where Did I See You Before...
A Holistic Method to Compare and Find Archaeological
Artifacts**
Vincent Mom 671

Keywords 681

Author Index 685

Equivalent Number of Degrees of Freedom for Neural Networks

Salvatore Ingrassia¹ and Isabella Morlini²

¹ Dipartimento di Economia e Metodi Quantitativi, Università di Catania, Corso Italia 55, 95128 Catania, Italy; s.ingrassia@unict.it

² Dip. di Scienze Sociali Cognitive e Quantitative, Università di Modena e Reggio E., Via Allegri 9, 42100 Reggio Emilia, Italy; morlini.isabella@unimore.it

Abstract. The notion of equivalent number of degrees of freedom (e.d.f.) to be used in neural network modeling from small datasets has been introduced in Ingrassia and Morlini (2005). It is much smaller than the total number of parameters and it does not depend on the number of input variables. We generalize our previous results and discuss the use of the e.d.f. in the general framework of multivariate nonparametric model selection. Through numerical simulations, we also investigate the behavior of model selection criteria like AIC, GCV and BIC/SBC, when the e.d.f. is used instead of the total number of the adaptive parameters in the model.

1 Introduction

This article presents the results of some empirical studies comparing different model selection criteria, like AIC, GCV and BIC (see, among others, Kadane and Lazar (2004), for nonlinear projection models, based on the *equivalent number of degrees of freedoms* (e.d.f) introduced in Ingrassia and Morlini (2005). Given a response variable Y and predictor variables $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^m$, throughout this paper we assume that the input-output relation can be written as $Y = \phi(\mathbf{x}) + \varepsilon$, where Y assumes values in $\mathcal{Y} \subseteq \mathbb{R}$ and ε is a random variable with zero mean and finite variance. We then assume that the unknown functional dependency $\phi(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$ is of the form:

$$f_p(\mathbf{x}) = \sum_{i=1}^p c_i \tau(\mathbf{a}'_i \mathbf{x} + b_i) + c_{p+1} \quad (1)$$

where $\mathbf{a}_1, \dots, \mathbf{a}_p \in \mathbb{R}^m$, $b_1, \dots, b_p, c_{p+1}, c_1, \dots, c_p \in \mathbb{R}$ and τ is a sigmoidal function. In the following, without loss of generality, we will assume $c_{p+1} = 0$. Indeed, the expression (1) may be written in the form: $f_p(\mathbf{x}) = \sum_{i=1}^{p+1} c_i \tau(\mathbf{a}'_i \mathbf{x} + b_i)$ where the constant term c_{p+1} has been included in the summation and $\tau(\mathbf{a}'_{p+1} \mathbf{x} + b_{p+1}) \equiv 1$. Therefore, results presented in this article may be

extended to the case $c_{p+1} \neq 0$ by simply replacing p with $p + 1$. We denote by \mathbf{A} the $p \times m$ matrix having rows $\mathbf{a}'_1, \dots, \mathbf{a}'_p$, and we set $\mathbf{b} = (b_1, \dots, b_p)$ and $\mathbf{c} = (c_1, \dots, c_p)$. The function $f_p(\mathbf{x})$ is realized by a multilayer perceptron (MLP) having m inputs, p neurons in the hidden layer and one neuron in the output. Such quantities are called *weights* and they will be denoted by \mathbf{w} , so that $\mathbf{w} \in \mathbb{R}^{p(m+2)}$. It is well known that most functions, including any continuous function with a bounded support, can be approximated by models of the form (1).

2 Preliminaries and basic results

Let \mathcal{F} be the set of all functions of kind (1) for a fixed p with $1 \leq p \leq N$. The problem is to find the function $f^{(0)} = f(\mathbf{w}^{(0)})$ in the set \mathcal{F} which minimizes the *generalization error*:

$$\mathcal{E}(f) = \int [y - f(\mathbf{x})]^2 p(\mathbf{x}, y) d\mathbf{x} dy, \quad (2)$$

where the integral is over $\mathcal{X} \times \mathcal{Y}$. In practice, the distribution $p(\mathbf{x}, y)$ is unknown, but we have a sample $\mathcal{L} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, called learning set, of N i.i.d. realizations of (\mathbf{X}, Y) so that we compute the *empirical error*:

$$\widehat{\mathcal{E}}(f, \mathcal{L}) = \sum_{(\mathbf{x}_n, y_n) \in \mathcal{L}} (y_n - f(\mathbf{x}_n))^2 \quad (3)$$

and estimate the least squares parameters by minimizing (3). A theoretical problem concerns the *unidentifiability* of the parameters, see Hwang and Ding (1997). That is, there exist different functions of the form (1) with a different number of parameters that can approximate exactly the same relationship function $f(\mathbf{x})$. Results due to Bartlett (1998) show that this is due to the dependency of the generalization performance of an MLP on the size of the weights rather than on the size of the model (i.e. on the number of adaptive parameters). Here an important role is played by the quantity $\|\mathbf{c}\|_1 = \sum_{i=1}^p |c_i|$, that is by the sum of the values of the absolute weights between the hidden layer and the output. This is justified as follows. Let \mathcal{X}_1 and \mathcal{X}_2 be two populations in \mathbb{R}^m and set $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$; for each $\mathbf{x} \in \mathcal{X}$ and $y \in \{-1, +1\}$, let $y = +1$ if \mathbf{x} comes from \mathcal{X}_1 and $y = -1$ if \mathbf{x} comes from \mathcal{X}_2 . Moreover let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a discriminant function of type (1) such that \mathbf{x} is assigned to \mathcal{X}_1 if $f(\mathbf{x}) > 0$ and to \mathcal{X}_2 if $f(\mathbf{x}) < 0$; in other words the function f classifies correctly the point \mathbf{x} if and only if $y \cdot f(\mathbf{x}) > 0$; more generally, the function f classifies correctly the point \mathbf{x} with margin $\gamma > 0$ if and only if $y \cdot f(\mathbf{x}) \geq \gamma$. For a given learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$, where $y_n = 1$ if \mathbf{x}_n comes from \mathcal{X}_1 and $y_n = -1$ if \mathbf{x}_n comes from \mathcal{X}_2 , with $n = 1, \dots, N$, let us consider *misclassification error with margin* γ $\widehat{\mathcal{E}}_\gamma(f, \mathcal{L}) = \#\{n : y_n f(\mathbf{x}_n) < \gamma\} / N$, where $\#\{\cdot\}$ denotes the number of elements in the set $\{\cdot\}$, which is the proportion of the number of cases which are not correctly classified with margin

γ by f . For a given constant $C \geq 1$ consider only those \mathbf{c} for which $\|\mathbf{c}\|_1 \leq C$, then we have the following result:

Theorem 1 (Bartlett (1998)) Let P be a probability distribution on $\mathcal{X} \times \{-1, +1\}$, $0 < \gamma \leq 1$ and $0 < \eta \leq 1/2$. Let \mathcal{F} be the set of functions $f(\mathbf{x})$ of kind (1) such that $\sum_i |c_i| \leq C$, with $C \geq 1$. If the learning set \mathcal{L} is a sample of size N and has $\{-1, +1\}$ -valued targets, then with probability at least $1 - \eta$, for each $f \in \mathcal{F}$:

$$\mathcal{E}(f) \leq \widehat{\mathcal{E}}_\gamma(f, \mathcal{L}) + \varepsilon(\gamma, N, \eta)$$

where for a universal constant α , the quantity

$$\varepsilon(\gamma, N, \eta) = \sqrt{\frac{\alpha}{N} \left(\frac{C^2 m}{\gamma^2} \ln \left(\frac{C}{\gamma} \right) \ln^2 N - \ln \eta \right)}.$$

is called *confidence interval*. □

Thus the error is bounded by the sum of the empirical error with margin γ and by a quantity depending on $\|\mathbf{c}\|_1$ through C but not on the number of weights. Two other important results for our scope are given below.

Theorem 2 (Ingrassia (1999)) Let $\mathbf{x}_1, \dots, \mathbf{x}_p$ be p distinct points in $(-r, r)^m$ with $\mathbf{x}_i \neq \mathbf{0}$ ($i = 1, \dots, p$) and $\mathbf{A} = (a_{ij}) \in [-u, u]^{mp}$ be a $p \times m$ matrix, with $u = 1/m$. Let τ be a sigmoidal analytic function on $(-r, r)$, with $r > 0$. Then the points $\tau(\mathbf{A}\mathbf{x}_1), \dots, \tau(\mathbf{A}\mathbf{x}_p) \in \mathbb{R}^p$ are linearly independent for almost all matrices $\mathbf{A} = (a_{ij}) \in [-u, u]^{mp}$. □

This result proves that, given $N > m$ points $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^m$, the transformed points $\tau(\mathbf{A}\mathbf{x}_1), \dots, \tau(\mathbf{A}\mathbf{x}_N)$ generate an *over-space* of dimension $p > m$ if the matrix \mathbf{A} satisfies suitable conditions. In particular, the largest over-space is attained when $p = N$, that is when the hidden layer has as many units as the number of points in the learning set. This result has been generalized as follows.

Theorem 3 (Ingrassia and Morlini (2005)) Let \mathcal{L} be a given learning set and $f = \sum_{i=1}^p c_i \tau(\mathbf{a}'_i \mathbf{x})$. If $p = N$, then the error $\widehat{\mathcal{E}}(f, \mathcal{L})$ is zero for almost all matrices $\mathbf{A} \in [-1/m, 1/m]^{mp}$. □

3 Equivalent number of degrees of freedom

For a given $p \times m$ matrix \mathbf{A} , let \mathbf{T} be the $N \times p$ matrix having rows $\tau(\mathbf{A}\mathbf{x}_1)', \dots, \tau(\mathbf{A}\mathbf{x}_N)'$, with $p \leq N$. According to Theorems 2 and 3 the matrix \mathbf{T} has rank p (and then it is non-singular) for almost all matrices $\mathbf{A} \in [-1/m, 1/m]^{mp}$. The empirical error $\widehat{\mathcal{E}}_\gamma(f, \mathcal{L})$ can be written as:

$$\begin{aligned}\widehat{\mathcal{E}}_\gamma(f, \mathcal{L}) &= \sum_{(\mathbf{x}_n, y_n) \in \mathcal{L}} (y_n - f(\mathbf{x}_n))^2 = \sum_{(\mathbf{x}_n, y_n) \in \mathcal{L}} (y_n - \mathbf{c}'\tau(\mathbf{A}\mathbf{x}_n))^2 \\ &= (\mathbf{y} - \mathbf{T}\mathbf{c})'(\mathbf{y} - \mathbf{T}\mathbf{c}) = \mathbf{y}'\mathbf{y} - 2\mathbf{c}'\mathbf{T}'\mathbf{y} + \mathbf{c}'\mathbf{T}'\mathbf{T}\mathbf{c}\end{aligned}$$

and for any fixed matrix \mathbf{A} , the error $\widehat{\mathcal{E}}_\gamma(f, \mathcal{L})$ attains its minimum when $\mathbf{c} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$. Thus the matrix $\mathbf{H} = \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'$ is a projection matrix since $\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ and \mathbf{H} is symmetric, positive semidefinite, idempotent and it results:

$$\text{rank}(\mathbf{H}) = \text{trace}(\mathbf{H}) = \text{trace}\{\mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\} = \text{trace}\{(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{T}\} = p$$

so that $\widehat{\mathbf{y}}$ lies in the space \mathbb{R}^p and thus to the model $f(\mathbf{x}) = \sum_{i=1}^p c_i \tau(\mathbf{a}'_i \mathbf{x})$ should be assigned p equivalent number of degrees of freedom (e.d.f). When the error is given by the following weight decay cost function:

$$\begin{aligned}\widehat{\mathcal{E}}^*(f; \mathcal{L}) &= \widehat{\mathcal{E}}(f; \mathcal{L}) + \lambda \sum w_i^2 \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{c}'\mathbf{T}'\mathbf{y} + \mathbf{c}'\mathbf{T}'\mathbf{T}\mathbf{c} + \lambda \text{tr}(\mathbf{A}\mathbf{A}') + \lambda \mathbf{c}'\mathbf{c}\end{aligned}$$

the equivalent degrees of freedom are:

$$k = \text{tr}(\mathbf{H}_\lambda) = \text{tr}\{\mathbf{T}(\mathbf{T}'\mathbf{T} + \lambda\mathbf{I}_p)^{-1}\mathbf{T}'\} = p - \sum_{i=1}^p \frac{\lambda}{l_i + \lambda}$$

which shows that p is decreased by the quantity $\lambda \text{tr}\{(\mathbf{T}'\mathbf{T} + \lambda\mathbf{I}_p)^{-1}\}$. Since $\mathbf{T}'\mathbf{T}$ is positive semidefinite, the p eigenvalues of $\mathbf{T}'\mathbf{T}$, say l_1, \dots, l_p , are non-negative. Thus $(\mathbf{T}'\mathbf{T} + \lambda\mathbf{I}_p)$ has eigenvalues $(l_1 + \lambda), \dots, (l_p + \lambda)$ and then the eigenvalues of $(\mathbf{T}'\mathbf{T} + \lambda\mathbf{I}_p)^{-1}$ are $(l_1 + \lambda)^{-1}, \dots, (l_p + \lambda)^{-1}$.

4 Model selection criteria

In the general framework of model selection, we suppose there are f_{p_1}, \dots, f_{p_K} models of the form (1). Since the estimation in statistical models may be thought of as the choice of a single value of the parameter chosen to represent the distribution (according to some criterion), model selection may be thought of in this framework as the estimation applied to the model f_{p_h} , with $h = 1, \dots, K$. The only special issue is that the set of models is discrete and has a finite range. There may be occasions when one model clearly dominates the others and the choice is unobjectionable, and other occasions when there are several competing models that are supported in some sense by the data. Due to the *unidentifiability* of the parameters, there may be no particular reasons for choosing a single best model over the others according to some criterion. On the contrary, it makes more sense to "deselect" models that are obviously poor, maintaining a subset for further considerations regarding, for example, the computational costs. The following indexes are generally used

for model selection since they be carried out easily and yield results that can be interpreted by most users; they are also general enough to handle with a wide variety of problems:

$$\begin{aligned} \text{AIC} &:= \log(\widehat{\mathcal{E}}(f)) + \frac{2k}{N} & \text{BIC} &:= \log(\widehat{\mathcal{E}}(f)) + \frac{k \log(N)}{N} \\ \text{GCV} &:= \widehat{\mathcal{E}}(f) \left(1 - \frac{k}{N}\right)^{-2} \end{aligned}$$

where k denotes the number of degrees of freedom of the model f . The AIC and BIC present different forms in literature, here we follow Raftery (1995). Some of these criteria obey the likelihood principle, that is they have some frequentist asymptotic justification; some others correspond to a Bayesian decision problem. It is not the goal of this paper to face the outgoing discussion about their relative importance or to bring coherence to the two different perspectives of asymptotic and Bayesian-theoretic justification. In this work, via Monte Carlo simulations, we first aim at describing the different behavior of these indexes; then, we wish to determine whether such values and the model choice are affected by how the degrees of freedoms are computed and by how the empirical error minimization is performed. In Ingrassia and Morlini (2005) a Monte Carlo study has been drawn with small data sets. For these data, BIC has been shown to select models with a smaller $k = p$ than those selected by the other criteria, in agreements with previous results (see e.g. Katz (1981), Koehler and Murphree (1988), Kadane and Lazar (2004)). A comparison with the criteria computed using the e.d.f. and $k = W$, where W is the number of all parameters in the model, has also be drawn and this shows that, when $k = W$, some indexes may assume negative values becoming useless. Values across simulations also reveal a higher variability and the presence of anomalous peaks. Another analysis concerning simulated data has shown the ability of the UEV to estimate σ^2 when $k = p$. In this work we present further results, carried out in Matlab, based on large datasets: the *Abalone* and the *Boston Housing* (www.ics.uci.edu/~mllearn/).

5 Numerical studies

The *Abalone Data* consists of 4177 instances with 7 input variables and one discrete output variable and the *Boston Housing* data consists of 506 instances concerning 13 input variables and one continuous target variable. Observations are split into a training set of dimension 3133 for the *Abalone Data* and 400 for the *Boston Housing* and a validation set of dimension 1044 for the first data set and and 106 for the second one. In order to avoid overfitting, we estimate the parameters both by minimizing the sum-of-squares error function with the stopped training strategy and by minimizing the weight decay cost function. To interpret the following numerical results, it is worth noting

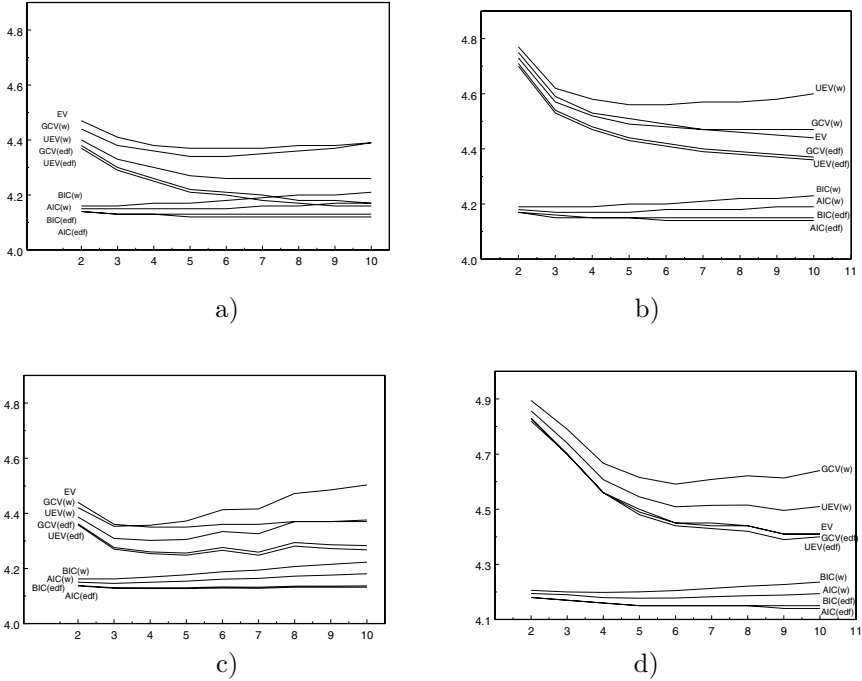


Fig. 1. Mean values of model selection criteria for the Abalone data set obtained with weight decay and a) $\lambda = 0.005$, b) $\lambda = 0.05$, c) λ chosen by cross validation and d) stopped training.

that when the weight decay function is used, the error on the validation set (EV) may be considered as an estimate of the generalization error since the observations are independent from those used for estimating the parameters. On the contrary, the error on the validation set is indirectly used for estimating the parameters if the stopped training strategy is applied and cannot be considered as a generalization error estimate. For the Abalone data, the mean values obtained by repeating the estimates 100 times, with different splits of the data in the training and validation sets, are reported in Fig. 1; moreover main results referred to the Boston Housing data are reported in Table 1. The first conclusion we draw, especially evident from Table 1, is that, for different values of λ (ranging from 0.005 to 0.01) model selection criteria computed using the e.d.f., that is with $k = p$ and $k = p - \sum_{i=1}^p \lambda / (l_i + \lambda)$ are nearly identical and lead to the same model choice. Since $k = p - \sum_{i=1}^p \lambda / (l_i + \lambda)$ is not readily available in software packages, the choice $k = p$ is shown to provide a concise, simple and reliable approximation of this value. The second conclusion we draw is that BIC selects smaller models, with respect to those selected by the other criteria, when $k = W$. Indeed, it leads to the choice of the same model selected by the other indexes, when $k = \text{e.d.f.}$ If the true

Table 1. Comparison among mean values of model selection criteria obtained with the Boston Housing data, with $k = p - \sum_{i=1}^p \lambda/(l_i + \lambda)$, $k = p$, $k = W$ and with $\lambda = 0.005$ and $\lambda = 0.01$. Bold values refer to the model selection.

		$\lambda = 0.005$								
		$k = p$			$k = p - \sum_{i=1}^p \frac{\lambda}{l_i + \lambda}$			$k = W$		
p	EV	AIC	BIC	GCV	AIC	BIC	GCV	AIC	BIC	GCV
2	17.89	8.56	8.59	13.08	8.55	8.59	13.06	8.76	9.18	16.08
3	17.92	8.42	8.46	11.29	8.40	8.45	11.26	8.68	9.23	14.96
4	17.59	8.32	8.37	10.31	8.31	8.36	10.25	8.65	9.35	14.77
5	18.36	8.32	8.38	10.29	8.31	8.36	10.21	8.71	9.55	16.00
6	19.20	8.39	8.46	11.02	8.37	8.43	10.90	8.85	9.82	18.66
7	20.10	8.32	8.40	10.31	8.30	8.36	10.17	8.84	9.96	19.10
8	20.68	8.38	8.47	10.96	8.36	8.43	10.78	8.97	10.23	22.31

		$\lambda = 0.01$								
		$k = p$			$k = p - \sum_{i=1}^p \frac{\lambda}{l_i + \lambda}$			$k = W$		
p	EV	AIC	BIC	GCV	AIC	BIC	GCV	AIC	BIC	GCV
2	17.90	8.54	8.57	12.84	8.54	8.57	12.83	8.74	9.16	15.79
3	17.27	8.45	8.49	11.64	8.44	8.48	11.60	8.71	9.26	15.42
4	17.00	8.30	8.35	10.07	8.29	8.34	10.02	8.63	9.32	14.43
5	17.95	8.31	8.37	10.17	8.29	8.35	10.09	8.70	9.54	15.80
6	19.20	8.39	8.46	11.02	8.37	8.43	10.90	8.85	9.82	18.66
7	20.10	8.32	8.40	10.31	8.30	8.36	10.17	8.84	9.96	19.10
8	20.68	8.38	8.47	10.96	8.36	8.43	10.78	8.97	10.23	22.31

underlying model is chosen to be as the one with the smallest validation error, using $k = \text{e.d.f.}$ instead of $k = W$, leads to choices which are never considerably different and sometimes are considerably better (for example, when λ is small and BIC is used). Another conclusion we draw from Table 1 and Fig. 1 is that the GCV is always larger than the other criteria and has a smaller spread with the validation error, which is a reliable estimate of the generalization error when the weight decay approach is used. Moreover, GCV has a less smoother pattern with respect to the dimension p of the model and a scree test based on the plot of their values against p may be used to choose the optimal dimension p of the model. If the graph drops sharply, followed by a straight line with a much smaller slope, we may choose p equal to the value before the straight line begins. Fig. 1 a), b) and c) clearly indicate to choose $p=3$ while Fig. 1 d) suggest $p=6$. In the scree plots obtained from Table 1 (not reported for economy of space) there is clearly a discernible bend in slope at $p = 4$ for $\lambda=0.005$ and 0.01 . In another case, with $\lambda = 0.05$ the bend in slope is at $p = 5$. In both data sets, when $k = \text{e.d.f.}$, these criteria are nearly identical and lead to stable estimates of the generalization error and stable model choices, for different p . By comparing the results obtained with

different values of λ , it is apparent that increasing the value of λ does increase the numbers of possible better models over the others and, in general, leads to less parsimonious models. In this case model choice should be based on the scree plot instead of on the basis of the absolute minimum value. The e.d.f. are still shown to work well, even if they are based on the achievement of the absolute minimum of the error function (3) which has a wider spread between the minimum of weight decay cost function, as long as λ increases.

6 Concluding remarks

Based on this computational study, we can draw conclusions about the comparisons of different degrees of freedoms given to nonlinear projection models of the form (1) and about the reliability of the model selection criteria routinely implemented by software developers. In particular, our study has shown that BIC tends to select more parsimonious models than GCV and AIC when $k = W$. The GCV criterion gives a larger value of the generalization error, which is in agreement with the empirical error computed on new independent patterns. The choice $k = p$ gives a good approximation of the trace of the projection matrix for projection models of the form (1); it leads to values of selection criteria nearly identical to those obtained with the trace. Using $k = p$ instead of $k = W$ leads to model choices which are never worst and sometimes are better (for example, when BIC is used). Using a scree test plot to select a single best model is increasingly important as long as the value of λ increases. Further simulation studies on the e.d.f. are in progress and the obtained results will be summarized in a future work.

References

- BARTLETT, P.L. (1998): The Sample Complexity of Pattern Classification With Neural Networks: The Size of the Weights Is More Important Than the Size of the Network. *IEEE Transaction on Information Theory*, 44, 525–536.
- HWANG, J.T.G. and DING, A.A. (1997): GPrediction Intervals for Artificial Neural Networks. *Journal of the American Statistical Association*, 92, 438, 748–757.
- INGRASSIA, S. (1999): Geometrical Aspects of Discrimination by Multilayer Perceptrons. *Journal of Multivariate Analysis*, 68, 226–234.
- INGRASSIA, S. and MORLINI, I. (2005): Neural Network Modeling for Small Datasets. *Technometrics*, 47, 297–311.
- KADANE, J.P. and LAZAR, N.A. (2004): Methods and Criteria for Model Selection. *Journal of the American Statistical Association*, 99, 279–290.
- KATZ, R.W. (1981): On Some Criteria for Estimating the Order of a Markov Chain. *Technometrics*, 23, 243–249.
- KOEHLER, A.B. and MURPHREE, E.S. (1988): A Comparison of the Aikake and Schwarz Criteria for Selecting Model Order. *Applied Statistics*, 37, 187–195.
- RAFTERY, A.E. (1995): Bayesian Model Selection in Social Research. *Sociological Methodology*, 25, 111–163.